OCTOBER 30, 2014
1:00 P.M. CST
IMLS MUDF WEBINAR

BUILDING FROM THE GROUND UP: A LONG-TERM APPROACH TO MUSEUM DATA COLLECTION

>> Recording started. Good afternoon, everyone. This is Susan, director of the Institute of Museum and Library Services. We are excited to welcome you to this webinar, talking about the Museum Universe Data File.

We hope that we can share some good information with you and we anticipate lots of questions. We would love to see who is in our audience. It would be terrific if you could in the chat box let us know what organization you represent, what state you are from. We would love to know who all we are talking to. Thanks for joining us today. I'm going to turn it over to our director of research, Carlos Manjarrez.

>> CARLOS MANJARREZ: Hi there. Thanks for taking the time to join us today and to hear a little bit about what we are doing with museum data collection, and also to share your thoughts and provide input, and help us shape the future when it comes to data collection.

I'm going to start out talking a bit broadly about data collection. I'll hone in on the Museum Universe Data File itself, talk about some uses of the data, and what our future plans, but I want to let you all know that I've tried to limit this to about 25 minutes, to leave plenty of room for questions and comments. Keep that in mind. Please feel free to use the chat box, if you have some questions. We have somebody here who's looking at the chat box and they will be, I'm going to move forward with the presentation, but somebody will be logging those questions and we will try to address them at the end of the presentation.

With that in mind, I want to move forward. I want to talk to you a bit about the data collection that IMLS is doing. It is a relatively new function at IMLS, gathering data on library and museum services. What I've done here is basically presented a number of data sources that we have available now. Above the line are two data collection, two major data collection efforts that we have adopted in a sense from the National Center for Education Statistics.

That is a public library survey which is an annual survey, it is a census of public libraries across the country, and the state library agencies survey which was annual -- it switched now to every other year -- that gathers data from every state library agency in the country.

Those were adopted collections, if you will. They were ongoing. The public library survey has been going on for 20 years. Below the line are new data collection efforts that were initiated by IMLS, or data that we have leveraged to make available. The IMLS grants database, available in some form already on our website, is now available as a data file, and the Museum Universe Data File which we will be talking about mostly during this presentation. I want to alert people to a data file that will be released later in the year, the public needs for library and museum services, the national household survey. You will see on the right-hand side I've indicated which of those are data collections about institutions, about public libraries, state libraries or museums, and which of those are data collections about individuals.

It is really the bottom one that gathers information from people, one of those pesky home surveys where we call you during dinnertime and we ask people a series of questions about their habits.

These are the data collection efforts that we have been engaged in and that we are making available. We have made them available primarily in a form that researchers can ingest and use. They are flat files, data files that people can use with statistical programs, with GIS programs, and to analyze and produce reports.

But we recognize that that is really not the norm, that there are many folks out there who want to grab and manipulate data and use information for their own program services but may not have a statistician on staff or may not have access to GIS tools. With that in mind, we are building a set of tools for the every day user. This is an image of our beta site, that makes data available and accessible to people just right over the web, with a web browser and a mouse you can create charts and graphs using this data tool. You select the data file that you want to manipulate. You determine the kinds of maps or the charts or graphs that you want to make. You can filter the data. Then it also has APIs for all of these data sets on the back end. So for example, people who are developing mobile apps and want to connect to IMLS's data, particularly location data of libraries and museums, they can connect to the APIs here, draw in that data for their mobile apps.

And so that makes, again, that makes the data more accessible, beyond access just to researchers.

Let's hone in a little about the Museum Universe Data File. I wanted to identify some of the sources for the Museum Universe Data File. We have used federal administrative data. That's our own data from our grants database, but also data from treasury.

So every year, nonprofit entities are obliged to submit nonprofit tax forms, also known as IRS 990 forms. Whenever entities that are designated within treasury as museums have filed tax forms, we have gathered that data. That gives us an indication of their size, based on operating budget. It gives us obviously some location information. It gives us an indication of the type. So we have gathered that data. We have gathered data from private foundations. And this we have gathered from an organization called the Foundation Center. The Foundation Center is basically an aggregator of information from the 1,000 largest foundations across the country. And those, they basically report every grant above $10,000 that they have made and they do that on an annual basis, mostly calendar year.

We have gathered ten years of Foundation Center data. We have also gathered data from social media aggregators, specifically Factual. Factual is a company that sells, that aggregates data and sells it, makes it available to groups like Yelp. Yelp is a purchaser of Factual data. 4 Square is a purchaser of Factual data. It is a interesting thing about these data aggregators. It is a cyclical relationship or symbiotic you might say.

While Factual sells information to Yelp and Four Square, it ingests information from those sources as well, so that when there is an update of a location, for example, in Yelp or Factual or 4 Square, that update goes back to Factual and they update their records.

So part of the reason why we wanted to go with this source as well is because we know that clearly not all museums are nonprofit entities. So they won't all be located within the IRS administrative data.

So we wanted to try and capture more entities that may be public entities, either underneath a university, for example, or part of a broader system like the National Park Service, or even at the county level. So that was the motivation for getting data from that source.

Why would anybody go through the trouble of gathering information from all of these different sources and trying to mash it up into one big file?

Well, as the name would imply, we are trying to define a universe. I want to be clear about the language here. This is really a very broad net that we are casting. And it's quite intentional. We want a very broad net because we need to go below that level, cleaning the files so that we can develop a museum sample, a sector sample, sampling frame. And then from there, we draw a sample for future surveys, whether that's museum count or other sample surveys.

This, the Museum Universe Data File I really want to point out is part of a broader process of gathering information, casting that broad net, reviewing that information, cleaning, adding, supplementing and then drawing samples from it.

It is basically part of the research process, and one that was really necessary in this case, because there is no single authoritative source that comprehensibly lists all museums in the United States. We looked.

This is just a graphic image of that process, and what one of the uses are for the file.

What happens, what happened after we released that initial file, well, we got a lot of feedback. We have got feedback from over 300 people, providing us input on the file, letting us know if an address location was off, if the name was wrong, letting us know that their institution wasn't represented.

And you know, I want to alert people. When you are gathering information from so many different sources, when you are trying to build one resource out of many different resources that frankly didn't really speak to each other before, it's a messy process. It's a necessary one. It is one that we take quite seriously, so seriously that we committed to making multiple releases over time, in a fairly aggressive data release schedule.

If you look across agencies of the federal government, it's quite rare for agencies to establish a 6-month data release regimen. But we wanted to do so because we know that data, making data available in shorter time increments is really important, not just for the museum field but for the public in general.

We really look forward down the road. We would love to be in a situation where we are actually making realtime data available.

That's particularly, that is quite far down the road, unfortunately. But that's the goal we are setting ourselves up for.

So the data cleaning, it's something that we started with cooperators at Drexel University. We have a contingent of data processors here at IMLS. Right now, the focus of this round before the December release is identifying operating addresses. Part of the reason why this is an issue is because for many of the IRS, for a lot of IRS data, the information that we have in there is a P.O. Box, or what we have found in some cases, it's actually the address of the person that's filling out the tax form.

So there are a number of entities, for example, across Oklahoma and Texas, that are using the same accountant. And that accountant's address has been put in about 50 different entities in their 990 forms. These are some of the anomalies, some of the issues that we are finding.

We are identifying phone numbers, we are identifying URLs, and associating those with the records, entering also-known-as names, or other names that entities are known by. I want to also point out that we recognize that there are, in a number of cases, multiple entities for the same institution.

An example that we have talked about here is the National Zoo and the Friends of the National Zoo. This is all part of the process of really casting the net quite broadly, and getting many different types of institutions, and then adding value over time to the record and differentiating the data over time.

I mentioned data cleaning. But you know, there's also been data use. People have already started using this data. We have already started using this data. I've got examples. The data has been used. People have been linking the Museum Universe Data File to other data resources. People have been merging the data with other data files. Here is a couple examples of that.

One example is a researcher at George Mason University's Rosenzweig Center for the History of New Media. His name is Patrick Murray-John. He has taken that data file, the Museum Universe Data File, and he's linked it, he was able to augment the data by making use of a simple linked open data connection, and connecting that data to wikipedia DB.

Basically he is merging the data based on URL, based on name. And when there is a link, a positive link that is made, then the information from wikipedia gets drawn in, and you can find, you will find something like this, a more detailed information about the institution, phone numbers, for example, web addresses. And if there is not a record within wikipedia, then what Patrick has programmed is, he's programmed something called a sub record. Basically, it's a record that allows you to start entering information on to wikipedia pretty easily using a simple form.

He's even provided a little bit of help information on how to add your museum to wikipedia. Once people start adding information into wikipedia, that information gets automatically associated with a record within the Museum Universe Data File. And data becomes, starts to get aggregated and is augmented in that way in realtime.

That's one use of the data, and as a linked open data system. Just this week, somebody from our Office of Museum Services came to us with a question about the location of Native American and Native Hawaiian museum program grantees.

So we have manipulated the data in that earlier, in one of the formats that I showed you earlier, our open data portal, but we have also put it into a system called cartodb, which allows us to zoom in and zoom out of different geographies. What I've displayed here are just two different data systems, two different data elements.

One is basically the geography of Native American identified areas, tribal areas, within the United States. That is the orange areas. The dots are grantees of the Native American Native Hawaiian museum program. This is the big view for the entire United States, Alaska and Hawaii, and no, Hawaii is not as large as Connecticut. I just changed the scale so that you could see it a little bit.

This second layer or second visual is basically, it's the Choctaw area, Oklahoma tribal statistical area. I've left a layer on that shows the tribal area in orange. I've left on the layer that shows the grantees, that's the larger dots here and here and here. But I've also folded in the data from the Museum Universe Data File.

So basically, you have got three layers of data here, and it shows you where the location of those grantees, within this Oklahoma tribal statistical area. It also shows you which records within the data file have obviously not received the grant. I've highlighted here the Choctaw National Capitol Museum. This simple visualization, three layers of data, shows you the way in which we can basically use this information, that has very clear programmatic implications. We see who within this tribal area has applied or received grants support in this program. We see other potentially eligible entities within the same area, that we can reach out to and let them know about the program.

I could basically repeat the same process with different areas, different policy areas. It might be areas with high concentrations of child poverty. It might be areas with many English language learners. Speaking to folks at the American Alliance of Museums, there were folks there who were interested in coastal flood zones, and which institutions are at potential risk in coastal flood zone areas.

So this type of layering, this type of spatial layering is something that planners, policymakers use readily, and it's something that we want to make available to our stakeholders as well.

This is institutional data, but it's a different visual. It's a different use of institutional data.

What I'm representing here are tweets that reference museums. It's tweets that are referencing museums in different countries. You have a cluster of museums at the top from the Netherlands. A cluster of tweets about museums in Germany, a cluster of tweets in France, the UK, and this big cluster on the top right is the cluster of museums in the United States.

This is a cross-sectional export of Twitter data. All that was needed was the Twitter handle of the institution, the reference of the institution, and then these sorts of associations were made.

So, for example, this person who tweeted about an institution in the U.S. also seemed to have tweeted about institutions in France, for example.

So I show this to you to give you an example of network data, ways in which one can use this institutional information to look at the structure of the museum sector in a different way, and the way in which the public is associating with those institutions, at least via Twitter.

And I show this to you as an example of really what the future of data manipulation, data visualization looks like. I know that many of the folks that are out there in museums that have called in for museums themselves are likely to have Twitter handles and are likely to have Facebook pages. And this kind of manipulation is something that quite frankly you could do, or you could set an intern to do. There are pretty easy YouTube tutorials on how to make use of your Facebook data and to visualize networks.

Let's get back to the uses of the Museum Universe Data File and some of the things that we know and some of the things that we don't know and that we are hoping to learn.

We have had the question: Are there 35,000 operating museums? We were a bit ambitious early on. I think in our first presentation. The truth is we don't know yet. Our original announcement was based on what we knew at the time. We gathered this data from these different sources. We obviously analyzed the data, and we didn't realize how much variability there was beneath this data.

There are organizations that file with the treasury on an annual basis under a U.S. museum code that care for institutions that are outside the United States.

Are they a U.S. museum? Well, if you are doing a economic analysis of the museum sector, some may say yes. If you are doing a analysis of museum services accessible to young children here in the Continental U.S., you would probably say no.

We have learned a fair amount about the way in which the organizational structures differ. Sometimes they differ from state to state. Sometimes they differ within the state. We have identified, for example, nonprofit entities that are not coded as museums, within, by the IRS that is, but yet they operate historic houses. They hire curators, they are open to the public.

So, these present interesting challenges to us. I might say they present interesting challenges to the museum sector as a whole, to learn a bit more about what are the boundaries of this sector, and that is something that we really want to work on together, and public input is vital to this process, as we leverage the data and we add additional information to the data.

Our next steps, well, I mentioned some of the cleaning that we are doing now, some of the record cleaning. But as we look forward, looking down the road, we want to identify attributes for these records. We want to identify whether or not these organizations are open to the public. Are they using collections, including living collections. Do they have staff or volunteers? Do they provide exhibits or programs? These are elements that are going to help us cull through the data, and sort the data in different ways.

It will help us address questions, for example, like how many collections-based institutions are in a given state or given region. How many of those institutions frankly have no sign of a staff or don't seem to have hours open to the public, but yet they have collections.

So, what I want to point out is that this process of adding additional attributes is part of an ongoing exercise for understanding the different elements of the museum sector.

We recognize that IMLS is not an island. IMLS is a public agency that, where public input is vital to the work we do.

And we want to assure people that we take very seriously the obligation to get input about what we do. In fact, it's written right into our statute. We are identifying formal, informal processes for public input. One of the formal processes is the Museum Research Federal Advisory Committee that we are establishing.

This is established under the Federal Advisory Committee Act. It's a formal process that requires notification in the federal register. It requires that the information that is obtained in these meetings be a matter of public record. But we also seek input from informally, most recently I was participating in a meeting of the international standards organization. We were talking about an international museum statistical standard. But to get input on that standard, I reached out to museum study professionals from across the country, shared with them the standard, got input from them, collated that and provided it to the chair of the working group.

We also present finding, research findings at professional conferences. Earlier this year, someone on my staff presented at the American Education Research Association. I'm presenting at the Museum Computer Network coming up in Dallas. We shared research results with the European Group of Museum Statistics, EGMUS.

And of course, we will continue to provide updates via webinars and blogs, and to seek your input in fora like these.

But I did want to talk a little bit, go off on a broad note and let people know that we really see this effort of gathering this information, collating it, establishing standardized procedures for collection and distribution, as part of a broader effort to make open data available, linkable.

And I like to refer to folks in the linked open data movement, and I've inserted cultural in there, because there are a lot of entities that are engaged in the linked open data, linked open cultural data movement. A lot of these folks are focused on linking content and interoperability of content management systems, for example. What we are doing with the Museum Universe Data File, you can think of it as an institutional corollary to that. We are making a network of institutions available that can link to a linked open data network on the cultural and the content side.

We think that both parts of that puzzle are pretty important, one clearly very important for understanding the care of collections, the quantity and quality of collections, the other, quite important for understanding the vitality, the size, the characteristics of the institutional, of the institutions that support those collections.

So, with that lofty note, I want to thank you for your time. I want to step back and ask you if there are some questions.

>> If you have any questions, feel free to enter them into the chat box.

>> CARLOS MANJARREZ: We will hold off for a few minutes to give you some time to think of a stumper. I know that it can be a challenge or has been a challenge for some people to access the spreadsheet. I want to let people know that we are working on different tools to make the data available and accessible in different formats, so it's easier to identify your institution on a map, for example, and to verify some of that data.

Oh, okay, Katie asked a question, for people interested in helping to update or scrub data, how can we help to identify attributes like open to the public or collecting institutions? I'm really glad you asked that question, Katie.

Right now, as I mentioned, we are trying to define those attributes and operationalize them in a standardized way.

What we are doing right now is we have developed a pilot process to do that testing. Once we have completed that pilot process, we are really going to be putting those attributes out and asking for comment on them.

That is something that we anticipate doing through the museum research working group.

But we are looking at other ways that we can do that as well. I know the agency has used different processes for public input, idea scale and some others. We are looking at different options for getting public input on those.

Once we have had public input on the way in which those are operationalized, the next step obviously is doing the work. We would really love to talk to people about how to go about that.

Good question from Debra. If we find an error in the information, on our institution, how should we correct it?

What we are asking people to do is to contact us at research@imls.gov. And let us know about what the nature of the error is. That E-mail address is being monitored by folks who are tied into this project. What we are saying to people, we can't promise realtime updates. What we are doing is we are gathering that information and folding it in to our release, our subsequent release.

We are basically, in order to process all the information we have gotten thus far for the December release, we really need to close the door by the first week of November, in order to get that data processed and updated for the December file.

But you shouldn't hesitate, if you still, if there is input after that, we will fold it into the next release.

There is a question about defined code that we should use. In fact, there is a code book associated with that data file. So, I'll go back to a page that shows you where the data is.

If you go to IMLS data, the page looks like this. Right here, if you click on the Museum Universe Data File, you will get access to the file but you will also get access to a full code book. That code book basically identifies all of the different variables. We don't anticipate that people are going to tell us the longitude or the latitude of their museum. There is no need to do that.

With the address information, we can associate that specific information to each record. Essentially, what we are looking for is a name, address, phone number, URL, and museum type. That is the kind of information that we are looking for, if it's not available in the file or if it's incorrect in the file.

But some of the other metadata associated with the record, we can really tag that.

So, Celeste said thanks for the effort. What are the expectations, if any, from other museum service organizations or associations in helping IMLS in this endeavor?

Well, we really do, as I mentioned, baked right into our legislation is the need for consultation with stakeholders, museum service organizations included.

So we are looking to them to help us as we operationalize these different attributes, for example. We have already had some input from folks about analysis that they would like to see, data that they would like to see layered with the Museum Universe Data File.

I talked earlier about the risk analysis of flood zones and what institutions might be at risk in the event of a flood zone. So for that analysis, we gathered data from NOAA, and NOAA has a number of data files that allow you to basically map out potential flood zone areas, coastal flood zone areas. Then we have layered that in the same way I've shown you with the tribal areas. We have layered that with the Museum Universe Data File.

That is something that we are sharing, we will be sharing with AM and that we will be making available to everyone on these data portals.

Good question from Arizona State University. How do you feel about including all of and each of the different museums that are part of the single university?

There are two, in my nerdy researcher mind I hear two questions implied there. One has to do with institutions that the public might identify as a museum, an average Phoenix resident. The other has to do with the governance structure of these institutions.

The governance structure as you all know as museum professionals can be quite complex. There are parent institutions. There are institutions that are associated institutions, friends of organizations. At the current time, we are not identifying those, the pairing of those institutions within the Museum

Universe Data File.

That is something that we are looking forward to doing. It's a complex set of tasks to identify different governance structures, and it's likely going to require multiple data files that relate to one another, relational data files.

So we want all of those institutions right now as atomized individual records, because we are trying to identify each one right now from the user's perspective, from the perspective of your average attendee. And for many people, they may not see the distinction between the different museums, even if they are on the same campus.

They may see, they may want to go to one institution, and one institution alone on the campus. The association, the governance structure, that is something that we are going to be working on over time. We would love to have your input on that. Katie asks are there ways to stay more up to date with what is happening with the Museum Universe Data File?

We can pledge to provide more webinars like these. We can also talk, we can also share updates as we are cleaning the file, a blog about the work that we are doing and share visualizations that are in progress. Does that address your question? I hope so. We will try to tweet about it too.

Okay. Oh, here is a question, from somebody who is viewing on their phone, a high tech person viewing on their phone. I don't have access to the webinar but my question has to do with the coding and the 990s.

Carlos said that it sounded like it was treasury. It is actually when you file, when an organization establishes nonprofit status, with treasury, that code is established based on IRS staff. So they take the information from the application, and they code it within national, basically the code is called the national taxonomy of exempt entities. In fact, you will find that entire coding system on the IRS website, but also an organization called the urban institute makes that taxonomy available.

It is obviously not just for museums. It is for hospitals, human service organizations, education organizations, mutual aid organizations, all manner of nonprofit entities.

And I think, I know the person who asked that question, so I'm going to say something that I think might be implied in that question, and that is, how reliable is the treasury designation of the nonprofit status?

Well, they are our colleagues and I will say that it is pretty good. But it really does need a lot of work. There are many, many entities. One of the things that we discovered, for example, is that some entities file as educational organizations, the nature centers, for example, we were contacted by the American nature center association. They said, what the heck happened? We are not in the Museum Universe Data File.

We learned that most of the organizations that fall within that association have identified themselves or have been identified as environmental education organizations.

So, we are working with them to incorporate those entities into the file. We have identified other organizations that are defined quite generically as A50, general museum. When they may not identify themselves as such. So people who have sent us corrections on the museum type have asked us to change that museum type, that they are not a general museum, that they are a children's museum or an art museum.

There is a bit of slippage with the treasury. Remember, this is administrative data. So the coding system is based on what you are using it for. And that instance, that coding system is for treasury and for tax purposes. Given the fact that these are nonprofit entities and this is a, there are no funds being transferred as a consequence of that 990 submission, there is not a lot riding on it.

So, I think there may be another question down below.

Oh, there is a question about on-line only museums. We are not restricting museum types. As I said, we have cast a very broad net. What we are doing now is that we are identifying different aspects, different attributes of those museums.

So, in the case of an on-line museum, we would be looking for a physical address. We wouldn't find one. It wouldn't be in the file. We wouldn't be able to geo code it. That is fine. It can remain in the file. It just won't get a longitude and latitude.

We will be looking for staff. The on-line museums I've visited, it's hard to determine staff for a number of them.

So we will be looking for attributes of all of the records that are on the file.

So, Victoria from ACM, I'm going to say the association because it is embedded in the statement, ACM appreciates the data file, we have gone through the list and identified which are open children's museums as well as those in the planning process for, in parens, emerging museums. We also identified those on the list that are not children's museums. And we will get this information to you.

We welcome the information and we also welcome the information of how you went about determining what was and what was not a children's museum. That is actually, that is going to help us a lot in the protocols that we develop moving forward.

So a couple people are leaving the room. It's 2:43. We have a little bit more time. If you do, please feel free to send in questions. We will hold out for a few minutes longer.

Victoria, thank you. Couple more questions. Somebody is asking what is the best format to get updates. They have asked Excel spreadsheet, list format. Excel spreadsheet is fantastic. Against any research@imls.gov is the place to send it. We can ingest it in pretty much any format you send it. But especially format would be great. Question about the slides being available. Yeah. You bet.

We will make those available on the IMLS site. Thanks for the question, Ben.

Thank you for your time, for those of you that are still on the line. I think we are going to wrap it up, with no more questions. Feel free to send us a E-mail at research@imls.gov with questions or comments. We would love to hear them. If you have specific updates on the file, that is the place to go to.

So, again, thank you for your time. We look forward to tuning in with you in the future.

>> Recording stopped.