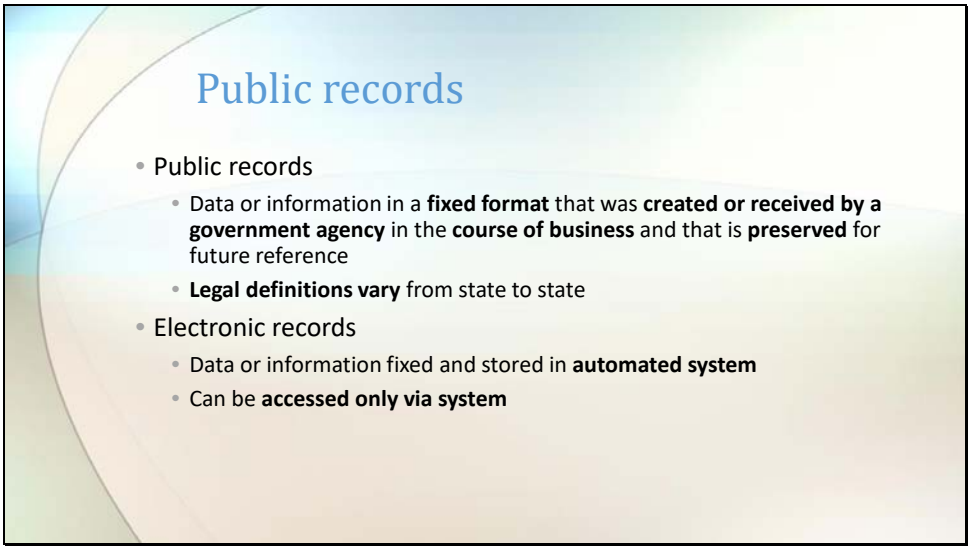
Slide 1

Welcome to Managing Digital Content Over Time. This training was produced by the State Electronic Records Initiative, in coordination with the Council of State Archivists. It was developed under a grant from the Institute of Museum and Library Services and based primarily on training created by the Library of Congress. It is to help archives and others who maintain electronic records understand the necessary steps of digital preservation. This is module 1, Identify.

Slide 2



Before we begin, let's take a moment to examine the differences between "public records," "electronic records" and "digital content." Archivists and records managers who work in state government archives are typically focused upon and work with public records, which the Society of American Archivists' *Glossary of Archival and Records Terminology* defines as: "data or information in a fixed format that was created or received by a government agency in the course of business and that is preserved for future reference." The precise legal definition of what constitutes a public record or government record varies from state to state, but virtually all public records laws focus on information that documents government business activities.

Within the government environment, electronic records are public records that have "been captured and fixed for storage and manipulation in an automated system and that requires the use of the system to render it intelligible. "[*SAA Glossary*, s.v. "electronic record"]

However, some state archives acquire not only state government records but also the papers of individuals and the records of non-governmental organizations. Other archivists and records managers work exclusively with non-government materials. As a result, in these training modules we're going to use "digital content," a much broader term.

Slide 3

What is digital content? Digital content is any content that is published or distributed in a digital form, including text, images, sound recordings, video, data sets, and software.

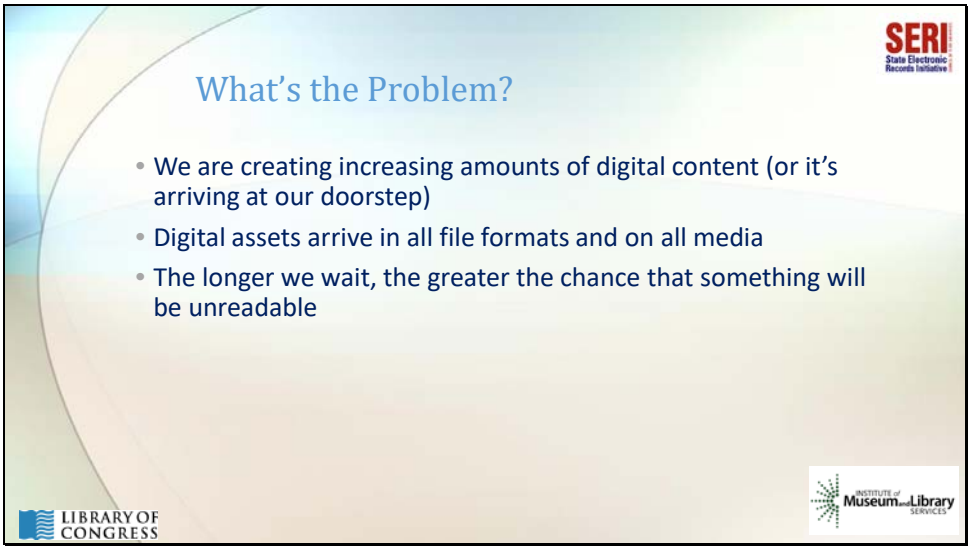
Anything we may encounter in a digital form is going to fall under the umbrella of digital content. This is going to encompass anything that comes our way and that we are going to have to think about preserving for a long period of time.

There are really two sources of this digital material –

- Those created from physical sources that we turn into digital items, such as digitization of maps or documents
- Those that are created digitally, such as items on our computers or digital photos

A lot of times our responsibilities include not only holding and preserving that digital content, but also making it available to the public. But this ever-increasing digital world is presenting us with new challenges. And what are those problems? …

Slide 4

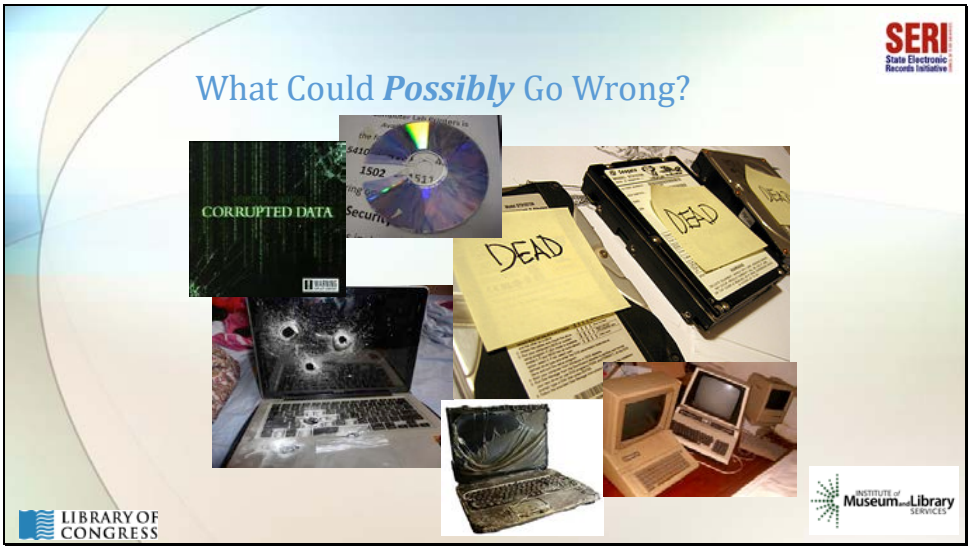


Everyone is

- **creating** digital content in lots of different formats
- **distributing** digital content
- **using** digital content

And **we are responsible** for managing digital content now or are expected to in the near future. The longer we wait, the greater the chance that something will be unreadable.

Slide 5

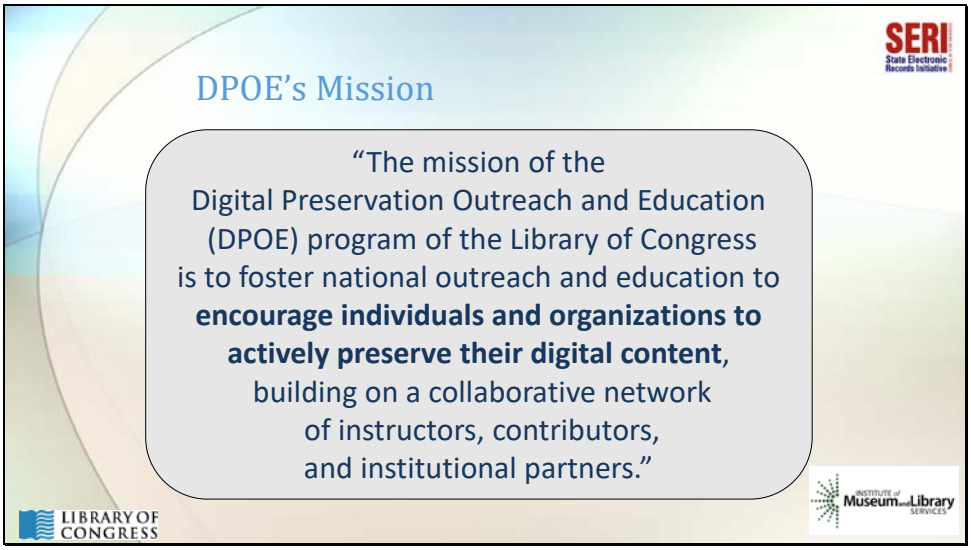


What could possibly go wrong?

Slide 6

Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.

Slide 7



Part of the curriculum in this program was developed by the Library of Congress as part of their Digital Preservation Outreach and Education program. DPOE is part of a national effort

The first 4-day National DPOE Train-the-Trainer workshop was held at the Library of Congress in Sept 2011.

Slide 8

The next six slides use an extended visual metaphor to explain the six modules for this training.

To look at the process metaphorically, the slides describe these steps in terms of getting oranges from the processing plant to market where people can access them.

The orange processing closely mirrors actions that you take with digital content.

Here the worker is grading the oranges, identifying their quality and variety.

Slide 9

Here a man is selecting one orange out of many.

The selection process requires that one choose only a small percentage of your content for full long-term storage to keep management costs in control.

Slide 10

A man is bringing oranges into refrigerated storage.

How are you going to store the oranges – sometimes for a fair amount of time – so that they make it to the public in the same state you packaged them?

Slide 11

A women is packing oranges to protect them from damage during transit.

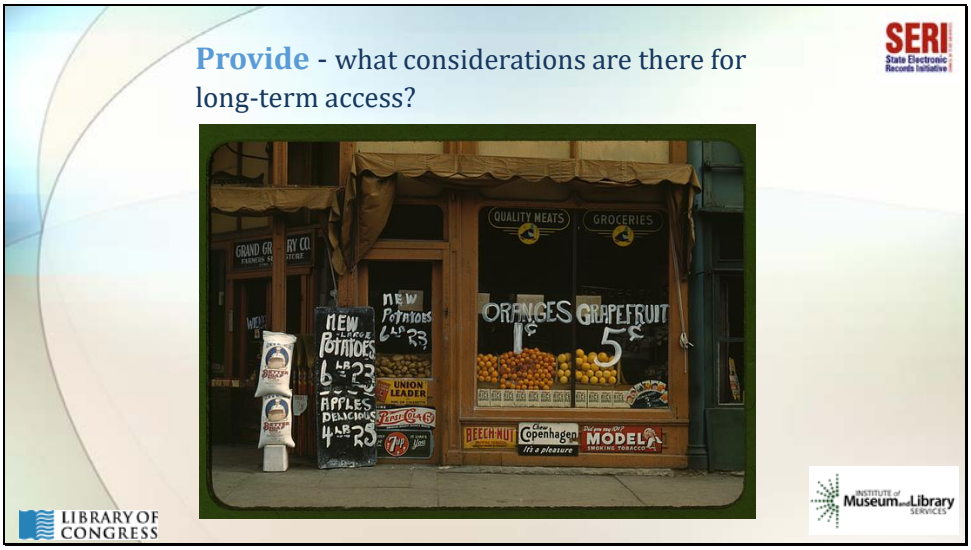
Aside from storage conditions, how are you going to protect them from any damage that may occur?

Slide 12

A man is stacking and organizing content in storage in a pattern to fit the most in the truck. Here you see all the elements that need to be supported for orange storage: labor, equipment, space, packaging, and of course basic organizational skills to get it to all work together.

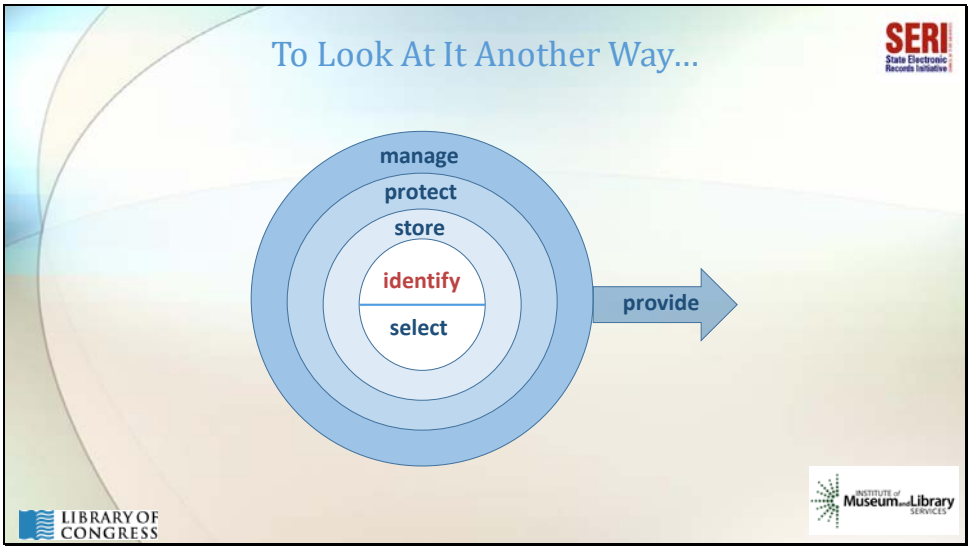
Note: **Admittedly, this is the most tenuous part of the metaphor, as management is so abstract.**

Slide 13



Oranges for sale at a store, the ultimate goal of the process.

Slide 14



Modules are not designed to necessarily provide specific technological solutions but they are designed to make sure the right questions are being asked about each stage – planning, policies, staffing, and advocacy.

Identify and Select are in the center – they repeat over time and often happen at the same time. It is here that we identify new/additional content and then select the portion to preserve.

Store the selected content.

Protect what is stored.

Manage it over time (preserve it) – including policies/rules for identifying, selecting, storing, protecting and providing content. Providing preserved content over time (long-term access) relies upon good management over time

Slide 15

So let's get started…
As stated earlier, the volume and kinds of digital materials we create or inherit are growing.

Much of it is useful and even necessary for our work, but much of it is not. Think of the string of e-mails created as people go back and forth discussing a topic. Or different drafts of a document. How much of that is really worth saving for posterity? The Identify stage helps us figure out what content we have, so we can determine what needs to be kept.

Good digital preservation requires an explicit commitment of resources in terms of staffing, finances and cooperation from others, which - for most organizations - means planning ahead. If you don't know the extent of the problem, you don't know what resources you need.

The first step in planning for digital preservation is to know where you stand with regard to your digital assets.

Slide 16

Identifying content is a foundational step for digital preservation. It helps determine the extent of what you may need to preserve.

The orange section is what the inventory will cover, so it's the first step in whittling down the vast amount of data that an organization creates.

The inventory process is also a powerful way to raise awareness of the problem within your institution.

Slide 17

In order to plan effectively for the future, you need more specific answers to several questions:

- What content do you have now? (this includes the files you are actively managing, as well as the stuff that's hiding on someone's hard drive)
- What content do you know you will have in the future? (digitization project that is regularly producing image or audio files)
- What content might you end up having? (potential large donations; content stored with a 3rd party – what happens if they go out of business, or you can't pay the required fees)
- What content must you have? (for a State Archives – if we know that an agency has gone mostly paperless, we should be receiving electronic records to document the activities of that agency. We may have legal/ethical requirements to keep particular records, like donor records. If they are digital then we HAVE to manage and preserve those files)

Again, an inventory can help raise awareness within your own institution in terms of where you are right now and what you will need in the future.

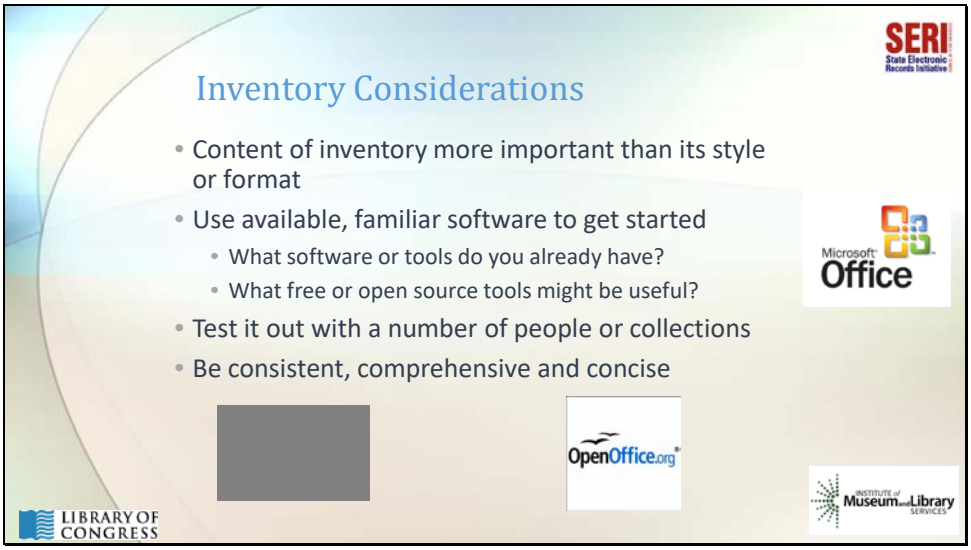
Slide 18

Does your institution have an inventory of your digital content?

If you answered “Yes” – Fantastic! You have a great head start. Now, do you know where the inventory is stored? When was the last time it was updated? Is it complete – or is it just an inventory of digital content of a particular type, or on a particular server?

If no, do you need to get permission to start an inventory project?

Slide 19



The content of an inventory is more important than its style or format. Don’t let implementing the software become the focus.

Use software you know and have available.

Stick with a single format; don't change once you've decided on it.

Be consistent, comprehensive, and concise.

Slide 20

An effective inventory is …

Scalable - You will be adding content to it over time

Available to others: This should be a group project within your own institution. One person can manage it, but there is no way one person will know where everything is and what it consists of. This is a tool to ensure you are all starting from the same place and collecting the same information.

Usable – anyone should be able to pick it up and understand it.

Current - This is not a static thing. Make sure you track it when it's updated. This helps keep it a "live" document. This is also REALLY hard.

Electronic – Easier to work with, you don't have legibility issues. It's also easier to report off of – showing change over time.

Documented – Again, easier to keep track of what you have over time, report off of and gives you something to show someone when you would like to request resources.

Slide 21

It's key to make your inventory answer these questions by targeting the inventory at the core mission and functions of your organization, and to also consider the value that preservation adds to the organization.

Slide 22

Here's an example from the California Department of Fish and Wildlife - ACE II – which means Areas of Conservation Emphasis. It provides data to help guide and inform conservation priorities in California.

Slide 23

You decide what details you need, such as…

Extent - How much is there? Both within your organization that you need to manage and in another organization that you might be getting records from.

This will help you identify digital content that may have made its way into your organization through different points and currently lives in disperse locations. Perhaps some came from a state agency, some came in with a manuscript collection, some came in from a vendor/contractor, or some from a digitization project.

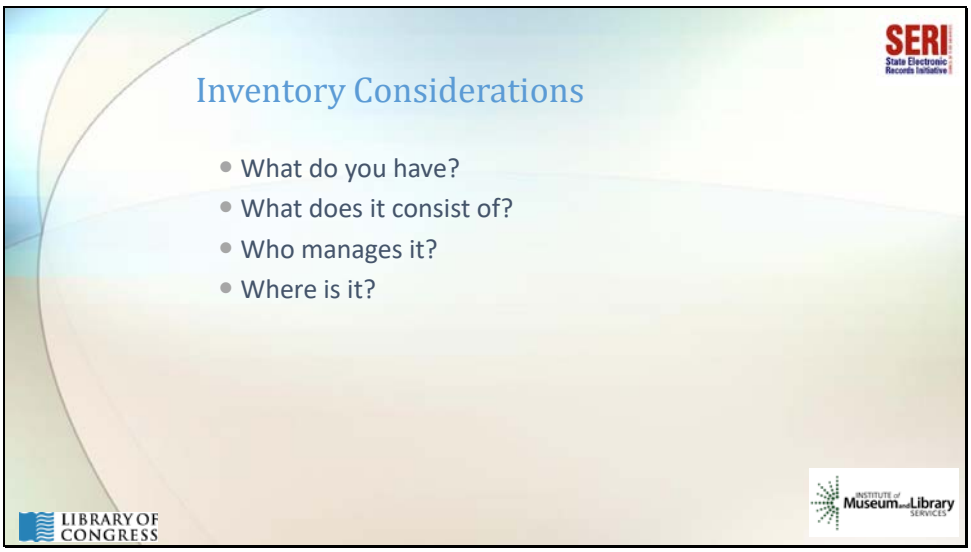
This tool can also be used at the contributor level when you are talking about what they have that should be transferred to you.

Nature and location – This answers the questions of …..Where is it? Can you access it?

Resources - How much money/manpower can you devote to collecting this information?

Timeframe - How much time do you have to do this?

Slide 24



Work at the collection or series level, not the item level. What is the familiar title for the collection?

Provide a brief description of what is in the collection.

You are collecting information about items that are known and may be in your catalog + items that have come in your door that are waiting to be dealt with + items that are being created (digitization projects) + things you may not even know about yet…

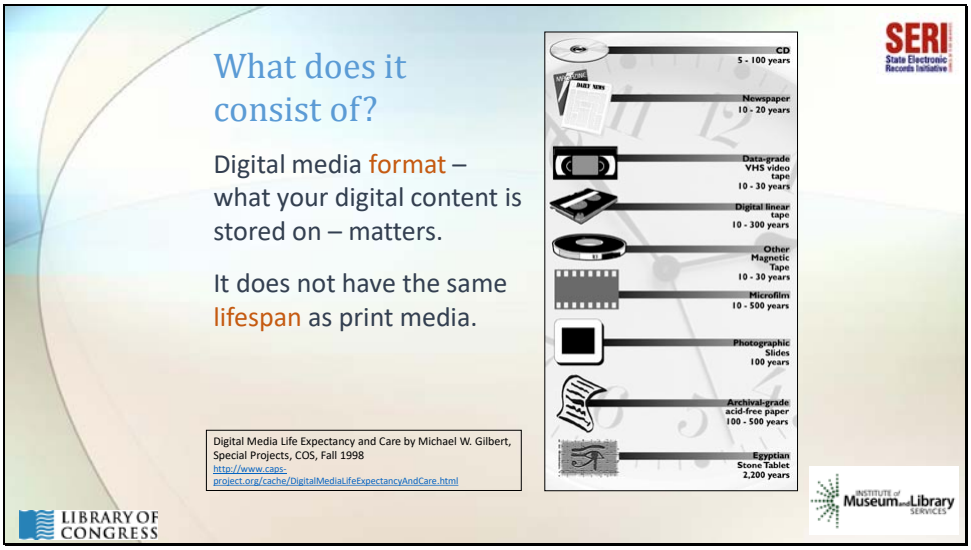
Slide 25

First identify records in broad groups. You may not have all types of content.

Slide 26

Then identify file formats. The format types will depend on what content you have. Don't worry about formats you don't have (or don't have yet) because your inventory will narrow down your scope of concern.

Slide 27



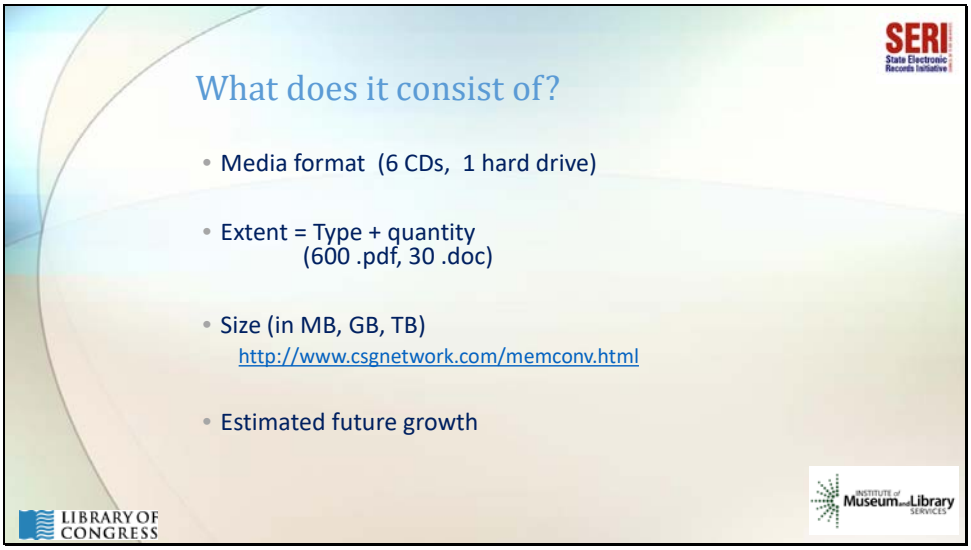
It's a good idea to note the media type, or what the digital content is stored on, since some media types last longer than others. Digital content on more fragile media might be a higher priority when you get to the selection phase.

What if, for example, the only final copy of a report or document is stored on somebody's flash drive in a drawer at their desk? Not only is the document unavailable to those that might need access to it, but if something happens to the flash drive, it will be lost forever.

This chart is nice because it really illustrates how much shorter the lifespans of digital formats are compared to the physical formats we're more used to, although it still has the lifespan of CDs as being 5-100 years, which is pretty optimistic. Five years is probably closer to the truth.

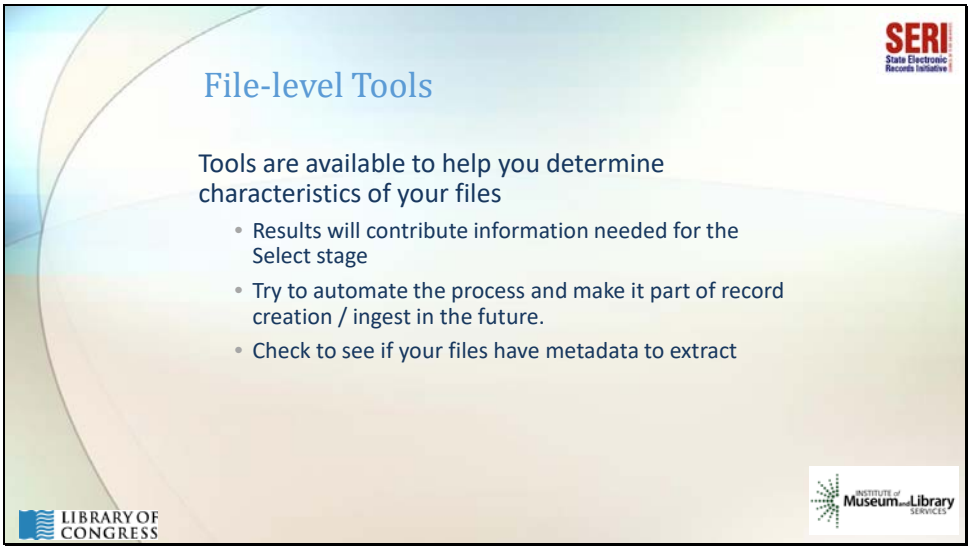
Software and operating systems: identify operating systems required to access files. Document specific software programs needed to read files.

Slide 28



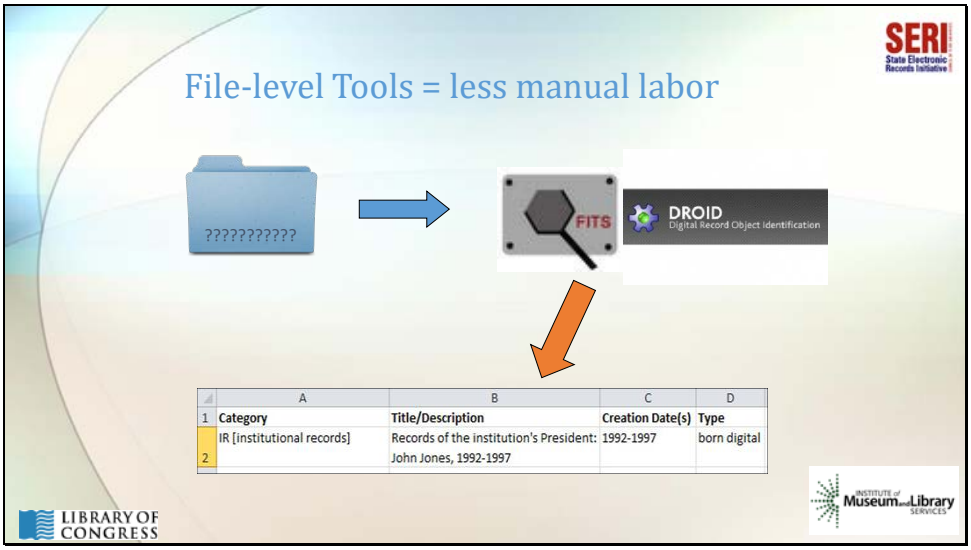
Count what you have, the number of physical media, how many of a given file format, and file sizes, to obtain a grand total.

Slide 29



To assist you in the inventory task, you may use tools to help automate the process. You can also wait to use tools until the selection step.

Slide 30



File level tools are free, not too hard, and pay off big when used correctly. Take an unknown file directory and use the file level tools to extract an inventory and get technical metadata and format verification on the files as well.

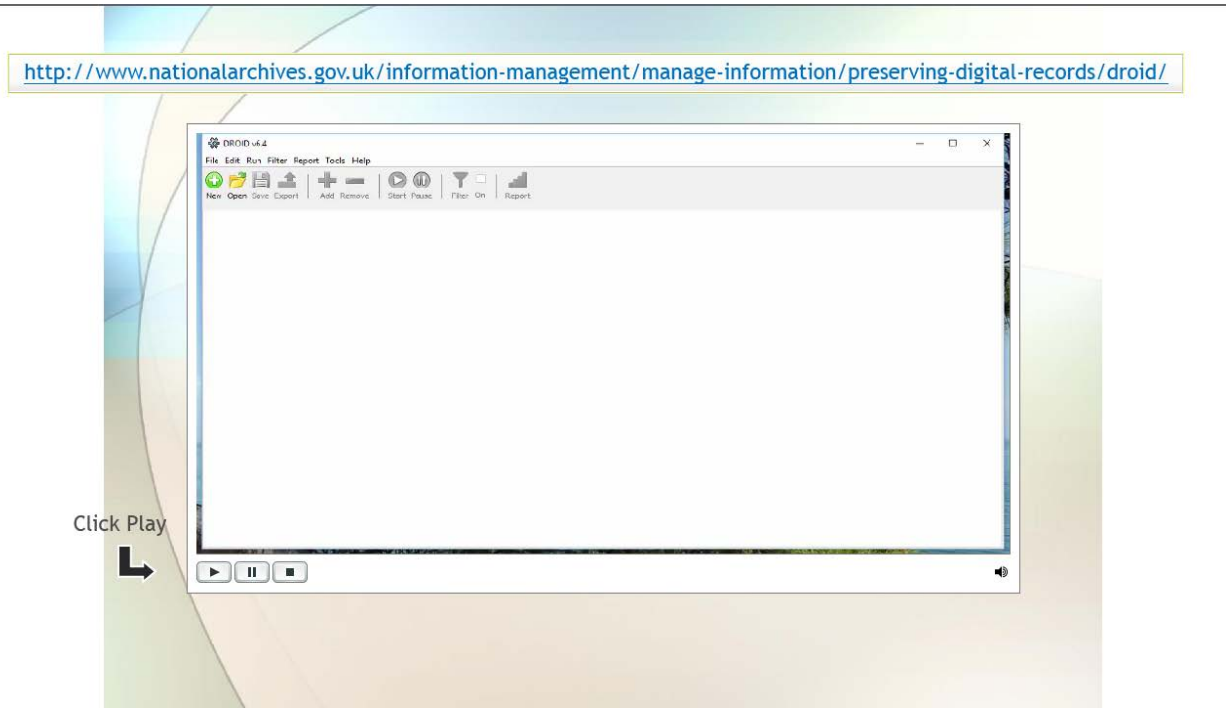
Tools include:

DROID (which stands for Digital Record and Object Identification). It is a software tool developed by The U.K. National Archives to perform automated batch identification of file formats. If it can identify it, DROID will tell you what versions you have, their age and size, and when they were last changed, as well as help you find duplicates. This is a stand-alone tool.

The File Information Tool Set (FITS) was developed by Harvard and wraps a number of open source tools (including DROID) into a package that can be used to more broadly identify, validate and extract technical metadata for a wide range of file formats.

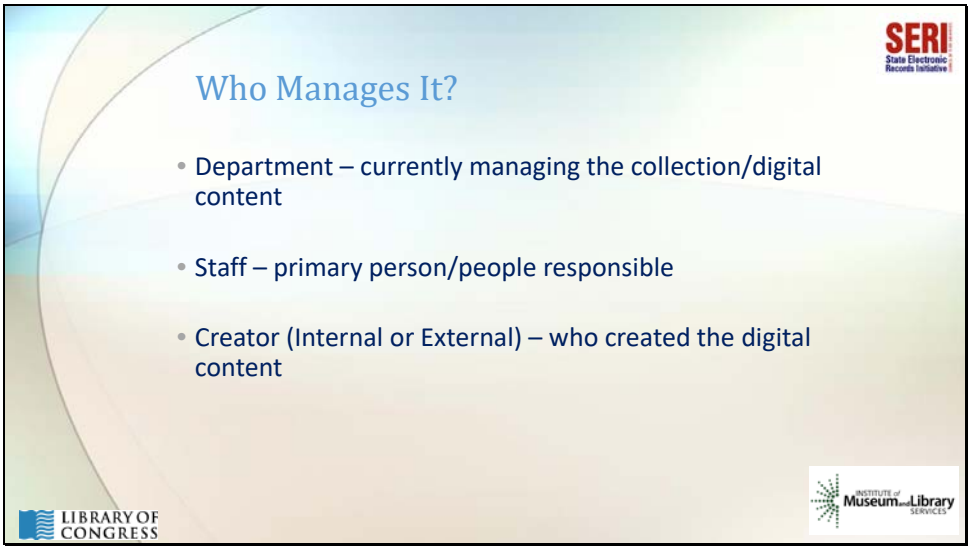
Identifying your file formats helps you to manage your information more effectively. It helps you to identify risks (and therefore plan mitigating actions). It also helps you to save money, for example by supporting data reduction. On the next slide click 'play' to see a demo of DROID.

Embedded Video



To run Droid, first download the files from the website, unzip the package and find the droid.bat file. When you run it, this window will open up. First you want to click the Plus button to create a new window. And then you want to add files to the window, such as these, or you can navigate to a different location. Click ok. So the folder is listed here. Then you want to click the start button and it will analyze what's is in that folder. If you click the little Plus by there then you can see the data that results from it—so the file extension, size, last modified, and format information, including versions. This will help you determine what types of files you have.

Slide 31



Moving on – The content we are trying to wrangle rarely lives in one location with one person. Others know and can provide information about these records:

Your Department – You can call this anything that's appropriate for you and your institution but it's who is currently managing the content.

Staff – Who is your contact if you have any questions?

Creator – so that you can go back to them with any issues if you have questions about content or format.

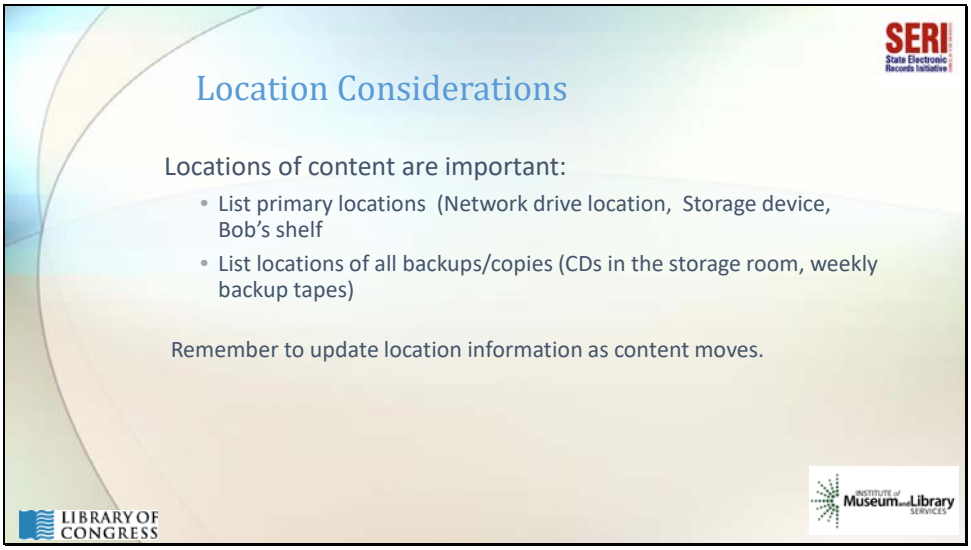
Slide 32

The location of the content should stay current. Documenting where the content is stored will be helpful when the content needs to be moved – such as, be deposited with the archives, or moved to a new server.

As you can see, there are many places digital content can hide! You should make sure to specify the location of digital content in your inventory. Some things you will want to consider:

1. How will you specify whether content is located online (meaning on your computer hard drive or a network server), or offline (meaning stored on some removable piece of media, like a CD or flash drive)?
2. Location in storage system. What is the path on the device where the files can be found?
3. Keep in mind that you will need to update the inventory whenever the content moves. If you get too specific you might spend all your time updating file locations.

Slide 33

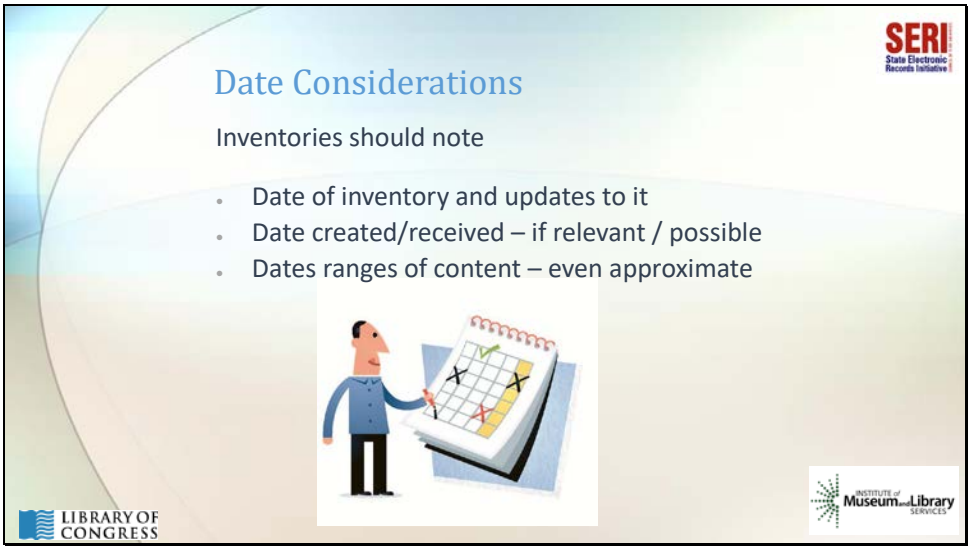


You'll need to strike a balance between being clear enough to find the content you're looking for, without going to extremes.

It is best to work at the collection or series level as opposed to the item level.

Be sure to include locations of backups and copies, and identify them as copies instead of additional new content.

Slide 34



There are several dates that you will need to keep track of in your inventory:

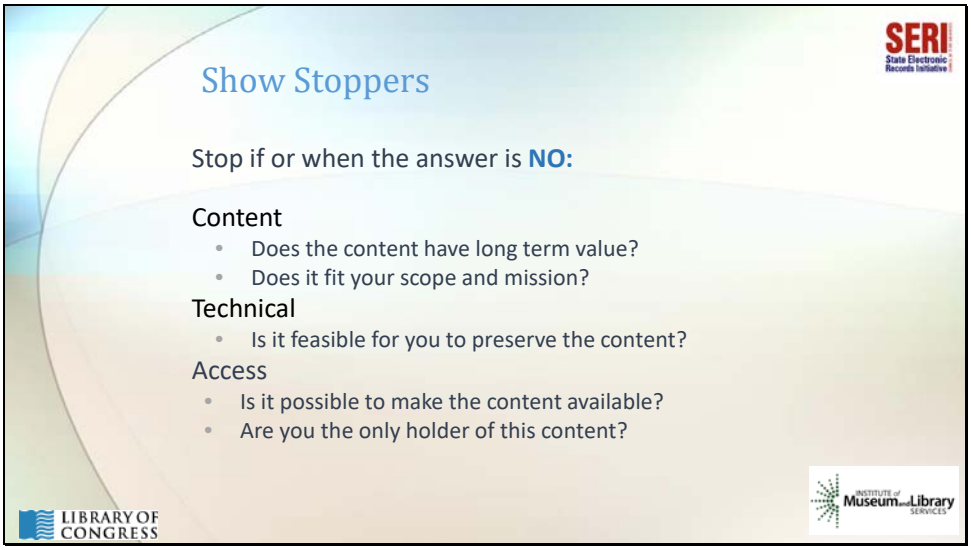
Date of Inventory – Dating the inventory and noting when it was last updated will help people using it and may ensure that it is updated more regularly. The inventory should stay current with your content as it changes and grows.

Date created and/or received – note how old a file is; since it's harder to migrate and open older files, you may want to prioritize these when you get to the selection process.

Dates covered in content – identify the date ranges of the content, for example, files created between 1998 – 2004.

The date of the files and the dates the content covers may sound similar but are actually very different. A creation date could be much newer than the content date (such as recent digitized images of historic photos). These dates will help in using the inventory and in prioritizing your review of the content.

Slide 35



Even if something fits your desired criteria, it still might **not** be reasonable for you to select it. You can use a decision tree or list of questions to help you decide what's practical to preserve.

You've already considered the **content** in view of your selection criteria. And you should already have answered 'yes' to both of these questions to continue considering the materials you hold.

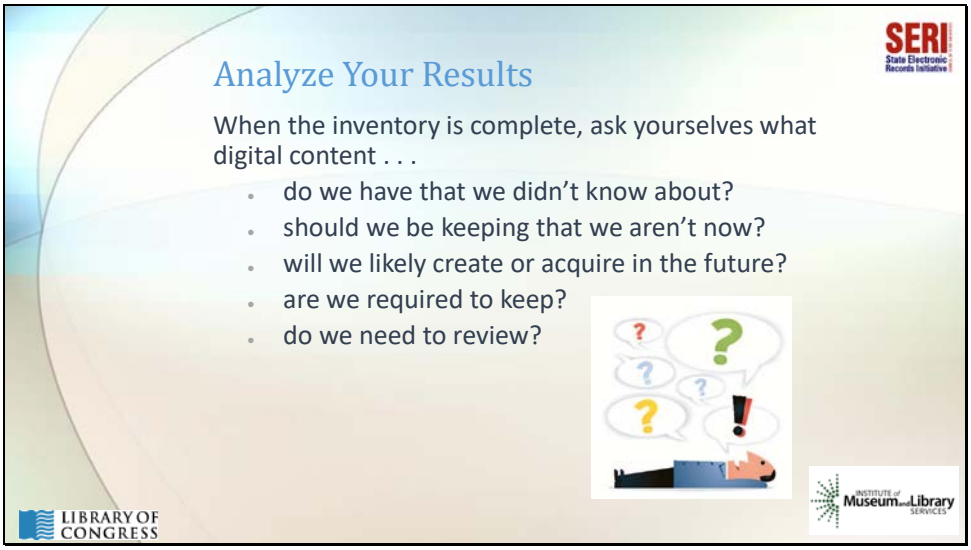
–does the content have long term value?
–does it fit your scope and mission?

Next you need to consider **Technical** issues:
–is it **feasible** for you **to preserve** the content? [Is it a **"digital time bomb"?**
–Some formats are a challenge to preserve, such as video or time-based media.
–Some may be too damaged to preserve. Do you have the skills and resources (either to undertake the preservation yourself or to outsource)?
–Some types of material may require far more expertise and resources than you have available.

And **Access.**
Even if we're not making it public, how useful is a server full of digital content that is safe, but that we can't access?
We need to ask:
–is it possible to make the content available over time?
–Are you the only holder of this content?

If it is not feasible to preserve the content, and not possible to make it available and usable, then it probably shouldn't be included in your selection –especially if you know you are not the only holder of this digital content.

Slide 36



The inventory allows you to identify the digital content you are already preserving and you can now ask yourselves:

1. What digital content do we have now? – this includes content you are currently managing, as well as the stuff that is taking up server space that you might not even be aware of. Is what we're doing optimal? What efforts are we already making?

2. Are there gaps in what we are preserving, such as things we should keep that we are inadvertently not, like important electronic correspondence?

3. What areas of our digital content do we know/expect will grow over time? Does your office regularly generate audio files during the course of business? If so, how much can you expect to add in the next few years? Are you responsible for an ongoing or upcoming digitization project? You will need to plan to manage those files as well.

4. Which items are dictated by Retention Schedules? Organizational/government requirements? Are there transfer agreements that state you WILL hold and manage a digital collection? Do you have gaps in what you should be collecting – are there things that you SHOULD have but don't have accounted for on the inventory?

5. Is there material that's been saved not because you wanted it but because it never got deleted? e.g. many draft versions of a document? What digital content can you discard?

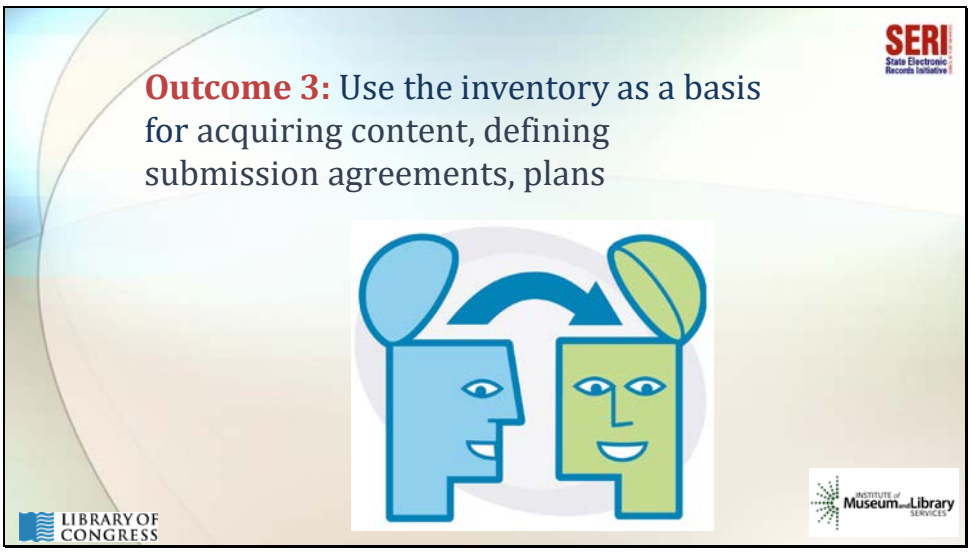
Slide 37

Identifying records has several outcomes. It identifies potential digital content you may need to preserve. Planning ahead is always better than reacting in the moment.

Slide 38

Treat the inventory as a management tool that grows as your program grows. Add to it regularly, so you aren't overwhelmed by a big project every few years.

Slide 39



Use the inventory as a basis for acquiring content, defining submission agreements, and plans.

Working with an existing tool helps to define what you are looking for and what information you need about that collection when you develop your transfer agreements. It provides your content creators some context and ensures you get the information about the collection that you need in order to manage it.

Slide 40

Use it as a planning tool, to prepare staff, for training, or annual growth. Keep it in good enough shape to share it with those who fund you.

Slide 41

This completes module 1, Identify. If you are using these modules in order, the next one is module 2, Select. For additional resources on electronic records preservation and management, please visit the State Electronic Records Initiative webpage. This link is on your screen.