>> So we're going to get started here. I'm really excited to introduce Dr. Merce Crosas who is the chief data science and technology officer at the Institute for Quantitative Social Science at Harvard University. She has more than 10 years of experience leading the Dataverse project, which some of you may have heard of. An open source repository framework for sharing and archiving research data. And more than 15 years of experience building data management and analysis systems in industry and academia. One of the reasons why we wanted Dr. Crosas to join us today is that we were talking about born digital curation. And research is something that I think all of the residents are intimidated by, but also excited to kind of take up the challenge. It's been gaining a lot of attention recently because of changing funding requirements. Encouraging scientists to share their research. And there's a big push in the community for openness in research. But the technical complexities of accessioning, preserving and making this data accessible makes people like Merce to us again another digital preservation rock star that we hope to learn from today. So thank you for coming.

>> Merce Crosas: Thank you.

[Applause]

>> Merce Crosas: Many thanks. I'm delighted to be here. Thanks for the invitation. Everyone has been very welcoming and nice. I appreciate that. So I usually like to start when I talk about data publishing with a little bit of history about scholarly publishing in general. It is relevant to show how data publishing in a way is an evolution of the increase of complexity of scholarly publishing. So we will start from 1665 to the 20th century and beginning with just one journal in the philosophical transactions royal society. We see that around 1700 there are about three journals. About 1800, nearly 1800, there are ten journals. 400 journals be 1900. And in 2000, 14,000 journals depending on how you count this. But that's based on peer-reviewed journals. So the increase of the research output since the beginning of the modern science, basically in the 1600's until now, doubles every 20 years. And this increase of research output is not only in the amount of journals or amount of articles in publications, but also in the complexity of what research is. So the moral of the philosophical transactions, the Nullius Verba, don't take anybody's word for it. Any research output needs to provide evidence of why you make some claims, why you make some conclusions for what those results are coming from. And that evidence a lot of times in terms of data, in terms of analysis, of understanding what have you applied to the data. So in the very early part of the modern science, the data were more description of the observations of the experiment. There was less interpretation or explanation of that data. But they were already present in the article. There were already some figures and some tables and very much mattered as part of the text. After the first 50-100 years, you start seeing some articles that are citing previous work. And it's again part of the showing evidence of your work. You want to say, "Don't take my word for it." You want to show that what you are saying is based on previous experiments, previous

observations. Until the 1780's there were more interesting data visualizations that tried to explain the data – well, to show the data in different ways, in a statistical way. They are the first with graphs and bar charts and then pie charts a little bit later. We start seeing more articles that contain figures and data in the article in form of tables. That complexity of research output also means that you need a methods section. You need to explain why the analysis has been applied to the data. The first scatterplot actually doesn't appear until the early 1800's. 1833 by Herschal and 1896 by [inaudible] are the very first ones that we see. And as we go into 1900's, then practically most articles rely on figures and data visualizations and on data tables. And often there are many within an article. And also all the articles rely on citation from previous work. And the citations then become part of the scholarly credit. So a way that grows is in a book Communicating Science. It's a history from the 17th century to the 20th century. So what happens with scholarly publishing is they say – it's a fantastic book by the way. It's very tedious, a lot of information. But a very good analysis of these centuries of scholarly publishing. So [inaudible] and others say that these changes in scholarly publishing, what we're seeing is an adaptation of the complexity of how research output changes and becomes increasingly complex. And so the article, the way that we provide again this scholarly output becomes more efficient to accommodate that. So in the 18th century we start seeing these first formal components are more structured in the article, right? You see it in an introduction, conclusions, these tables and citations. In the 19th century we see there is an increase in complexity. It shows us the results and tries to explain the data instead of being an observation. To accommodate this we see more data visuals and sections like the methods section to be able to explain that data. And in the 20th century we start seeing a little more structured quantitative data with the rise of statistics. Interestingly enough, the data visualization golden ages of the 19th century decreases because there is a focus on statistics. And it starts to increase again in the late 20th century with the growth of the amount of data in the world. So we start seeing a lot more types of data. Every time, more difficult to contain that data, to be able to explain the data in the scholarly paper, right? Also the groups of scientists has increased from the very few hundreds in the earlier times of modern science where the scientists were mostly enthusiasts of science more than professionals of science, right? To a few million now. I think there are about 7 million scientists in the world as of now, although maybe it's really half of them that are actually in practice. And then so science comes with that growth of a number of publications, of the number of scientists themselves. Becomes more specialized. We see that not only – well, we go from one journal in 1665 to 14,000 peer-reviewed journals. But we also see that every new journal appears more or less from every 150 authors, and is read only by 100 in the community within that discipline. So it's very specialized. So what happens now in the last decades and how it is relevant to data publishing. The scholarly output increases even at a higher rate. If we were seeing that the number of publications were doubling every 20 years, now they are doubling more or less every 15 years. The latest estimate I found is there are about 80,000 known peer-review journal now in 2016. So at the same time we see that with an increase of the amount of data, amount of specialization and repositories. We started with the first social science archives from the beginning of the 20th century at the University of North Carolina and [inaudible] and others. And then we also see the first large biomedical databases like [inaudible]. Now in 2016 if you look at the number of research data repositories that exist, you can go to [inaudible]. And there are 15,000 registered. And we don't know to what extent every one of them are active, that you would consider that they actually preserve data. But there is a boom of data repositories for research. So how do we define data publishing or how data publishing is born? So is the union of this scholarly publishing that we've been seeing, the data increasing in complexity and the type of output, inputs and outputs that we have? And the data tidings of the 20th century? So from scholarly publishing, we want to distribute the research output, right? And we do this with a division and credit to authors through dissemination process and providing ways of finding and reusing the publications. At the same time, data archiving focuses on the long-term access to data, right? And so they are focused on accessibility, preservation and the finding and reuse as well. So when we merge them, we see that we start seeing this concept of data publishing that is really a new form of publication that exists only in the last 10 years. And what I argue is that data publishing in a way is an evolution, a continuous adaptation. The same way that Gross et al argued that a lot of the forms of the scholarly article and the way they change in terms of the structure of it, the style, the type of argument, that's to become more efficient as the research becomes more complex. The same way that in the last decades research has become more complex with a lot more data, a lot more software that accompanies the research claim, right, the results. That means that the publication has to adapt to that. And by adapting to that, needs to support a type of data publishing. And the same starts happening for software. You have a software publication as a form of publishing. So what we see is that data and software have become common. Input and output of research in scholarly article cannot [inaudible] anymore with the vast amounts of data and software. As an input and output of research, we need to consider the data then also as a citable product of research, right? And that's important not only for attribution, but for being able to provide persistent access to data, validation and reuse of the data in various forms of the scholarly product. So what is needed for what I call FAIR data publishing? FAIR is not a term that I made. It refers to – how many people here have heard about the FAIR principles? Maybe they are more common for people that work on research data. But it's Findable, Accessible, Interoperable and Reusable. And you can apply that to software. So in order to have the fair data publishing, you need data citation. I've been part for I think the last three years of a group that is led by Force 11, a scholarly communication organization, right, that's community-based. That has defined a set of what we call joint data citation principles, a set of principles for data citation. It's now been adopted very broadly by publishers, by funders and by journal editors. So for data citation you need a persistent identifier to reference the data uniquely. You need to support versions of fixity. Attribution to authors and attribution to the repository. Then you need metadata. And that metadata while it is at a high level is a catalog to discover location data. But then you need a lot more extensive metadata than that if you need to reuse the dataset, right? You need sufficient information to be able to understand what the data values are about. And then you need a repository. Another difference when you're talking about data publishing, you cannot anymore just distribute the data set and then that's it. Because in order to be able to access it, you need a repository that will house that data. So it is the place where you have digital access to metadata and data. It is the responsible archiving and preserving for long-term access. And it provides them the compatibility needed and the standards needed to be able to work with other systems and be able to provide API's for others to use the data. So on the FAIR guiding principles, if you haven't heard about it, you can go in the scientific data. There is a publication on that with a large number of authors. As you can see, somewhere in the middle of this. So let's move into dataverse, the dataverse system and how the dataverse helps with data publishing. So the dataverse is an open-source data repository system. I argue here also that it serves as a solution for data publishing and a way for publishing FAIR research data, right? Research data that is findable, accessible, interoperable and reusable. And it's installing number of universities, organizations around the world. About 17 installations now in production. They come for each other through harvesting metadata from one repository to another. Within one dataverse repository there can be hundreds of dataverses, and each one is its own repository. The Harvard dataverse actually is one of the largest ones that has a – well it's open to all researchers in the world across any discipline. It started with the social science, since we created it for the quantitative social sciences. But it's now used by medicine and astronomy and across different disciplines. And it has more than 1,500 dataverses and 60,000 datasets I think. Maybe 400,000 or so data files with that. And there's a number of almost 2 million downloads. It's getting there. From other researchers who are using those datasets. And just to move quickly, the dataverse can be used within an institution as an institutional data repository or an organization. Also for a scientific community for their research repository. Or for a journal or within a research group, right? And dataverse contains datasets and what we mean by the dataset is more than just the data files, but the data under metadata. And any other accompanying files that could be part of either the codebook or scripts or codes themselves that will help you understand and compliment a dataset. So it can contain as many files as one needs. And of course the metadata is a big part of it. So how do we do the data citation first? In order to support data publishing properly and support data citation? How do we do that in dataverse? So this is a dataverse landing page, a dataset landing page which this is a random dataset that I just picked. But it has – well it's titled the numbers of the metrics. Then it has a citation that contains the authors, the published year, the published year in the repository, that is. The dataset title. A global persistent identifier. In that case we provide DOI's. Those are DOI's that are registered in data site, either through the easy ID system at the California digital library, or the Data Site of API, if you're a member of Data Site. Data Site is a recent organization that provides DOI's for data and is already supporting DOI's across many repositories. So then the data citation also includes in a way the publisher, but in this case it's the repository that houses the data. And that includes some form of versioning. Because another difference between publishing data and publishing something in print is that the dataset can be dynamic, continually changing over time. There might be new versions, new updates on the data files. And we provide a versioning system basically that would allow you, if there is a major change in your dataset, it will include that in the citation. So when you reference that dataset, others know what version you're referencing too. And we're working on it for more dynamic datasets or streaming data, ways of including that. Instead of a version number, being part of a time range or time stamp to be able to say that that dataset was accessed at that time, that included up to that amount of data. That's actually a little more challenging than it seems. How do you do side-streaming of dynamic data? But it's one of the projects we have now. So based on these data citation principles that I talked about before, we created a set of guidelines for data citation implementation. And these are the basic elements of data citation. You have the citation and this is the documents citing the data. That citation includes a persistent identifier that results to a landing page. And the landing page is the page in the repository. The repository might not be the same repository that stores the data, but it's some form of repository. It has a page where you will find enough information to find what that dataset is about, how to access the data themselves, right? And contain all the metadata about that dataset. And any other terms of use or information you need, over your main information. And that metadata is also registered into Data Site so Cross Ref and others can search the metadata and find the dataset more easily by having more information about the data that is in the repository. And then there is the storage part that again can be in the repository or separate or remote. It is important also to think of the data citation and the landing page as the minimum a repository needs to guarantee that exists. So they might be at some point that data file is restricted or cannot be accessed for some reason. But still you have enough information of what happened to that data. The citation will not be a dead citation, right? So how do we do the second feature of data publishing, supporting data publishing, is metadata. How do we handle metadata in the dataverse? So we have three levels of metadata. First, the citation metadata, what we call citation metadata. But it's very high-level descriptive metadata of the dataset. And that maps to standards like doubling core and Data Site schema. And includes the author, the title, maybe an abstract, et cetera. We have another level. There is more extensive metadata that is domain-specific metadata. In this case you might need that information to understand how was that data collected, what was the methodology? If it's a survey, how it was performed. If it's an experiment, also details about the experiment as well. Then we map these two standards, if they are existing as standards in that scientific domain that exists to describe data. Not in any case we find standards so that we are allowed for data repositories to create a custom metadata block that will describe more accurately the type of dataset. But we followed the DDI, the Data Description Initiative, that was created by CPSR for social science and humanities data. We also work with the ISATOP that was originated in Oxford. And it's part of the NIH [inaudible] project, which is part of the Big Data to Knowledge Program at NIH. And also we map it to some fields from the virtual observatory metadata standard. And then when possible we also extract metadata. That's more challenging but it's actually a very important piece because it helps to understand – well, more than to understand – to be able to reuse that data. Be able to reformat it. If we extract that information, as I'll show you for some cases, then you can have the data values independent of the original software that the dataset was created with. And that top-level metadata can be combined to recreate the dataset in any other format. So for the case of tabular metadata we use for statistical tables basically, we extract the variable metadata and we map it to DDI to follow a standard. So all the variable information is captured in that DDI. Are people familiar here with the DDI and Data Site Schema? Okay. We also have creating adjacent schema, probably will be adjacent LD, Lincoln Data schema, to capture all the dataverse metadata for any dataset. So how do we do the information extraction of these tabular files? So you upload a data file in dataverse, that is on our data frame, in our format. In S Data, in SPSS, in Excel. Then it has basically columns. Each column is a variable, each row observation. We will automatically extract the variable information. And that will include what is the variable level, name, type, et cetera, into a metadata file. And then have another metadata file that just has the values. We can recombine those to provide a new type to be able to have a data file basically that is independent of the format it was uploaded with. So that's important for preservation of that file again, to make it independent of the software. So for accessibility, also providing more formats for download. And also for search and reuse of that dataset, right? Because we can search now all the variables in that data file. We apply also a calculation on the data values. And it checks some calculation for fixity, just to guarantee that later in another time the file has not changed, right? And that's a universal numerical fingerprint. It's different than just an MD5, because it would apply to an S data file or an Excel data file. Because it again is independent of the file format. It just applies to the content of the file, not the metadata part. We do something similar for FITS files that are used in astronomy. If you have a FITS file, there is scattered information that is part of basically metadata that you could use for discovery, for finding that data set and knowing what it's about. So those could be coordinates about the observation and other automatic information. So we separate those again as metadata of the file and also the actual data objects inside the file. And we are applying that same process of extracting information from data files to other files. There are some that are working from instruments in biomedical – well, for some of the biomedical experiments – that we're also trying to figure out the best way to convert it into a standard format. Sometimes it's a challenge because for example, in the case of astronomy, they have decided on a standard format. Most astronomers will be familiar with FITS. But in biomedicine there is practically one format for every new instrument that gets created. And finding a standard format across all of those datatypes is not easy. But their OMERO project standard is one that we're considering to use and do something similar that we do for astronomy data. The same for geospatial data files. We are extracting also the geospatial information. And be able to then, if you have that information, you can visualize that data into a map or into some data visualization tool that deals with JES data, right? So in addition to these main features of data citation and metadata, dataverse provides a lot of additional features to help you with data publishing and data sharing. One of them is tier access. Permissions are set in the dataset. The first default one is completely open. So when you upload the dataset in dataverse, since we encourage open data, the default would be open and you would have a CC0 waiver apply to that dataset. That means that at the time that you publish that dataset, the metadata are open and the files are open too. No rescue chance at all. You don't need to do anything to access the data. Just click and load the data. But that doesn't fit for all researchers or for each scenario. So there are cases where we provide ability to set up a guest book. That will mean that you still have the metadata and the files open for the dataset, but you are asking the users before download to enter some information about what they want that data, and be able to track that information. Then one step further is applying some terms of use to the dataset. In that case then you cannot use CC0 waiver anymore for the dataset because you are asking users not to use it in a certain way, right? In that case we provide a click-through agreement to be able to access the data. So even though it's open, it's not fully open. Then we have also the ability to restrict the data files. But the metadata is always public. From the point that you publish the metadata, the metadata has to be available to others to understand what that dataset is about. And the data restriction, the first level is just simply requesting the data through the user interface. But in some cases the data files are restricted under terms of access that would require to send an application and a longer process. And also with data publishing work flows for being able to curate that dataset, reviewed by others. So when you create a dataset, you have what it was showing about, the landing page of the dataset is restricted. Only your collaborators or the data authors can view the dataset. You can invite reviewers, even anonymous reviewers. This is a new feature we're adding now, because journals that use the dataverse, the repository for the data that accompanies the article, were asking for having a data review as part of the peer review of the article. So there is this step of reviewing. In some cases the dataset is not associated to a journal because it doesn't have to be associated necessarily again to a publication, right? You might invite collaborators or others to also curate the additional metadata, reformat the data before you publish it. And once you publish it, if the metadata is open, you have version one in your data citation. But you can edit at any other time. And there are two types of edits. A minor change that would include a minor version increase. And in that case the citation doesn't change, right? You might have only a fix in the metadata. You don't need to change the citation itself. Or you have a major change, or in this case your version changes completely and the citation is different. And again, that is important because if you're using the citation to reference the work that you've done to validate some analysis in your study with that dataset, you want to make sure that others are accessing the right version, or the version that needs to be on the citation. There are a lot of features and API's that are described in dataverse that will work. I won't go into detail here. Just want to do a check on time. Okay. So one of the projects that we have started with [inaudible] from the Harvard medical school and their Hanfley Foundation grant. It's a biomedical dataverse. In this case what we're trying to address is handling much larger datasets, the one that we can publish in a dataverse. We have a limit just because they are uploaded through an HTTP upload. If the file is more than one or two gigabytes, it starts to be too slow. Also if you have hundreds of files in a dataset, it's not very viable to download it this way. So with this grant or this collaboration that we started as the use case for structural biology data. It might have datasets with hundreds of thousands in each one, a few gigabytes. We are building a new way through the back end. You can publish a dataset with the metadata in dataverse, but you have a way to upload the data and hold the active structure of that data in dataverse through non-HTTP work flow. And we are also including as part of this collaboration with not only Peter but also Caroline in the medical school the LINCS dataverse or database to also provide a persistent repository for this library of integrated network based cellular signatures, right? The LINCS. So we are just in the first year of this project. We have two more years, as the pilot project has got a lot of attention already for the biomedical community to use, as a way to use dataverse for their data repositories. We also have another grant from Sloan working on social science big data. We are able to provide also big datasets or large datasets through dataverse, as also a way to publish them with the appropriate metadata, with dealing with the streaming data issues and the citation. But what is the additional challenge that we encounter with data publishing and research? The same way I was showing, research becomes more complex in many ways, including more data, more software. It also becomes more complex in the sense it deals with sensitive data. And even more when you start thinking about big data and more and more ways to be able to aggregate or to connect data that are made public with datasets that are nonpublic. And you can more easily – even when you are thinking that your data is de-identified, you can re-identify it if you have sufficient information. I worked very closely with Zlatana Sweeny, who keeps proving over and over how easy it is to re-identify a lot of datasets that were thought to be de-identified, right? So we work with her. Well, before showing you the project that we're working with Zlatana Sweeny, I wanted to show that even though many repositories are collecting research data, well are very widely-used already, they still all say – maybe not all, but most of them – still say don't upload any data that can be identifiable. Don't upload sensitive data. That's what we have in our policy for the Harvard dataverse. It's the same for the dry repository for example. It represents and warrants that the content does not contain any information which identifies or which can be used in conjunction with other publicly available information to personally identify an individual. This becomes harder and harder to be able to guarantee, right? So even though we all give these types of statements in our websites, how do you avoid that? The same with [inaudible]. Or maybe in this case it's easier to avoid, but it says that it does not include any data that could reveal the personal identity of the source. Again, when you start combining the existing dataset, it becomes harder to guarantee that. So still what we don't want to say is, "Well, we restrict all data. We don't make it available to others." Because it can have sensitive information. But we want to guarantee as much as possible that we're mindful of privacy regulations and other issues that could cause identification of datasets that have been previously de-identified. So just recently we published with Zlatana Sweeny, Michael Barnes and I a paper on sharing sensitive data with confidence, the data tag system. And since then we've been working on this system. It's not available yet in production, but it is under development and testing. So what a data tag is, is a set of security features and access requirements for file handling, for data file handling. And a data tag repository is a repository that stores and shares data files in accordance with a standardized level of security and access requirements. Basically it is compliant with a set of data tags. So these data tag levels, we've defined them this way so far. This was based on security levels that were done at Harvard for sensitive data. And also we reviewed with other universities that had similar security levels. But we added the access requirements onto that. There is one more column that is not here that is for terms of use. So you apply some security features and you basically create a tier layer as well. Different layers of security for your data set. And you would classify that dataset in one of those levels, right? So security is one, access is one and then how you use the data is another one. But that one, the terms of use, is a little harder to codify. They might not be falling to those levels necessarily. But it might be some terms of use that you will need to have with your dataset after you access it. Basically these data tag from blue, green, yellow, orange, red and crimson with all the metadata that goes with that data tag. It's a machine-readable policy that lives with your dataset. So wherever it goes, it has this policy. It goes from public, and green is what we call controlled public. This we are already supporting in the dataverse. But now we are implementing support for yellow, orange and red. The crimson case is a little more difficult. We don't know if we'll be able to support that. But the idea is that you can provide different levels based on the sensitivity of the contents of your data. And how it works with our repository. This doesn't need to work with dataverse only. It could work with any data management system or data repository that deals with data files that have sensitive data. So you first provide it – well when you upload the data, you can go through an interview that will ask you questions about your data. And based on the answers to those questions, we'll classify it as blue, green, yellow, up to crimson, right? In some cases this process would be too difficult to do remotely. Or maybe the interview will tell you that you need to go to the IRB or so to be able to classify your dataset. So with that, once you have that data tag, it gets uploaded into the repository with that policy, machine-readable policy. And the repository knows if its data tag is compliant. It knows what to do with it. So it has to store it in an encrypted form. It can only allow access in a certain way based on the data tag that is associated with the dataset. We work on collaboration, the privacy tools website in the school of engineering at Harvard. We're providing additional tools that when that dataset, if for example in this case it has a red tag and it's hard to ask or request access to the data, we would provide some tools that are based on differential privacy algorithms. You would still be able to ask some questions about the data in the sense of provide summary statistics that would preserve the privacy of the dataset but still will give you some information about what the data set is about. This is also under development. This is part of an NSF grant that we have one more year to put it into practice with those software tools available in the repository. So here I'll give you an example of this data tags interview. In a way it's sort of like – I apologize, it is hard to read. This is similar to TurboTax. So we build the user interface so it would ask you questions about your dataset, but underneath there is an interview engine or decision three engine that would codify regulations like, for example, HIPPA. You take HIPPA and say, “Based on this revelation, what you need to do to the data based on those questions.” We've been doing that for HIPPA and FERPA and a few other regulations. You just have to know there are a couple thousand privacy laws in the US. So fortunately you can group them into about 30 types. So we are going through those types of laws to see how we can codify it so that interview will help you to provide an answer of what data tag would correspond to your dataset. And so in this case it will ask you – you think of TurboTax, right? It will ask you a question and then based on your answer you will go to the next one. So here, “Does the data concern living persons?” Answer yes or no. Then, “Do the data contain health or medical information?” You keep asking and you're creating this policy basically that lives with your dataset. Here it shows all these. The final user interface doesn't need to show you all the code that is behind it. But it provides a policy. And one more question. There are explanations to answer the questions. So at the end you get an orange tag for example, and the repository knows what to do with that dataset. And basically I'll finish here. That's again for inviting me here. I just wanted to share also that we have an annual dataverse community meeting mid-July and I hope that if you're interested that you will join us. Thanks a lot.

[Applause]

>> So we'll have to save – if you have questions for Dr. Crosas, we'll do it during the part at the end of the symposium.

>> Merce Crosas: Yes.

>> Okay. So up next we're pleased to have Caroline Catchpole. Caroline is the metro mobile digitization specialist for Culture In Transit, which is a project funded by the John S. and James L. Knight Foundation. The project aims to bring mobile scanning equipment to smaller libraries, archives, museums and the communities they serve. The outreach center digitization model aims to democratize and diversify New York City's historical record. I want to say on a personal note that I'm really excited to have Caroline here and to hear more about Culture In Transit for my project that I presented before on the memory lab. We're interested in doing something mobile, and they have done an incredible job with documentation. Everything from how heavy their equipment

is to carry around, to workflows of how they accept pictures from people's smartphones that they bring in for one of these events. So really happy to have her here today and to hear more.

[Applause]

>> Caroline Catchpole: Hello. Right, sorry. Okay, good afternoon. It's a pleasure to be here. It's been a great day so far and hopefully I can live up to the other presentations. So as Jamie was saying, I work on the Culture In Transit project. And Culture In Transit was a 2015 winner of the Knight Foundation's new challenge on libraries. So we are an 18-month project. We started in February. I guess really the activity phase of the project where myself and the other two mobile digitization specialists started was in May. So really since last May we've been in the activity phase of the project. And I guess our USP is a mobile digitization kit. So what we do is we provide free digitization services to community members in Queens and Brooklyn through the public library systems. And myself at Metro, I provide free digitization services for Metro Library council members. So I will go to an institution who is under-resourced. Maybe they don't have any staff to do digitization. They don't have any money. And so I will go and digitize a collection for them. And then we at Metro will put that on our digital hosting platform. So this was the question that was posed by the Knight Foundation: how might we leverage libraries as a platform to build more knowledgeable communities? So as I said, we are answering that question by providing free digitization services. And we're coming at the project from completely different angles. So at Metro we have always provided traditional digitization grant programs. So members could apply each year for a pool of money for digitization projects of their archive and library collections. So this project has allowed us to think of digitization in a new way. Often institutions don't even have the time or the resource to put into writing a funding application to get money to digitize. So we just want to take the digitization to them and just streamline the whole process and make it easier. Queens Library. So Queens and Brooklyn are doing the community strand of the project. So Queens Library, they had a community scanning program in place really since about 2011. It's called Queens Memory. And so Culture In Transit has allowed them to kind of just a massive upscale in their productivity with community scanning events and digitization in the local community. And Brooklyn Public Library had nothing at all. Both Queens and Brooklyn have oral history programs. So they are on a regular basis collecting oral histories from community members. And so this project has allowed Brooklyn to create the kind of digitization side, where they already have the oral history side. And so the Empire State Digital Network is New York State's hub for DPLA. And so this is our overview of the work life of the project. So we go out into the community. We gather information from community members. We then host the digitized collections on our individual digital platforms. We feed all the data into Empire State Digital Network, and DPLA then harvest from ESDN. And so eventually someone's family photo from 1950, say from their grandfather's 50th birthday party that they came in to scan, will end up in

DPLA and will hopefully be used and reused in a variety of different ways. So before I jump more into the different strands of the project, I want to first introduce how we do these projects. And this is how we do it. With mobile digitization equipment. So we have three kits. We have the scanning kit, the copy stand kit and the outreach kit. Now the outreach kit is only used in the community scanning events. So the outreach kit comprises of a tablet and – well two tablets actually. One is preloaded with historical images from the archives collection of images from that neighborhood. So for example, say you're going to Flushing in Queens. Before the community scanning event they'll search for all the images from Flushing, Queens from the archive already. And we found that to be a great way to like engage people. That would be just like "Wow, really old photos of Flushing right here with my old photos of Flushing." It was kind of a good bringing together. And similarly, with the other tablet, it's an oral history listening station. So people can listen to other people's oral histories. And really I think it's a way to – a lot of people are kind of like, "You really want to hear my story? Like why am I so important?" So these two methods we found were a really great way at community events to kind of engage people. And to make them try and understand that their photos and their memories are just as important as anyone else's. And so the two other kits – if you're really, really interested in putting together these digitization kits – and as Jamie said, we like to share everything on our blog. We have written reviews for all the equipment that we've purchased and the ways in which we've used the equipment. For example, the copy stand. So you think of a copy stand in a library or an archive. It's going to be a flat board. It's going to have – it's just going to be really big. It's going to be really heavy and you can't dismantle it every time you want to take it somewhere. So we were like, "How can we get around this?" So as you can see, we bought a tripod which can invert. And just a normal DSLR camera and some lights. And so putting the initial kit together wasn't too hard. Because there are scanners on the market. So we're coming at this from an archival preservation standpoint. Our minimum baseline standard for scanning was a 600 DPI TIF. That was just going to be our standard for creating historical masters. And then from that we create JPEG derivatives of everything. So there's a surprisingly small amount of equipment that is both mobile and can produce the standards that we needed. So I can't even begin to describe the amount of time I have spent searching for equipment that could work. And the great thing about this project is we had the money to experiment, like it was fine. We had the money to be like trial and error. But for me like it wasn't – I just couldn't find the equipment. Like the scanners that we settled on are great. We settled on some Epson models. We have the V-600 and the V-800. And we only have the V-800 because Queens Library already had the V-600. So they were like, "Why don't we just get another model?" The V-800 just has a bit more functionality and capability than the V-600. But that comes at a price because it's $500 more expensive and it's significantly heavier than the V-600. So that has its drawbacks in and of itself. So I couldn't find another scanner on the market that I wanted to buy for the project. What you'll find is you'll find $30,000 scanners and copy stands. And then you'll find, you know, light and portable scanners for public consumption that people can use in their homes to digitize – like they're not looking to digitize 600 DPI TIF for example. So I came up against a lot of obstacles. But the equipment that we settled on is great. And the other – so the final deliverable for the project is a toolkit. We're going to basically produce a toolkit this summer that is going to hopefully allow people to replicate our model. So we want someone to just be able to go through the toolkit, say, "I want to do community scanning in a community," anywhere in the world hopefully. And with our documentation and our recommendations and just everything, they can hopefully do this. And so with that, it's also important to think about weight. So the original kit was 66 pounds all in. I modified my institutional kit. I added a different tripod because the lightweight tripod that we initially bought was really flimsy, given that it is lightweight. And I also ordered different lights. We ordered $40 lights. Maybe that wasn't the best decision first of all. But they're fluorescent and they just don't produce the right light for digitization. So I've ended up with a set of $800 lights. You know, the other end of the spectrum. There has to be a happy medium somewhere between $40 and $800. So my modification to my kit has meant that my kit is 86 pounds. So I'm scanning some at the moment and I needed both kits to take with me. So I have to take an Uber. If I can take one kit in the backpack, let me just show you. See? If I can take one pelican case and the backpack which has our laptop and all of our other sundry items, I can take that on the New York subway. And the New York subway is not a nice place. And I survived. So it is mobile. If you want to take both kits, then you're pushing it a bit. I have done it, but I walked to the institution. It was very close to Metro offices. And I survived. But that is a consideration, that if you were going to replicate this model, you would have to have a budget for transportation. So that's been an interesting occurrence through the project. Because I don't think – so when you read the project proposal on the Knight Foundation's website, it says we'll be taking this mobile digitization equipment through New York City subways. And we have. But not 86 pounds, we don't. And so yeah, the blog has been great. Just to talk through our process of buying our equipment and just the challenges that we've come up against. But I just want to jump now into talking about the two project strands. So community scanning. It's undertaken by my colleagues, Maggie and Sarah in Queens and Brooklyn. As I said, Queens had community scanning events prior to this project, but they've really seen an upscale in the number of events that have taken place. And in every location, we're talking 2-3 events at one branch library, and then the same in Brooklyn. They've been held all over. Over the past 10 months it's been close to 50 events in the two areas with over 1,500 unique items digitized from community members. So what do we do? We basically just ask people to bring in family photos, old mementos, old menus from restaurants. Just anything that communicates the history of the neighborhood to them and something that is meaningful to them. So we are not taking the physical objects from them. What we're doing is we're digitizing them, and we give them back. And then we also give them a flash drive of the digital copies. So we ask the donor to sign a consent form and then we work with them to create a metadata form. So that's just basically who's in the photo or what the document is, what year it's from. Just any information that we can use just to give some context and describe the item. And so while the donor waits, it's actually a surprisingly long process. Especially when you have donors who bring in a lot of material. Because this would be fantastic if we could just replicate our equipment five times over, and if we could just have more people at events. We generally try and have a member of staff present that could also take oral histories. So while we're working with one donor to scan their material, another donor might be undertaking an oral history. And then we found that to be quite good work flow. But we will scan the photos or the items to TIF files and then we will convert them to JPEGs and the donor will go away with a thumb drive with all the TIF and the JPEG files. And Maggie at Queens also produced a personal digital archiving brochure, which is fantastic and which donors can take away. And so what we found was a lot of the older generation don't even know what a thumb drive is, so it's meaningless to them. They don't even really know what a TIF file is or a JPEG is. They don't know what to do with the digital copies. You know, they're happy to share their memories and they're proud to tell their stories, but they don't understand what they're getting. So we found the personal and digital archiving brochure is really great to just explain what's on this thumb drive that I'm taking away. What can I do with it? You know, I can share this with my family in Philadelphia now. So that's been a really great addition I think to the project. And then after the event, in the background, we will work through a catalog, describe and upload all the digital copies onto the institution's archive. So this project has been fantastic I think for diversifying the holdings of each institution. A lot of where we undertake community scanning events and the people that come to these types of events are often underrepresented in our institution's archives. So this is a great way I think of diversifying not only the institution's holdings, but the holdings of you know, the wider city and state. Because these items are represented in DPLA, which is fantastic. And so thinking back to the Knight grant's challenge question: how might we leverage libraries as a platform to build more knowledgeable communities? We wanted to think about the community beyond the library. We are operating the project out of libraries, but that doesn't necessarily mean that we have to have our community events in a library. So we explored with different spaces. In Brooklyn, Sarah initiated a school scanning program. So she went in, worked with the teachers. They sent home metadata and consent forms ahead of time for parents to sign. And basically it was just to show and tell. People could come in with their old family photos, show their classmates and then work with Sarah and she would show them how she digitizes. So that was great. In Queens we had an event in a bar, which was great. It was probably – I think it was the most relaxed and the most fun event because of the setting. And people came with stories of the bar. Someone's parents had gotten married in the bar, and so they went home and got photos of their wedding. It was great. And I think having it in other spaces has really helped. And also one of the great success stories of the community scanning side of things has been working with community partners. So community partners have really helped to bridge that gap. We found that events help with our community partners. Sometimes people just don't hear about the event or they don't know anyone else going to the event, so they feel a bit intimidated. But where you have say a friends of the library group or you have a local history society who know members of their community, you know, they can drum up interest. So this was a really successful partnership we had in Queens called My Borrowing, My Borrower. So a local artist had received grant funding to hold a series of public art events. And so we partnered with him and whilst he held public art events, we also trained him to undertake oral histories. And so they were community events for the Filipino American community in Queens. And they were held in two neighborhoods where there's a large Filipino-American presence. And so this community is really underrepresented in Queens' archive collection. So this was a fantastic project. There were three community scanning events and Maggie even went – because of the partnership – went to quite a large Filipino American family to hold a private digitization event. And basically there were a lot of siblings in that family and they all contributed photos, which was great. I think from this partnership we got 16 oral histories and over 100 items digitized, which has completely diversified Queens Library's holdings with the Filipino American community. I just think it goes to show the value in such a simple way of what we're doing. I think the results are really tremendous. So going on now to speak about my strand of the project. So I work – again back to the question. I work at the Metropolitan New York Library Council. So the community in this sense is the Metro member community. So Metro is one of nine regional library councils serving New York State. We serve New York City and Westchester County. We have over 250 member institutions including names that you're going to have heard of, like the Guggenheim Museum of Modern Art, New York Historical Society. But we have a vast percentage of our members who are just small medical libraries, archives, school libraries. You know, we've got the public library systems. But there's a lot of underfunded and under-resourced members that we can help with this project. So our goals for this project were to support the needs of our smaller and underfunded members and pilot small scale digitization for them. Just to see, you know, how it looked compared to the traditional digitization grant program. We also wanted to contribute to DPLA. As I said, ESDN is the New York State service hub for DPLA and it operates out of Metro. And we also last April launched our new digital platform for member digital collections. We had I guess a historical legacy of collections come over from a previous content management system. But with the launch last April we wanted collections to kind of seed our new digital platform. And we also just wanted to find out about our member needs. It's been a great project just to be able to talk to our members about digitization, about what their concerns are, what they're doing, how they feel there's barriers to participation. So what we created was a service model that looks like me going on-site to an institution for up to two weeks. I will digitize a small archive or library collection with my mobile kit. And then off-site work continues with finishing up the metadata. The donor at the community scanning event gets the flash drive with the images on. So the institution gets all of the TIFs and the derivatives and the metadata. So they're free to do with that what they want. We host the collection on Metro's digital site. And then that's harvested by DPLA. So we're focusing on metro members who just don't have the staff to do digitization. They don't have the money, they don't have the time. And a lot of these things are kind of like, you know, with one comes another. They don't have the money so they can't purchase the equipment. They don't have the staff or they don't have the time. Or they don't have the means to host the digitized content online. I think you know, digitization for preservation is excellent, but then I think the added bonus of digitization for access is even better. So that's how we've managed to help them. So these are the 10 institutions that we are going to be working with. I've worked with eight so far. I'm at my ninth at the moment and we have one scheduled for June and we may even try to squeeze one more in before the project ends, just because there's a lot of people out there who are interested in taking advantage of this. So we opened up the project to the membership last year and we just basically said, "We're going to begin this project. Tell us if you've got things that you want digitized." And so we had an overwhelming response. And we had to strike a balance between picking institutions who really needed our help, but also institutions who kind of had some level of organization and metadata with their collections already. And that's purely because of the time constraints of the project. Because I can only be with people for up to two weeks, there's only a finite amount of stuff that I can digitize in those two weeks. Digitization is exhausting and it's time-consuming. And we're not even talking about metadata creation. So having to spend time on metadata creation as well has been, you know, time-consuming. So we kind of settled on these institutions because it all just kind of clicked together for them. So yeah, I've digitized over 1,400 unique items. We have seven collections in Metro's digital culture so far. And it's not just me at Metro, I might hasten to add. A lot of work goes on behind the scenes. So once I am done digitizing and finishing the metadata, I then hand off to Metro's metadata specialist who will then ingest all of the images and the metadata into our digital platform. And then we have the digital services manager who works to get the collection published and up online, what you see. So there's a lot of man hours that go into digitization at Metro for these projects. And we have four collections harvested by DPLA so far, which is fantastic. And it's been really quite a proud moment for us because that was the ultimate goal, you know, of the project, just to get these – we call them hidden collections because they are. Because they're hidden in archives in community members' homes across the city. You know, for only them to know about. But now, you know, someone in Australia searching DPLA's website can find them. Which is the beauty of digitization. So I just want to show you, so this is how the collection shows on our site, on Metro's site. And then in DPLA, what DPLA does is it pulls in the metadata record from our site. So if someone is interested in a specific object, they will click view object and they'll be diverted back to Metro's site. Which is great for us. And so I just want to talk a bit more now about what I've digitized. You know, you can talk forever about the process of digitization, but I think the stuff that you digitize is more interesting sometimes. So I was at LGBT center in December. They are completely underfunded and under-resourced. They have an archivist, but he's a retired archivist. He's there basically out of the goodness of his own heart. He's there only part-time and they really just have no resources to put into any digitization efforts whatsoever. So I digitized 10 years' worth of center newsletter records which were fascinating resource about what was going on in the LGBT community in New York City at that time. And I also digitized these three guides. So the guides, the one on the left and the one on the right, from 1968 and 1969, these are two of the most heavily-used guides in their collection. And so the gay guides contain information about just community information like the social scene. They're often a good indication of where LGBT community members lived at the time. The gay scene guides, the one on the right, contained a lot of information. There was a spate of killings of gay men in August 1969 on the subway. And so that contains a lot of information about safety. And the one in the middle is a center publication from 1989. And that contains a lot of information about the AIDS crisis and what they can do to help prevent AIDS and HIV. So these guides, from a preservation point of view I might just add, the guide on the left is the most heavily-used guide in the collection. And the paper was so thin and it was so worn, and it's the only one that survived. So from a preservation standpoint, fantastic that it was digitized. But from a historic point of view, these guides are just so completely important and relevant culturally and historically to the LGBT community and to the gay rights movement in New York and America and the world. And considered, especially the 1969 guide, with the Stonewall riots in the city at that time, and how that kind of communicated the wider gay rights movement in America. These guides are I guess a no-brainer to digitize. They are fantastic. We've just published them this week on Metro site. So we're really proud to get those up. And I was at Yeshiva University in January this year and I digitized an oversize poster collection. So it was 177 posters from the Soviet Jury movement. And these, as you can see, the poster on the left, this is pretty indicative of what the whole collection was like. Just a series of rolled-up posters that were torn. So how do we digitize them? So the one stumbling block and the one thing I wish I had in our digitization kits that we don't is a large piece of Plexiglass or optically pure glass to flatten material. So what you often see in digitization labs, you know, in an archive. We don't have that because I can't transport it. I haven't really yet thought of a way to easily transport that. So what we have instead is I brought archival weight bags and archival book tape with which to flatten down either end. And I'm going to say it's not as perfect as glass would be to flatten, but it's pretty good. You know, we digitize these posters that otherwise would just – they weren't even accessible to researchers. The archivists there said they just couldn't even allow researchers to use them because they had to provide constant supervision. And again, they're under-resourced and underfunded. So they just couldn't really provide that information to the researchers. So this was a really great institution to partner with on the project, just to provide access to something, you know, that would otherwise just be in a storage room, not getting used. And we've actually seen – I started digitizing last June and we've already started seeing examples of how the digitized content is being used. So I digitized at White Plains Public Library in September. And this year they're celebrating the 100-year anniversary of their incorporation as a city. So they're holding a series of events to celebrate that occasion. And so they've used the images that I digitized in community trivia night and also in a research workshop. So people were able to go along and research the history of a building or the history of a street or, you know, where their family grew up in White Plains through the content that I digitized. Which was great to see. And the first institution we worked with was the Wildlife Conservation Society. And the design department at WCS has been very interested in the series of pamphlets sand brochures that I digitized of the zoo exhibits. And so you can just see on the right there, I think last week was the anniversary of the opening of the African Plains exhibit. And so that was a brochure that I digitized. And so that's been great to see that design department want to reuse images or have drawn inspiration from the digitized stuff. Because they didn't even know it existed before. Even though they know the archivist, they know they have an archive. You know, until they see the physical thing like, "Oh, we have those. That would be great to reuse." So that's been really, you know, heartening to see that it's already being used and the project isn't even over yet. But I just want to move away slightly from the projects and just spend a few minutes talking about digitization. Why do we do it? I guess I was taught when I was training to be an archivist, you know, two things: access and preservation. I guess if you talk to any archivist, any librarian, you know, any cultural heritage professional, they're going to say "This is why we digitize." So access. I just want to talk about access in the context of a previous project that I worked on. So I was the archivist at the Natural History Museum in London, and I worked on a multi-year initiative to digitize the correspondence of a naturalist, Alfred Russel Wallace. So he comes up in conversation, no one really knows who he is. He actually co-discovered the theory of evolution by natural selection at the same time as Charles Darwin. Charles Darwin always gets the credit because a year after in 1859 he published The Origin of Species. Takes all the credit. Wallace was a fascinating individual, okay? He was a naturalist but he was also a socialist. He basically just got himself into all sorts. And he was really intelligent and he was a conversationalist. And we know that there are around 5,000 letters that survived to and from him in about 150 institutions across the world. So the aim of the project was not just to digitize – so in the UK you've got the British Library, Natural History Museum and the Royal Botanic Gardens at Kew who hold the majority of Wallace's UK correspondence. But because Wallace spent time in South America and in southeast Asia, he also in the 1870's basically sailed around the world and did a lecture tour. So there's just correspondence from him all over the world. So we wanted to scan and transcribe and put online every known piece of correspondence. And this is what this is. Wallace Letters Online is that archive of letters. And so the fantastic thing about Wallace Letters Online is you can read entire conversation chains. It's just like you can read letters back and forth. We have the complete set of correspondence between him and Charles Darwin, which is just fascinating.

Because they're each going, "Well, no, you're the better scientist." "No, I think you're much better than me." It's just fascinating. And you know, it was tough. It was tough to locate all of it. And I'm sure we haven't even – you know, we knew there were 5,000 letters. You know, what about all the other letters that we don't know about? And that's the problem with archives. It's all the stuff that we don't know about. But it's just so valuable to researchers. And the other side of the coin, preservation – and I'm just going to preface this by saying I know that I've picked the most dramatic headlines. And this isn't an everyday occurrence thankfully, but this stuff happens, okay? Like paper is not indestructible. There will be fires. There will be floods. After Hurricane Sandy happened, a woman came in from a neighborhood that was completely flooded and devastated by Sandy. Came into the Queens Memory Project and said, "Please digitize all my family photos." She lost a portion of her family photos in Sandy. But she's like, "Up until that point, I never even realized you know, this could happen. And it happens." But I am fairly aware that this is a worst-case scenario. And when we talk about preservation, digitization for preservation, we should be thinking you know with disaster in mind. But I think it's also in terms of the fact that paper isn't indestructible. And this is kind of, "I'm going to go back to the Natural History Museum here and Alfred Russel Wallace. The digitization of his archive and his correspondence I think shows how digitization can be mutually beneficial. It can provide access, which is just – I could talk about it for hours. I'm so passionate about digitization for access. But it can also provide digitization as a means of preserving. We shouldn't use digital surrogates as, you know – like we should always keep the original. But we can use digital surrogates in place of it. So Alfred Russel Wallace's notebooks. He spent eight years travelling southeast Asia. And he recorded every single bug that he encountered and collected. Every single local community that he visited. All of the tribes people that he encountered. So people could take these notebooks that he wrote in the 1850's, go to these communities now and basically see what's changed. And a lot will have changed because he went to places. He went to Borneo. He went to Siloazy. He went to all of these countries that are now being affected with sugar plantations and rubber plantations. And you know, deforestation. And he went to the Amazon. You know, and he basically recorded everything that happened in the Amazon in notebooks as well. But unfortunately, on his way back to the UK from South America, his ship caught on fire and all of his records were destroyed. Again, disaster. But he couldn't digitize in the 1840's. So it just goes to show I think. So these notebooks were some of the most heavily-used items in the Wallace collection at the Natural History Museum. And as a result, they were being badly degraded. I mean, you can see the front cover. But the oil from people's fingers was rubbing away the pencil marks. And when you use gloves to open the notebooks, it was damaging the paper, so you couldn't win. So digitization as a preservation tool has been really, really, really useful I think with these notebooks. And just a note on – I spent a large portion of my career working with history of science archives. And the volume of records that was produced in the 19th century by men exploring – I'm going to say women as well because there was a really amazing artist at Kew who basically travelled the world and painted all the flora and fauna of the world. But these explorers who went out and collected botanical material. They collected animal specimens. They observed cultures and they observed tribes people. The amount of correspondence and records they sent back is humongous. And the problem is that they were often on the road for years at a time. Like Wallace spent eight years in southeast Asia. He took cheap paper and cheap notebooks with him. And as a result, they just degrade. Like they are so acidic, you know. The iron ink that they used eats away at the paper. You know, a lot of these record aren't even here anymore just because of the natural structure of the way the paper was made has just – you know, it's degraded. And so digitization, you know, has provided a way for all of this to be opened up. And without the need for researchers now to go in and view the physical objects and you know, handle them more, which is going to destroy them more. And you know, once they're gone, they're gone. So really I think the access and preservation standpoint for digitization is you know, a perfect combination. So that being said, digitization is great. I think we can all agree that. Why isn't everyone doing it? I think now we come down to the crux of it. We have those who can and those who can't. And again, it comes down to money basically. Money for equipment, money for staff. And I think working on the culture and transit project has just really highlighted to me the absolutely massive divide there is, you know, in people who just can't digitize. It's not that they don't want to. It's just that they can't. And those that do. So for example, just to give you an example of how expensive digitization can be, this is just one scientist's archive. The American Institute of Physics had all the papers for Samuel Goldsmith who was – I'm not going to go into Samuel Goldsmith. But let's just say he was an important scientist. But it took two years and $120,000. That was external funded dollars, to digitize 67,000 pages. It's fantastic that they've digitized it. But you know, it's money and it's time. Similarly, the New York Philharmonic archives – I love this project. They're basically digitizing their whole archive and putting it online. But it began in 2007. It's not going to be complete at least until 2018. It's been funded externally by two different organizations. You know, it's going to take more than 10 years to digitize one archive and countless dollars to digitize. And I think this is the crux of it. I think these two examples are extremely indicative of the landscape of digitization. These projects, ours included, Culture in Transit, is externally funded. But I don't think it's sustainable. In 2005 when the UK won the chance to host the 2012 Olympics, fantastic. Great. I had an amazing time. But they diverted funding to the Olympics from the Heritage Lottery Fund. And the Heritage Lottery Fund was basically one of the key ways that libraries and archives in the UK get money to digitize. Just completely gone for like six years. Just no money. And everyone's just like,"What do we do now?" Like, you don't think – it's almost like funding circles come around. You know, like the NIH. Their funding circles will come along annually. You'll think, "Great, I can apply for that next year." But then it's gone and then you can't digitize. So I think that really shook up the sector in the UK. And we were all actually really pleased when the Olympics was over and they started diverting more funds into digitization again. But you'll see again and again it comes down to money. And I just want to kind of bring it round and just finish up with an example of where I'm digitizing at the moment. So I'm digitizing at the General Society for Mechanics and Tradesmen. And it is two blocks away from the New York Public Library. And I love the New York Public Library. I think it's fantastic. They have – the NYPL labs is phenomenal. I went on a visit there, and the amount of equipment they have and what they're doing with digitization is fantastic. And the way that they are using the digitized content in different tools, in different apps, as a way to educate people is amazing. They have over half a million items online. The general society have zero. And I just think that it's so ironic that institutions that can be two blocks away from each other in New York City can just be so far removed from each other in terms of digitization. It's just crazy. Like NYPO has digitization as part of their core digital library operations offering. I think so many institutions are so far removed from having digitization as a core collection management policy. It's crazy. But the general society has a long and illustrious history. They are the third oldest organization in New York City. They were incorporated in 1785 and their library, which was founded in 1820, is the second oldest in the city. And it's also one of the first public libraries in the city. So they were rebranded as the Mechanics Institute in 1858. And they continue to provide to this day – they're the oldest privately endowed tuition-free school in the city providing trades-related education. And their archive is fascinating. I have digitized some glass plate negatives of the construction of the 59th Street bridge. And it is just phenomenal. And they're just sitting in their archive. Again, because they have a part-time archivist who's in two days a week. He spends the majority of his time replying to collection research queries that he gets via email. There's no time, there's no money, there's no budget for digitization. But their archive is just a treasure trove. And I think that's basically what we've tried to do with this project, is just to open up digitization to everyone. I don't think anyone should be discriminated against not being able to digitize. And I think with this project – well I guess I hope we've tried to show that digitization doesn't have to cost a lot of money. Yes, it still takes time, but I think we've added a lot of value in a very simple way I think is what I'm trying to say. And hopefully underrepresented communities, underrepresented archives, underrepresented museums and libraries can now hopefully have their voices heard and have their histories heard on a wider platform. And I mean, digitization for the win. That's what I say. Thank you for listening to me ramble on.

[Applause]

>> Well you for the win. It's 3:00 exactly.

>> Carline Catchpole: Woo.

>> That's it. So we have a break. Stretch your legs. Eat food if there's some left out there. I think there's a lot. And then when we come back we're going to do the panel and we'll ask the burning question that you may have been keeping

inside for the past couple of hours. So 320 is back here.