Event ID: 3215896
Event Started: 4/12/2017 1:45:27 PM ET

Please stand by for realtime captions. [Captioner is on hold, waiting for event to begin.]

Hello everyone, it's Jamie at GPO doing another audio check, feel free to adjust your volume.

All right good afternoon everyone welcome to that FDLP webinar Planning for and Managing Storage for Digital Collections, my name is [Indiscernible] -- Jessica of Lincoln Illinois, with her specialization and data generation from the University of Illinois, sponsored by the Library of Congress, served as the digital stewardship resident at the government publishing office, to certify the digital system, as a digital depository, to help achieve the goal of certification from 2015-2016, presently the preservationist of the 163163 audit. I cannot roll those off my tongue like you do.

I will walk you through it, if you housekeeping -- A few housekeeping items.

If you have any questions please check them in the Chat Box, any questions that come in at the end of the presentation we will answer them. We are recording today's recession -- Today's session, the webinar will be available along with the PDF slide deck on the website. Under FDLP Academy. We will send you a certificate of participation using the email you registered in today's webinar. If multiple people watch the webinar with you please email, along with the email, include the names of those who are attending.

To exit the full screen mode, to chat questions mouse over the blue bark at the top year screen, then the return button for the default view.

We will be sharing a webinar satisfaction with you, we will let you know when the survey is available the URL will appear in the Chat Box we appreciate your feedback after the session is through. I will hand the microphone over to Jessica.

Hi this is Jessica, my presentation today is an overview and introduction to preservation storage for up digital storage. This is low-level for a librarian, or librarian who is working with an assistant librarian or preservation storage and storing content, this is a broad information hopefully you can refer back in the future as you go through any individual projects or assessments. In terms of infrastructure and requirements. To begin with, I do want to go over the broader responsibility of preservation as accepted by the preservation community. As a review, what is the digital preservation and what role does storage play? The definition I think is most pragmatic with the technical services, a division of ALM, they provide three descriptions at the basic level, the definition on the sly, digital preservation come by and -- Combine strategy and policies overtime, they provide a medium definition which concludes digital preservation combined with policies and strategies and actions to ensure access for re-formatted digital contact regardless of the media failure and technical change.

The accurate rendering of authenticated content over time, in the definition they go through the measures of considering changes of media failure and technology change, this is where the requirements come in.

As you can see on the slide I provide a GPO's summary, as these overarching summaries, taking content from a submitter or producer or the long-term management of the content or repository, the consumer at the end getting the information.

Digital preservation strategies and actions, content creation, integrity and maintenance.

Clear and complete specifications, production of master files, sufficient descriptive, metadata for future access, and control processes. Contract --

Content integrity, strategies and procedures, uses and identifiers, change history for all objects, verification mechanisms, the security requirements of routine audits.

Maintenance includes a robust computing infrastructure storage and synchronization of files, continuous monitoring of fire loose -- Of files. Creation and testing for recovery plan. Periodic review and updated policies and procedures.

We are talking about content maintenance and integrity.

As you can see digital preservation described as a responsibility, in order to determine good preservation and evaluate and assess activities, we have to fully communicate a commitment to individual assess -- Access for any amount of time. Going from left to right, these become more specified. The first column which is in red, generally speaking the minimal level of preservation for digital collections. As you go from level I, two, then three and four you are providing more substantial metadata, services like emulation, and not just access is being provided, but also complete rent ability to those -- Render ability, and accessible in the way it was intended to be access in the environment it was produced.

Because all digital objects are presentation of binary zeros and ones, at the basic level, the responsibility of what differentiates digital collections to perform trued preservation for any, this provides content over time, with proper monitoring strategies, it eliminates the destroyed overtime, refreshing preservation hardware, through this process. It is not sufficient to backed up content, they must be proactively checking, preventing redundancy, monitoring risks as no storage media provides full reliability, this can be copied to a reliability of errors, for this rate of implementation and storage instructions. I will go into more detail on how that might occur.

On the slide I will present the most common forms of media for digital preservation, tape, optical, digital and flash, these are the most common used, there are others. Tape is chosen, the form of tape linear tape open or OCR tape, developed in 1990s, optical storage, on the readable medium this is precisely focused on a part of an optical disc. This storage mechanism, for data recording a mechanical changes to one or more rotating disk, a disk drive is implementing this type of storage mechanism, hard disk drives, HCD's, or floppy disk drives -- This means no

mechanical parts involved, this allows flash drives to read at a different rate. Storage at a service model, possibly backed up remotely or across data centers, made available to users over the network, may or may not be hosted over service, this slide presents a table for the portable storage media types, tape is low-cost. Reliability performance is lower, takes a lot of time to read tape, if you are simply trying to provide optical copy, and you don't need to read right away or quickly, tape is a good choice, or perhaps you have a small collection, it may be more relevant for something like that. It is not portable, it is very large and heavy crate, quite robust. You have to think about where you will store it. Large systems are disk, they can read much faster, these are some of the trade-offs, more often, they are going with cloud services. Often these have the same benefits, however all of your services are dependent on whatever vendor relationship you have with your service provider through the cloud, you have in consideration, you have not complete control on it being monitored, how quickly you can quickly monitor this coming off of the cloud, the benefits are on the cloud but many risks to cloud storage.

Now I will go into concepts, to have an awareness of if you are going to have conversations with vendors and what type of storage is most appropriate for your institution.

The first term is areal density, the amount of data that can be stored on a given unit of physical storage unit, measured in gigabits per square inch used to describe the capacity of your hard disk drive, how much you can put on space, synchronization, is one or more data, if you are doing some backup synchronization.

Failure rate is used interchangeably, this is the calculable weight based on the given life span, when you are buying storage, the failure rate, an indicator of how your drive will fail, how often you will replace it, and make sure that your data is backed up. This is the number given from the vendor, this is mean time between failures. A measure of how reliable a hardware product or component is, this measure given in thousands or tens of thousands between failures, a hard disk drive might have a meantime of failure between 300,000 hours. You can potentially estimate that you will have 300,000 hours for use before something might fail or between failures, redundancy checks, a system of area testing, calm code to detect raw data, a process automated by software in the storage systems, you do not necessarily need to turn on, or tell them to perform them it results in a lot of large-scale, big systems often have.

Now we have array, the user lysed storage preservation for strategies. I will provide a overview of how RAY works.

This storage allows for two discs, and allows for it not to lose any data.

This is not the only one on the server we have a certain number of discs in the system, which are data discs, then parity disks in addition, the controller serves as the brain of the system how it data is stored which one these are data discs or parity disks, how these are being stored across the discs, and organizing the bits. On this slide you can imagine. I will play with a highlighter to show this column represents a discs. And this column represents a discs, then you have a Perry disk each of these colors you have the red, the white, the purple and the blue, you can think of those individual bit that belong to a single digital file, all these purple blocks represent the bits make it up the digital image of my podcast, all of these are across they can be calculated on a

parity disk. In encoded way. You go and you can see a system going on, 3+1, +2. So on your double parity disk we can buy track -- Backtrack -- The fourth column should be 2, that is how double parity works, these coded blocks across these discs can be recalculated, even if we lose 2 of those discs, we can recover that data, you have the dual parity disk even if you lose a complete disk. -- I can revisit this concept and go over in detail.

A lot of people are concerned how many copies should I be keeping? The national digital starship recommends at least two copies of all digital copies, in geographically dispersed locations three or more are better, or if the digital objects are replicated across data centers, it is great if you have three copies, two of them are on a disk, one on tape, the only reason it's good strategy to have is some have different threats, so that all copies are not exposed to the same types of threats that could cause error on the reading or writing of the data or geographic issues where there is an earthquake in one area, or perhaps all of our disk centers are in a room that might get overheated but the tape is stored elsewhere.

Determining how many copies you need, is how much your risk and institution is able to count for, -- Account for.

Natural disasters or sound vibration, can alter the data at the point in which it is written to the storage data, it is important to recognize many copies will increase your cost, most can recalculate formulas depending on the pack up -- Backup parity you are using. Only 50% of your disk space might be usable raw storage, MIP 50% of a disk is raw data the other 50% is where there is backups, keep that in mind when you are purchasing storage and how much you need.

The next topic fixity, to remain unchanged over time. Repository managers will ensure that they have it's that are not degraded, typically unjust, these are unique to the digital object, this is the cryptographic signature representing the bits and some files, if any of this changes, so will the cryptographic signature refer to as a hash, if these change the repository managers know the bits have changed, currently there are a romantic -- There are 120 bit values, designed to be fast on 32-bit machines, where the V6, 256 hash requires computing capabilities to generate. If you have a large amount of content keep in mind it is going to require a much higher capable computing the process could be slow, if you have a smaller collection you might choose to use MD5, however the 256 provides greater.

I have a link on the slight to a publication -- Slide to a publication digital content when should I be testing at, a good resource to refer to, it talks about why some check more frequently than others, and for your own institution how frequently you want to implement these checks, and algorithms you will need to check the system.

The most complex of digital preservation storage I would say is cost, this is affected by a wide range of factors, files in the collection, disk array, and relying upon IT department, the operating systems and many other fact is. Because of the factors that could contribute. I cannot request any generalized, but I say that the libraries pay attention to the market trends, they are not always consistent with the expectations, so because of this I encourage everyone to follow the blog of dog are David Rosenthal, he was a leader in the storage program providing resources and insights of technologies and cost reductions for repositories and library requirements. Another

great resource is the Library of Congress events play age -- Events page they have an annual event for storage, notes and slides about the meetings all of these highlight opportunities for the technologies cost reductions, architectural design advancement, in general I would say it is important to have two major concepts as they pertain to preservation concepts, outlined on this slide, is more slot -- Morsla -- The cost of preservation will go down, this preservation storage cost will be less of the barrier over time the other Crider's law, it is not as important, is the increased capability of technologies to store more bits onto a hard drive, an example obviously most evident in smart phones, they have become smaller, they can hold more data, the areal density intends to decrease. Historically a 30 year history of prices dropping 40% per year due to these, having higher areal densities. There is doubt that these laws can provide a lot of certainty, going into the future this is decades of time, for most into tuitions -- Institutions, you will have to refresh multiple times, you have to take that into consideration for cost, it might be lower, but you will be repurchasing storage replacement hardware five or 10 years later, for some institution, cloud storage has become a solution for lower cost. You have to keep in mind the market is dominated by very few providers, it is uncertain going into the future what that means for cost. There is very little market are few people out there to purchase cloud storage from.

At the same time. Even if the storage has been cheaper this has been increasing 60% Dr. Rosenthal, I will reappear -- Refer you to his article if you have questions about more detail.

This is a case, just to keep in mind, this is not saying that storage will get cheaper, we do not know for sure what those projections will be.

Speaking of risks and cost, we can go through the managers repository, and storage preservation, this spot model, these are two methods for risk assessment, the spot model stands for simple process risk assessment, applying six I have highlighted on the slide properties most relevant to digital preservation and storage requirement. This is persistence that frequently includes, improper negligent handling of storage, useful life of storage media being exceeded. The equipment being unavailable.

Bit sequences.

Or damage to media by hardware or software, the other is risk assessment, providing a link to Drambora, which is a useful way you can download it as a toolkit, you can drag and drop which ones are pertinent to your organization, and fill it in within the toolkit, you have this full range of comprehensive list of threats, you might see risks you never would have thought of on your own.

Obviously I have been at GPO since external [Indiscernible], if you refer to the standard, this certification, you can view the criteria as a means to gains, what kinds of threats you need to consider over 20 of the standard is dedicated to them for structure and security risk assessment, when I was preparing GPO stats for assessment internally, I developed a list of questions to interview everyone in terms of monitoring our threats to our preservation systems, and how we are considering costs and financial risks, and monitoring backups all of this stuff. Because of this webinar is recorded and you can download the slides. Hopefully you don't take the time right now to read the slides word for word. I wanted you to have those list of questions to talk with staff to determine risks how we are monitoring them, are we doing this appropriately, are we

doing what we say we are doing? One thing to say we have a backup plan, it's another thing to actually ask these questions and see, perhaps I thought I knew the answer to this perhaps I don't. If you are meeting with your IT staff, or administration or just curious on storage, in general, these questions are useful as a starting point. Before you do assessment. As some of the questions on here. That might be good. That might stick out, does the repositories the any potential for licenses to change or any time when the license has changed and the repository has maintain cost? Referring to changes in the virus software, perhaps there is a system that perhaps you do not realize has a major impact for your system, maintaining awareness of the license or anything that causes edger O-Matic change is important. You should also have a relationship with your IT department documented, do you have something like a service, something like shared responsibility who is doing the checks, does the system administrator realize the content on the storage systems are [Indiscernible], ideally data that needs to stay there, well into the future. Another question to consider, do you have a disaster recovery plan, who oversees the execution of the plan? If you are seeing the disaster preparedness plan on your computer, and tell your personal computer breaks, then this is not any use. You can understand in addition to how this plan is important we can refer to some of these questions that might be useful, what would happen to data being ingested, or this is to test out with your data systems people, if you pulled the plug right in the middle, see what happens, plan for appellate -- Power failure just so that you know what impact it has, it is possible you can pull the plug and not lose any data these are things to keep in mind. I also have more questions. I hope you can refer back to this in the future, download the pages as a handout.

You can think about some of the questions on here like, do you monitor the threat of silent data corruption? Mrs. corruption taking place after the content has been copied. One example, if your storage filer, or storage systems are located next to a really loud device. Or perhaps really close to a highway even. There is a high variety of things that can change the way our data is written to different systems, the data could get corrupted just by moved or overheated, you might not know it unless you are checking, your institution may check only once a year, a whole file may be impacted by data corruption, that is a lot of data that you did not know until you talked. These are questions to lead through, talking about scenario planning.

Were cases things like that. All of this information you can read in much more detail, I have included a risk of references, from storage systems in low-level detail, to risk assessments, checks, these are all good articles, that I encourage anybody who has wanted more in-depth systems, this can spur you into more reading. If you have any questions I can take them. I appreciate you coming very much.

's -- Feel free to chat in your questions that you have for Jessica. I don't have any submitted during the presentation, I will give you a few minutes to chat them into the Chat Box, while I have you waiting for the questions, I will go over the webinars that we have coming up. We have a measuring America series, access in finance school data, 2 PM, April 17. GP has a better way to assessed April 19 2 PM, library and drive -- Guide to codes on April 20 at 2 PM.

If you are interested in any webinars, head over to FDLP.gov, sign up for any webinars you are interested in.

We will give it a little more question time, we will set out the Satisfaction Survey while we are waiting for questions.

We have a comment. This is a new area for me, very helpful thank you.

We also have another comment, I appreciate the eight approachability of your presentation they can be very useful to bring to administrators.

Ashlee has pushed out the Satisfaction Survey, if you do not have any questions, go ahead and fill that out. We will hang out in case anybody has any questions.

If you need anything further. You can contact Jessica at her email. [Audio disconnected-Please stand by while reconnecting]

[Captioner Standing By] Thank you for visiting WebEx, please visit our website.

[Event Concluded]

Actions Save as Text Save as HTML