Federal Open Data - 2022 Update – Transcript of audio

[please stand by for realtime captions]

good amp, welcome to the FOD academy. We have another terrific webinar for you today entitled federal open data. 2022 update. Again my name is Joe PASKOSKI from GPO. My colleague Ashley for tech support and with us our presenter, HYON Kim. Let me read a little bit about him. He's the program director for data.gov, the United States Government's open data website operated by the U.S. general services administration. For workers focus on expanding the scope of public federal documents data sets to assist the public businesses in academia. Prior to GPO HYON was a program manager at the office of national intelligence. Other federal experience working as assistant general council at the federal bureau of investigation and serving on the staff of the Senate select committee on intelligence. Prior to her work in Washington HYON was a litigation associate at white and case LLP in New York. So with that, I'll hand the virtual microphone over to HYON to begin her presentation.

Thanksgiving, Joe. Can you hear me okay? Before I start. .
Yes, fine.
I'm sure you'll step in or Ashley if there's any issue with internet getting faded or anything like that. Thank you very much for inviting me to present in this FDLP webinar series. I think I last presented in person around this time in 2018. I think you had a conference and I recall presenting in one of the break out sessions and even though it doesn't seem -- it's not a long time ago but somehow seems like a long time ago in 2018. Having a big in-person event but thank you very much for inviting me. We don't generally do a great job of seeking external audiences to get the word out and I've always thought there couldn't be a better group to get information about what is happening with the federal Government data site or federal open data in general than to this group. So, we determined that I couldn't do a screen share somehow. We couldn't get past that so we do have a lot of screen shots of the sites that I'll show to you and I'll try to paste the links to where I'm looking each time so that if you like, you can go see the live site while I'm going over the presentation. If you heard this presentation in the past, you know that there's actually pretty SUSH substantial history at this point. Data.gov, the U.S. Government's federal open data site has been in existence since may of 2009. It arises from the open Government executive order and later an open Government directive that mandates the creation of the federal open data site. And the big policy development since then have been a 2013 OMB order on making federal data open by default. That had a lot of detail and established the frame work of a Meta data catalog where the agencies keep the Meta data and maintain it in a central federal catalog. So that model of this harvest model of populating a federal open data catalog has been in existence since then and since the last presentation, the big change is the open Government data act which is title 2 of the foundation for policy making act which was enacted in January of 2019. So it may be current federal open data catalog.gov a statutory mandate and also required federal agencies to do what they have been doing which is making Meta data for a federal data catalog but also made those provisions mandatory and also established a lot of structure around how that was going to happen. So I'll go over those provisions in a little while in more detail. To give you an overview, data.gov is the U.S. Government's open data site and as I mentioned, it was launched in may of 2009 GSA was the agency that launched it. It's mission is as it says, there is so much federal open data that can impact or be useful almost every aspect of daily life so it's to help agencies achieve their mission, drive innation and promote open and transparent Government. So when it was launched in 2009 it had 47 data sets and right now we'll go over the catalog a little bit and the number at front is over 300,000. There are numbers all fluctuating because of the constantly updating catalog. There are around 60 federal agencies and there are many sub agencies and we also include data from state, cities and counties if they volunteerly make their Meta data available for harvesting. the outdated comparison with the catalog of years past

when there was an entry for every book, there's an entry for every data set but it's actually hosted generally at the agency's website so mainly a catalog to help you discover data sets. As I mentioned, it's been this harvesting model where we pulled the Meta data from agencies that are supposed to keep them constantly updated and the data catalog harvested by checking those inventories in most cases on a daily basis. As I mentioned, it became all of this the agency keeping the Meta data and operating the catalog became a statutory mandate as of January of 2019. And I'll go over the main provisions but it also expanded the scope of data.gov in that prior to the statute it was mainly focused on the chief CFO agency, the 24 biggest agency that is have a chief financial officer but it increased the universe of federal agencies that are supposed to have Meta data and data sets. The components I'm showing you today, the heart of it is the Meta data catalog. It runs on an open source platform called CKAN. There's a landing team and private site called inventory dot data.gov and I'll show you briefly that's used by some agencies for managing ma ta data and the team is also involved in policy support for the open data efforts by coordinating with all the agencies involved in data.gov and by hosting a monthly meeting where we share best practices and latest updates on what's going on and open data. We also manage a listSERV for anyone interested in keeping up and it has about 1,000 subscribers. The new part of data.gov which I'll also show you, there's a repository of best practices guidances at resources.gov and that's a new site required by the open Government data act. As far as users, federal open data is just so broad so to the extent that we find out about the users which is through the request and questions that they send to us, we find that it runs the GAM met so it's general public that has questions about a variety of topics. It's businesses that use the federal data to support their own apps or business processes, also a lot of questions from researchers and academia. So this is a little small. Hopefully you can read it. I will put in the link. if you want to see that. as I mentioned the big development in the last couple of years is the open Government data act title II of the foundation which enacted a number of provisions from the foundation for evidence based policy making commission so it was a bipartisan commission that had a number of recommendations and the open Government data act which had been around as its own standing legislation for a while got grouped into the foundation for policy making act and became law in January of 2019. So, what it requires is what most of the agencies, the larger agencies had already been feeling which is maintain Meta data to make their data sets available in open format and make that Meta data available to a federal data catalog which is.gov. So requires the OMB director to issue guidance on what is required in the Meta data. And that for the agencies to develop and maintain these data inventories and continually update them, it makes the GSA operation of a federal data catalog a requirement under law. And also required GSA to work with OMB and the office of Government information services within the national archives. Actually, it's focusing on issues requiring three to work together and develop and maintain an online rePOZ TIR of tools and best practices and resources. It created the position of chief data officers and required every federal agency to have a CDO and established a CDO council under the OMB and then I can give a link for them. The open data responsibility fall under the data offices but the officers have more that they're responsible for under the legislation. And the open data aspects. So the chief data officer council is very active and has a number of subgroups that you can check out on their site. The support for the council is out of GSA but in a different division than mine from the office of Government wide policy. okay, so this is the main site. we're going to have a new version of the site. Just trying to do more design OB it. So the main data.gov site I think in a couple weeks will look different but it's where we have the most basic information about federal open data efforts and also a link out to the other parts of data.gov. the most heavily used part of the data.gov site is the actual Meta data catalog. I think if you go to the site now, you can see that the number of total data sites is already different from when I did this the other day and that shows it's something continually changing in that we go to all of these agency sources every day with the exception of massive inventory. For most agencies after business hours eastern time the catalog goes to where their ma ta data is kept to continually update the catalog. The number 300,000 or so is just a way that the software called CCAN groups it. If there's a collection meaning a data set that has many similar data

sets grouped under it and it's called a collection, that counts as one in the 300,000 number so the actual number of data sets is much larger so I haven't checked but they are somewhere between 1 million and 2 million, generally. so this is a view that if you wanted to look at if you clicked on the organization tab that's on the upper right, you would see the whole list of organizations and you could scroll through it. And then it's an easy way for instance, you just wanted data sets from a particular agency, you would search in that box where it says search organizations and get to just that data set. So I am going to do that to use the link or the next. so this is a typical data set entry page. Let me put the link here. We checked Google analytics fairly recently and I'm not quite sure why but when we looked at the most data sets in the last couple months it's steadily been this FDIC failed bank list and I haven't figured it out so if you go to that link, this is a typical data set page and you can see that it has a description about what it is and if you go to the download button, it should just download that CSB of the failed banks and then the visit page button should take you to a landing page on the FDIC site that explains it so everything that's on this catalog page it's maintained by the agency in this case, FDIC and what the catalog is doing is reflecting whatever we got most recently of Meta data and then if you scroll down on that page, you will see that there is additional information. With some standard Meta data fields and then there's a button at the very bottom that says show more. And that should give you the full list of the required Meta data for that data set so generally, it's about the agency that its from, when it was last updated. There's a contact. Unless it's not in here. Generally, there's a contact for the particular data set. I think I might be missing it here. We would like to have a contact email address associated with each data set. So that if someone is looking at it and has questions they can easily get to the right person in our office for more information about it. okay, so if you noticed in the prior screen shot, in the bottom half of that screen there is a Meta data source section and it says harvested from FDIC data. And that's just showing you where the agency's Meta data inventory is and I'll find the link here and put it so you can see a typical Meta data source. here's a link for that site. So every agency is going to have a similar site where it's agency.gov and however many data sets they have they have described the Meta data for each one. So the catalog is scrolling through the list of agencies and locations like this on a daily basis so if FDIS or any agency wanted to add or delete a data set or edit something in an existing data sets they make changes on their end. Once an agency is set up and we have their harvesting location, there's very little that needs to be done. We have set it and generally, we can forget it unless something goes wrong in the harvesting process. this is the site that I mentioned before. That's new. This was required by the Government data act for OMB, GSA and national archives to work together and you'll see there are a number of different headings and tabs so if you go through the various parts of the site under resources or standards, communities there are a lot of different open data resources including pass the 1-B policies, schemas, and standards outside standards. And guides for agencies who are working on open data. So there's a lot of resources on this site. You can see that the very central thing is the schema. I'll get you the link in case you don't see it. if you click on that link or go to the link I just put in there, that is the schema that is being followed by all the agencies that describe their Meta data. And if you scroll down on the page, you can see that there is a lot of detail on every single possible field. And if you scroll down further, there's a section where on the right hand column under required, it will either say no or always or as required or so the only fields that are required are the ones that are marked always meaning that if an agency describes the data set but doesn't have something that's in the always category like you have to have a title. Then when the data.gov catalog goes to harvest it, it won't pick up the record. It will not have validated against the standards with these required fields but ultimately, there are not a lot of required fields so for an agency that is just starting out, it isn't that heavy of a burden to find the fields that are required for each data set. If you click around in there you can see for each field there are links and there's much more detailed information about what each of the fields mean and what the standards are for what might be acceptable for the field. If it's a date like what formats are acceptable and so on.

so this is the site that's not public. Having an inventory to meet those standards that could be a burden and so when we launched the current version of the catalog years ago using this schema, we created this separate private square. If you go to it you will get a "forbidden error" but there's a log in button for registered users of this site and this is something that we operate to give agencies a tool to create that Meta data. So if you were to log in and if you had an account log in, this is what it looks like for me. I have access to all the agencies. It's very easy to create an organization in this. And I'll show you, for GSA, as an agency uses our tool to create their Meta data inventory meaning that the document that they need to most at GSA.gov is created by using our tool. So when you log in for particular agencies, you see all of its data sets and if you see that there's a way to add a new data set and then if you click on that, there's a Meta data entry form with three screens so this schema I showed you from resources.gov is already built into there. Anything that's required would be in the first page. And then an agency would add the title description tags and the fields are already built-in and the required Meta data is on the first page knowing you have to provide at least that. And so for agency that is don't have another tool available to create their Meta data, this is a useful service. I think we have more than 12 agencies using it to generate their Meta data. So adding a data set means filling out these fields and we also have validation so that if you enter something that's not a proper value that it will tell you okay, it doesn't need a certain requirement and then what happens is that once you added the data set, there's a button up here on what they do. It lets you download a data set so GSA and all the other agencies using it, they have users who log in to add or edit data sets whenever they're finished with that update, they will download an updated data and post it at the same location. so we're a pretty small team. Right now we are three Government FTE with contractor support on the technical maintenance and operation of the site. We work completely in the open and the screen shot here is of our current board and you don't need an account to see it. So at any given point you can see exactly what we're working on so what we're working on right now in progress which is what is probably going to happen in the current sprint meaning the first two weeks and from the product backlog, you can see what is up next. So if it's up at the top, that's probably the thing that we have prioritized and will bring in next as soon as we have finished the items in the backlog. The code for data.gov is also in the open and so the depository is open for anyone to suggest or if they notice something is not working properly. We'll review those and also see what we have done. Most of our effort in more in the last year or so has really been solely in the back end if you're familiar with federal website requirements, we are required to be accredited and go through the security accreditation process at our level every three years so we have that in the last year or so and we're also shifting infrastructure of where our system is hosted and so most of our work has been not visible to any users. It really has been on the back end of making our infrastructure stronger and also trying to improve that harvesting process. so this is the hub issues and then there's also the code for anyone that wants to take our code and set up their own data catalog or do anything similar using what we have developed. So as I mentioned, because our focus is so much been on the back end, we're hoping for the fiscal year we can do something more visible to users. One of the things that we have done in the last year though that is actually doing user testings with users from the public. People we don't know from all walks of life to see what is easy and difficult about finding data sets and using data sets so we created a number of issues in our backlog to address that. There are features that we have meant to work on for more than a year. Specifically on metrics agencies would like to know out of 300 data sets that TSA has, what's the most popular one as far as what's the most viewed and downloaded so we do have that information but not in a report or something that's readily available and that's something that we would like to share with agencies and also make public so that anybody can see what is being accessed and what's the most popular data set. There are some stylistic things also to make sure that we comply with the U.S. westbound design system standards and also the 21st century idea act so it has to do with accessibility consistency and everything that applies to all agency websites and any major updates. The noncatalog part will be launched pretty soon with a new look. We have not been able to foe can say on getting the content and current content for that so we hope to be able to direct more

attention to that since we have been so focused on the back end and also the proper functioning of the catalog. That harvesting process that I mentioned we have even though it's not hundreds of agencies, the way that Meta data is maintained it's not as neat as one source for department of commerce. It isn't possible just because of how existing data sets and existing systems are at agencies so for instance, you get census data, we have to harvest hundreds of different census sources. Not all the time but it's the complicated process to be able to accommodate the way the data is at agencies and feeling with hundreds of harvest sources so as you can imagine, that process of harvesting and doing it on a daily basis, it's not something that is just running smoothly without any intention to it. We know we need to work on some improvements on the harvesting. We have tried to engage with additional agencies so there are smaller agencies that have come into data.gov as a matter of the open Government data act being passed but I'm sure won't surprise you, many smaller agencies that have not gotten their Meta data inventories together to be able to be harvested into the catalog and as I showed you, we do have a tool that we do, it's very basic but we offer it to agencies without charge if they don't have anything to create a Meta data inventory so we occasionally do outreach to smaller agencies to say we're here and can we help you get your data sets, however many you have into the catalog. We work closely with the chief data officer council considering their roles in the statute and also just in operating all of this. I think the biggest thing is that the Meta data schema I showed you has not been updated since 2014 so by virtue of a couple of provisions in the statute itself and the passage of time and experience and different fields that agencies want out of there removed we have to do an update OCHT schema which would be pretty sizable undertaking to agree and arrive at a schema that everyone agrees to but also mechanically getting that change happening at all the agencies and in our catalog so we're just as far as a big project as far as significant overall open data development across federal agencies, that's probably the next most likely thing that will happen. So our contact information, we do have a section and contact form that is very basic for any kind of comment you have or people use it to report okay I found a certain data set but the link is broken. At that point we would work with our agencies contacts and let them know, look, someone let us know this ALTH SI URL is broken so can you update the Meta data. We have the Twitter account and this generic email goes to the core team so we tend to offer that as the one for public inquiry general questions. Then we have different addresses for our agency partners where they're getting into the true technical questions about the harvest didn't pick up three things and we want to figure out why. So yeah, if you have any other questions or comments, suggests for improvement, that would be many ways to get in touch with us and it's a small team so whatever you use, you can get to us quickly. I think there's a question, Joe, you will share the PowerPoint?

yes, Ashley can comment on that. That's on our website either tomorrow or the next day, the recording. And the PowerPoint. it will be a PDF. .

Okay, I see a question, do I have a stick to get it. No, not speaking from the data.gov program GSA end because I'm sure you can tell, we're basically the imp meanters of all this. Our primary mission is operate the catalog and to work with an agency. I think as far as any I would say the stick is thinking the open Government data slide I mentioned there's an GA overport that is written into the statute so the GAO issued a report in the end of the 2021 as required by the law. And actually, I have the link for it here. So that for us there were a couple recommendations and we have been working on the findings for those. There are a couple agencies that were called out in that. So the one report that GAO was required to do under the law they have done. Whether there's additional GAO audits on this issue, we'll see. And then the other thing is basically OMB. They have been -- it is EBB and flow. As far as levels of involvement in the true compliance monitoring so it's been higher and lower in the past. So the stick part I guess would be from OMB but we work closely with them but our job is to make sure that the catalog is running and be here for the agencies that are trying to get data sets in and to the extent of our ability, help the agencies do that.

thank you, great presentation. Anymore questions? ?

will PL need to be reviewed? What does PL mean? ?
oh so there's no time limit on the statutory requirement. There are some time limits on the CDO council. I want to say it's like five years so there comes a point where the CDO council they have to decide this council is something we keep running. I don't think we're there yet so I know there's something on the CBO aspect of it but nothing in it that says the open data catalog expires or anything like that so it's there until statutory revision otherwise.
okay, great. Anymore questions? Pretty good on time. Jennifer, what do you mean by GPO question mark? .
Oh I see GPO data sets. Work with GPO to find sets? Okay. GPO is an agent to find sets? I get GPO colleagues here in the audience if anyone wants to chime in.
I would say overall the law and policies have always been very expansive. The intent is for agencies to put Meta data about everything even nonpublic data sets to share with the public there's an existence of a data set so everything they're publishing and letting people download but also things that they -- and there are many nonpublic data sets especially from DHS where they're telling you about the existence of a data set that might have to do with transportation security and there's a description of it so they're sharing it exists but it's not something that has a download or access link because it's not intended to be a public set. So the intent is to have a really comprehensive inventory but the reality is it varies by agency so we're always trying to get improvement pretty clear there's nowhere near really fully complete accounting every single thing agencies make available or have in existence in the catalog.

Thank you. Ashley just put the webinar satisfaction survey in the chat so please fill that out, if you would. Jennifer says so you have.
That's exactly what is in the law. The law says that unless there's an exemption for it, the data set should be in the catalog. and it also says that one of the Meta data fields has this gone through for review. And that's actually in the law so that's one of the small number of fields that are indicated as in the law as definitely part of an update to the schema but that has not happened yet. .
Okay, I see SE seal YA says great information and I would love it if the FDLP could create a similar infrastructure so the FDL's can add records to the CGP. That's very interesting comment. I don't know if any GPO colleagues want to comment on that. We'll think about that one. Okay Jennifer says you need more money.
yes, she's very familiar with that. So interestingly in our -- there's always support for it. But you know, it's for instance, in the -- you all know the reality of how things have gone. We are still in part of GSA in the annual appropriated section and the house and Senate appropriation bill for FY 23 is wildly different in the amount of funding so you know, as usual, it's the CR until December and then who knows whether we'll get the high or low or somewhere in the middle. I'm sure you have all went through it and it's a similar experience for many years.
thank you. More good comments in the chat. You might want to take a look there. Okay we have time for more questions, please send them in. Ashley, if you could put links in the chat to our webinar repository, I would appreciate that. This presentation for the last two or three years you can find them on our training repository so Ashley will put that in the chat. I'll make a few comments while you're putting questions in. First off, I would like to thank HYON for a terrific webinar. I remember that presentation at that conference and that was great too. That may or may not have been record but a great presentation today also and also I would like to thank my GPO colleague Ashley DAHLEN for great tech support and running smoothly. We have one more webinar scheduled for October so far. It's next Thursday October 13th. Titled "knowing NAICS and understanding the way the Government classifies industries." Also don't forget to register for our fall virtual depository library conference October 17th and 19th. And if you go to FDLP.gov you can find details there. Also, we have this year so happy to do this. We have a pre-conference which means the GPO updates will be on a separate day. The regular conference is pretty much programs that you can get a lot of information from. And the preconference is on Wednesday October 12th next week so please sign

up for that also. You'll receive notice of all of our upcoming webinars when they're announced if you sign up for the email westbound service FDLP.gov and from the academy webpage, you can link to your Cal ler of upcoming events and webinars and access past webinars from our train training repository and there's also a link to volunteer. That could be on any Government information topic or FDLP specific managing depository and things so all good for the FDLP academy.

let's see if we have any other questions. Or comments. often times the presentations are so good the questions are answered so thank you that could be the case.
That email address is very easy to remember. So if you take time to click over on that and have suggestions, questions it gets to that email address goes to me and my colleagues and we'll see them right away so any time.
okay. We'll wait another minute or so to see if there's any last questions for you. good comments about how to lobby people to get more money for that. Looks like the questions have run out. So I think I can safely close things out. Fantastic webinar and thank you very much. Please come back any time to provide us an update webinar or any information related topic. We would be happy to have you present again and thank you Ashley for great tech support work and thank you audience. Please come back to the FDLP academy and the next webinar will be a good one. It's complimenting the one we get on the census tool that we're offering to the depository with pass word access and also, don't forget to register and attend the pre-conference and regular conference in October. Those should be great events. Always a great presentation and a very large audience. Thank you. So I guess I'll close it. Have a great rest of the day.
thank you. [Event Concluded]