MUSEUM UNIVERSE DATA FILE WEBINAR
JUNE 11, 2014, 3:00 P.M. CST

Hi, everybody. We are going to get started in a few minutes. (pause).

>> The conference is now in silent mode. (pause).

>> Recording started.
(silence).
(standing by for audio).
(sorry, if someone is speaking, I can't hear them).

>> Can everybody hear me now?

>> Okay. I think we worked out the audio issues. I'm going to do a quick recap, if you didn't hear me before. This is the Museum Universe Data File webinar.
My name is Justin Grimes, with the Office of Planning and Research Evaluation at the Institute of Museum and Library Services. We are going to talk about the Museum Universe Data File. It includes what is the file, how it is created, what museums are included, how can the data be used, and a little bit about what is in the data file and what is next and what you can do to participate.

As I was saying earlier, what is a Museum Universe Data File? It is a list of known museums in the United States actively maintained by the Institute of Museum and Library Services to support our agency's mission, which is to support libraries and museums in the United States. This list is not just a count of museums. It is more than that. It is a public resource that we provide to the community for research purposes, for scholarship and for, as a tool for innovation. We talked about what the Museum Universe Data File is.

Let's talk about what is included in the Museum Universe Data File. The Museum Universe Data File, the way we constructed this is it's an aggregation of several data sources that we, administrative data, public data sources that we pull together. What is in this? This is, we pulled sources from IRS, nonprofit data, this is the Internal Revenue Services' 990 forms which requires nonprofits to file. Private information includes information from museums from our own administrative data. This includes grant applicants that have applied for funding to the agency.

This also includes private foundation grantees. This is information from the foundation center, which is a source of data that provides information about private foundation grant-making that is over $10,000. It is the top 1,000 private foundations in the United States. It includes information from third-party sources. This is crowd source information, web-based information from primarily a company called Factual, that scours the web and aggregates information.

From these four lists, we pulled together any records that occur between 2009 and 2012 of the data sources that were self-identified as museum.

Below there, you can see a table that shows the breakout of museum disciplines for the museums that are in the file. You can see a count, percentage of art museum, and aquarium, historical societies, history museums. That gives you a feel for the types of museums that are included.

I mentioned earlier that we pulled information from several sources, four sources, the four sources together. It is not that we just merely took these four sources, put them together and re-released the file. We did several steps to try and remove possible duplicates. We tried to reconcile the records. How we did that was for each of the data sources, we attempted to identify common identifiers. And this could be names, this could be geographic information like address, this could be a more stable identifier such as EINs, which are employer identification numbers, which is used by the IRS.

What we did was, when we aggregate these four sources together, we looked to find these common identifiers to connect information together, to remove duplicates and to append additional information when possible. Through that process, we ended up with the 35,144 museums that are in the file that you see.

I'll give you an example. On the right-hand side of the screen, you can see where it says data were combined using common identifiers and duplicated were removed. A good example of this might be the Bird museum.

Say the Bird museum, which is not a real museum, the Bird museum in Toledo, Ohio, so the Bird museum could be one record in the Museum Universe Data File. There could have been another from a different source such as Museum of Birds also in Toledo, Ohio. It is clear these two sources of information are the same.

What we did was merged those two. We reconciled those two records together to prevent double counting. The process was largely automated. We use algorithms at the agency here to do most of this process automatically.

When there was a degree of uncertainty, we had human staff, human staff trained with very specific protocol on how to reconcile those records together. We feel fairly confident in the work both through the automated process and through the human verification that we did our best effort to reconcile and remove possible duplicates in the file.

That is a little technical on how we constructed the file. Now you might want to know, what is in the file itself?

What we have in the Museum Universe Data File is first basic identified information. This is the name of the organization, this includes address information, city, state, zip, for some of the record we have phone information. We also have web URLs. If they have a website, and that information was made available, we included that as well.

You also see that I have alternate name up on the slide. This was when we were in the process of reconciling the records, there was some cases such as the example I gave earlier, the Bird Museum versus Museum of Birds. When we reconciled those records, we preserved the alternate names on the field so that to improve the searchability and findability of the information in the file. Other information that is included is administrative and financial information. As I mentioned earlier, one of our primary sources was IRS 990 data, as part of this submission process for tax-exempt status. Museum nonprofit organizations are required to submit their income and revenue. We have appended to each file where we pulled in IRS 990 data, which is a large percentage, we have their most recent income and revenue that was associated with their most recent IRS 990 tax form.

Additional information that we also have from the IRS 990 form is something known as the national taxonomy of exempt entities core classification code.

IRS classifies all nonprofit museums according to a specific classification, very specific taxonomy. This includes not just museums, but includes education resources, health. It is quite expansive. This is one of the ways that when we said earlier of organizations within the

data sources that self-identify, that were identified as museum, the national taxonomy of exempt entity core classification was one of those ways that we identified museums. That information is preserved in the file as well, and income, revenue. Also we provide in the data file geographic information.

Part of our process was during the post editing process, we ran every record through geocoder. When there was not information associated, we put it through the geocoding process. By doing that, now we have specific geographic information such as lat, long, census tract and block. We also have the American Alliance of Museum region codes as also associated with the file.

We have something known as the NCES urban centric locale codes. This is a methodology used by the Department of Education to provide location information on a rural, urban continuum. This would put every location that is in a city, in a town, rural area.

We also, so that's basically the information that you can find in the file. We also have one thing that is not on the slides that I mentioned earlier, which is flags that are associated with the data sources.

So for every source of data, you can see where the information came from, and how many times that entity appeared in the different data source, which is a useful tool if you are interested in trying to determine additional levels of verification and scrutiny, so if it appears in multiple sources, that provide more confidence in the information that we are providing in the file.

That tells you a little bit about what the Museum Universe Data File is, how it was created, and what type of information are in the Museum Universe Data File.

What does that mean for us? What are the next steps? For us, we have committed as an agency to maintain this list on an ongoing basis. We use this to support our agency's mission. We are committed to produce this on a semiannual basis. We will have another upcoming list later this year and going forward. This was May, 2014.

With this, as you can imagine, it's difficult to construct a list of known museums in the United States. So the process as we go forward, we are going to continue to improve and tweak. And so we are always looking for feedback, always looking to improve our methodology, our techniques for making this data available. Every iteration is going to get better and primarily going to get better based on community feedback.

Through this process what we are looking for on the next release, which will be later this year in the fall, is, as I mentioned earlier, one of the primary purposes of the Museum Universe Data File is for research and scholarship. We will, we at the agency, will be using this as a universe file for to construct a sampling frame for upcoming museums count survey, which will be deployed in the field, and provide additional information that can't be gathered through this type of aggregation of data sources that we are currently doing. There is limitations to this process, because we are limited by the source of information we can find.

This museum count survey is greatly important because it can provide us additional information about collections, staff, additional details about financial information, to better help us understand and help the community understand the current state of the museum landscape in the United States.

While we are doing the parallel tracking museum count survey, we will continue the work on the Museum Universe Data File. We will be releasing research briefs, reports, and data products for almost on a continual basis.

As I said earlier, we are going to be doing data products and research. We already have

some available on our website. You can find this on the IMLS.gov website. We provide descriptive information about some of what you might find. This includes a quick distribution of museums in the United States by state. That would be what is on the right. On the left, what you are seeing right now is a core plus map that shows the number of museums by state per capita for 100,000 population.

You can find all this on our website. We will be releasing more descriptive analysis, more descriptive research in the coming months.

Other than research and scholarship, there is another thing that I mentioned that you might also see which is data products. We are not by all means limited to, as I said earlier, this is a public resource. This is a community resource for other people. This is not just limited to research and scholarship. This is also related to innovation.

You can imagine a lot of people could find this as a tool for application development, software development. You can imagine one taking this resource, this data, and making, say, a mobile museum finder which can help people, assist people to find museums in their neighborhood, or different kind of spatial or geographic application. Here is one that we are working on here at the Institute of Museum and Library Services. You can give me one second to set this up.

Okay. Hopefully, if everything goes right, you can see a map. It should be displaying the Washington, D.C. area. You are seeing the points. You are seeing a map of the United States. That rainbow that you are seeing is every museum in the Museum Universe Data File. This is 35,144 museums.

Now, what you are looking at now, the reason why it's color coded is it's color coded by discipline. You will see the historical sites are blue. Art museums are green. Arboretums are red. This is the tool that we are hoping to make available in the coming months. It is basically a nice interactive map visualization, where you can type in either a museum or a geographic area. I'm in Washington, D.C. so I'll type in Washington, D.C.

It will take me down to Washington, D.C.

So you can see all the museums in your neighborhood or area. Let me get a little closer here. Washington, D.C. have a lot. We are basically the museum capital of the world. One of the things here, you can click on this, let's find, where is -- Patrons of the Arts in the Vatican museums. What you will be able to do is click on any of the points, pull up information, so this will provide the name, address, basic information from, about the organization. If the phone and web address is available, it will provide that as well.

This allows to you click and explore to see museums in their neighborhood. You also, as we progress forward, you will see other additional layers of information pulled in, possibly census measures of poverty or other quality of life or economic variables pulled in. This will be hopefully available in the next couple months.

This is just one example of some of the things that could be done with the Museum Universe Data File.

Let me switch back. One second.

Now that I talked about some of the things that we are working on, here is how you can get involved. This data is already publicly available on our website. The documentation is available on our website. You are free to start digging in, looking into it.

Like I said, this is an ongoing process. We are always looking for feedback. One of the things we are asking everybody to do, specifically those in the museum sector or the public as well, open the file up, look to find museums in your neighborhood. If you see something

wrong, if you see something, say something, suggest it to us. You can E-mail us at research@imls.gov. This can be, as I said, we are pulling from public data sources. There could be misspellings, typos. And some of the data source points, if you see any kind of information that you would like to add, please submit to us. If you know a museum is closed in your neighborhood, feel free to tell us. If there is a new museum in your neighborhood, send that in as well.

I'd also suggest, not just for checking resources, but feel free to open up the file. If you have a programming background, load it in some applications, do some visualization. Feel free to send those to us. We like to see any work being done with our resources.

So, wrapping up, I'd like to point out three links for you. The first link is a link to our website, where the data files are available. They are available in CSC Excel format. There is also, we provide data documentation. If you want to know more about the technical process for how we constructed the file or any limitations, feel free to look at the documentation. If you have questions, feel free to E-mail research@imls.gov. Also we have on our website an FAQ, questions and answers, that might help in terms of answering any of the questions you may have about the Museum Universe Data File.

Now that I've gone through the presentation, we can do questions and answers. Feel free to type in any of the questions in the chat room, and I'll read them back out to you and answer them.

Okay.

>> The first question we have is, do you realize that the list is inflated? There are still duplicate listings for some museums, plus other listings that are clearly not museums; example, neighborhood associations, foundations, friends groups and arts and crafts associations.

>> Hi, this is Carlos, the research director at Institute of Museum and Library Services. As Justin said earlier, we are using administrative data sources, multiple administrative data sources. And we have combined those.

We do recognize that on some occasions, there is both a museum listed and a friends of listed. Those are separate. When you see those, that is because they have separate 501C3s. We did not go through a process, for example, of determining whether or not the friends of the friends of historical association was itself, had a physical historical site attached to it. That is a more detailed process than we were going through with the Museum Universe Data File. There are type one errors and type two errors in this file.

There are some records that exist that may not meet your definition of a museum. But there are also things that we have noticed that some people have told us that we didn't capture because we didn't use a particular NTE category within the IRS.

As we mentioned, this is a process. And the process will shift and improve over time. I think it is important to say that this is a stake in the ground. And as Justin said, this is a statement about the vitality of this particular portion of the arts and culture sector, of the cultural sector, and the STEM sector.

And so, in other words, it was important to put a stake in the ground, and to improve moving forward.

>> The next question I have is: Why did you not consult with state museum associations

prior to publishing this data?

>> In fact, we did consult with state museum associations, about a year ago. We developed a tool for verification of data that we had. People's time is precious. There really wasn't much participation in terms of verifying the institutions. We understand also, people, we didn't want to burden folks too much. So we moved forward in the most efficient way we thought possible.

>> The next question we have is, on what basis will the sub sample of organizations be selected for the museum count survey? Will it be likely that meaningful but small categories of museums such as contemporary art museums or culturally specific museums or outdoor history museums would not be sufficiently represented for those constituencies to have a useful profile for, of their type?

>> The important thing about having a universe data file is that you can sample from that file in many different ways. So, for a national representative survey of museums, I can say that it's likely that very small subgroups may not be represented.

But that does not preclude in any way from doing a national representative study of small art museums, or a nationally representative study of nature centers. This is a foundation on which one can build. And one can select sub samples in many different ways depending upon your interests.

>> The next question I have is: Are there plans to put [inaudible] or pull requests?

>> This is Justin Grimes again. You are not the first person to suggest this. This is one of the things that we are actively looking at, and we will let you know when it is up. I think a version of this is already up on get hub. The folks that, hackers from Iowa posted a version of get hub.

>> I think it may.

>> JUSTIN GRIMES: The great thing about public data is that many people can grab the file and manipulate it and post it in many different ways. I know our plans are to not only post files in different formats, but also to have an API for the file for developers so that they can pull from the back end.

>> The next question I have is: Is the IRS form 990 data available in more detail?

>> JUSTIN GRIMES: Absolutely. We are not the source. I think perhaps the best source for that data, well, there are two main sources. One is the IRS itself. Department of Treasury makes that information available. But also the National Center for Charitable Statistics, which is part of the Urban Institute in Washington, makes that data available. They collect from the IRS, and then make data files available.

And then the third site, which I actually almost forgot to mention, is guide star. Guidestar.org, one can look at individual 990s, the full document. So by the way, one of the things that we retained on the data file was the EIN. And the part of the reason why we retained that is because we want to do financial analysis of the museums sector using the IRS numbers. So that we can also impute and develop estimates for those entities that we

don't have certain financial information for.

>> The next question we have is: Will your office verify changes?

>> JUSTIN GRIMES: Yes, but I guess if you can throw in a clarification, certainly when changes are recommended, we are going through a process here. If for example, somebody suggests that an institution is closed, we will obviously go through a process here to determine whether or not we can verify that an institution is closed, or if an institution is open, for example.

>> The next question I have is: Should we remove friends groups?

>> JUSTIN GRIMES: Friends groups sometimes operate historical sites. So that is part of the rationale for keeping in friends groups. This is true also in the library sector, is that they may be 501C3s that run in parallel and can also do programming.

So we were, we included, this was a consideration. We included them because we did not want to exclude friends groups that were actively running historical sites, for example.

>> The next question is: What is the process to update changes? How long will it take to update an entry?

>> JUSTIN GRIMES: Well, we cannot promise for resource reasons realtime updates. We have promised semiannual updates. So we can certainly take your recommendations on a rolling basis, but in terms of updating a file, we have committed to doing semiannual updates of the file.

>> Next question, could you talk a little bit more about -- could you talk about the back end APIs tools under way?

>> JUSTIN GRIMES: Sure. This is another thing that, much like get hub, that we are looking to proceed forward to making API and developer tools available. We are in the process of, we have to investigate what would be the best option for us. But that is something that is highly under consideration and something we will keep you posted for. We have on our website, I think it's the IOS/developer, we have a developer page. Definitely bookmark that. We will make more information available when it is.

>> Yeah, which is actually part of a broader open gov strategy, not just that IMLS is undertaking but federal agencies are undertaking to make statistical data available and also administrative data available in different formats and with APIs.

>> JUSTIN GRIMES: We are getting ready to wrap up. Does anybody have any more last-minute questions before we end? Feel free to submit them in chat.

>> Do you have a recommended date to submit the first wave of edits?

>> JUSTIN GRIMES: At any time. As I said, we are happy to accept comments,

suggestions on a rolling basis.

Okay. Thank you very much for -- you have another question? Thank you very much for participating. If you have any questions, feel free to submit an E-mail to research@imls.gov. We will be happy to take feedback questions or any suggestions.

>> We will quickly have one more question. How about umbrella parent organizations that are listed in addition to their satellite sites? Should those be removed?

>> JUSTIN GRIMES: Well, the organizational structure, as you all know, is very complex. There are parent sites that run an institution, an organization. And with satellites that also can function as their own independent sites. And there are parents that are administrative organizations.

We have defaulted to leaving in parent sites for that reason, and moving forward. If you have identified a parent site that is little more than administrative office, please let us know. We will look into it and adjust the file accordingly.

>> We have another question. What are the uses of the database? If for federal statistical analysis, shouldn't the list be cleaner first?

>> This is not a federal statistical file. That is a very important distinction. Thank you for asking.

However, there are many, many agencies that make lists available. This is a first step to getting towards a cleaner file that we are drawing a sample from for a statistical survey.

So thank you for that question. It's an important distinction.

>> The next question is, would you consider updating the database more than semiannually in year one to clean it up more quickly?

>> It's a pretty resource-intensive process that we have gone through. I can commit to updating right now semiannually. Frankly, as a business flow, my preference would be to have something closer to realtime. But that is really not something I can commit to faithfully to this group now.

>> We have a question related to one that was asked earlier. Can parents and child relationship be included in the data?

>> JUSTIN GRIMES: I would say probably not at this moment. This is something that we had discussions about. I think to maintain that relationship between an organization and the entity associated with that organization, so for example, if you are a naval museum in Hawaii, there might be reference to the naval museum itself, and every battleship will have a reference in the field, in the data file itself, but there won't be any way to capture the relationship between those two.

It is one of the things that we are looking for in the upcoming museum survey which will allow us to get more structural information to better connect those relationships.

>> Yeah. I'll add something as well here. This is Carlos again.

Now, there are different reasons to have organizations separated out individually, even if they may be related administratively. So for the purposes of way finding, if you are putting every institution in a map so that individuals can find nonmuseum folks, for example, can find an institution, then that is a good reason to put every single entity regardless of the administrative relationship between them. That is a good reason to put every entity in there, so that people can identify that physical resource.

Now, the governance structure and how they are connected to each other, that is an important question, obviously, for the field. It's an important question for IMLS to understand the characteristics of the field. And so when we are doing an analysis of the organizational structure, the financial conditions, those kinds of issues are particularly important. So for different, understanding those relations are just, are important for different reasons and different purposes.

>> I want to ask one more time before we wrap up if there is any additional questions. Feel free to ask now. You can always send us an E-mail at research@imls.gov.

Thank you very much for your time.

>> Looks like another question came through. When we build things, is there a way to share what we have done with the community?

>> That is excellent. I'm assuming you are talking about building applications, presentations or visualizations. IMLS is working through a possible get hub site. There would be a get hub site where some of that information could be posted.

>> Please, by all means send that to us.

We would love to showcase and highlight the work that people are doing with Museum Universe Data File. I can see us doing a post to draw attention to the work, fantastic work you might do. Please, yes, send that to research@imls.gov as well.

Okay, I think that is it. Thank you very much. As I said earlier, I appreciate you taking your time out to listen, and if you have any questions, feel free to send an E-mail to research@imls.gov.

>> Recording stopped.

(end of webinar at 3:42 p.m. CST)