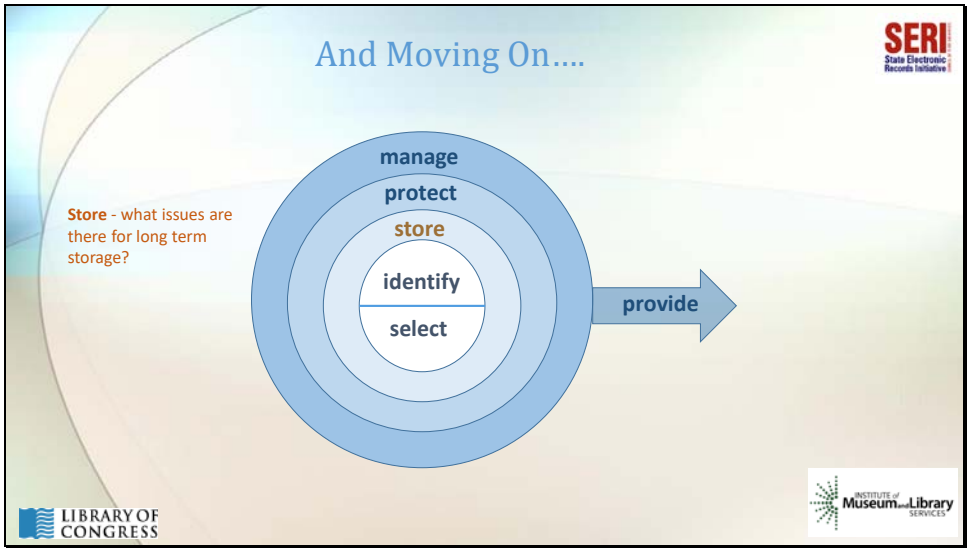
Slide 1

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 designed to help archivists and others who manage digital content understand the necessary steps of digital preservation. This is module 3, Store.

Slide 2



Looking at this another way…..remember that the identify and select processes pretty much happen together.
Now we are taking a step back and asking ourselves – how are we going to store this content for the long term?

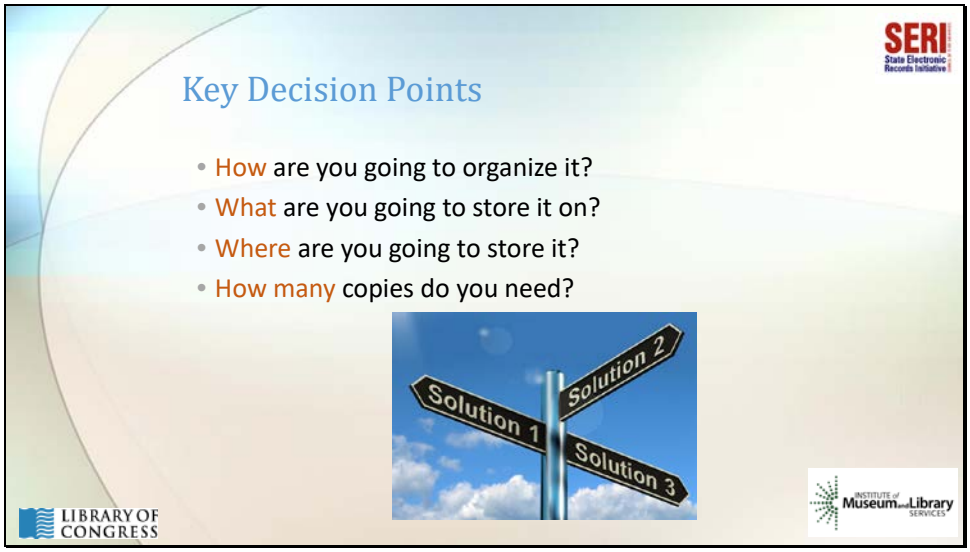
Slide 3

So, More specifically, here are the objectives for this section.

1. I won't be telling you which storage method you need to invest in, but I will help you to develop a better understanding of long-term storage requirements and possible options for meeting those requirements.

2. And, we'll be looking at some factors that you need to weigh in developing long-term storage management practice and policy.

Slide 4



So, once the selection process is complete, you need to decide:

HOW you are going to store and organize it so you can find it?
WHAT are you going to store it on?
WHERE you are going to store it?
HOW MANY COPIES are you going to make?

Slide 5

Starting with the ”how” question – This will involve taking a look at things like file and folder naming, folder organization, etc.

As you are going through the inventory, you will find things in many places and named in many different ways depending on who worked on the item. Establishing some consistency in how files are named and organized will help you manage and find the items over the long-term as well as help prevent accidental overwriting of key documents.

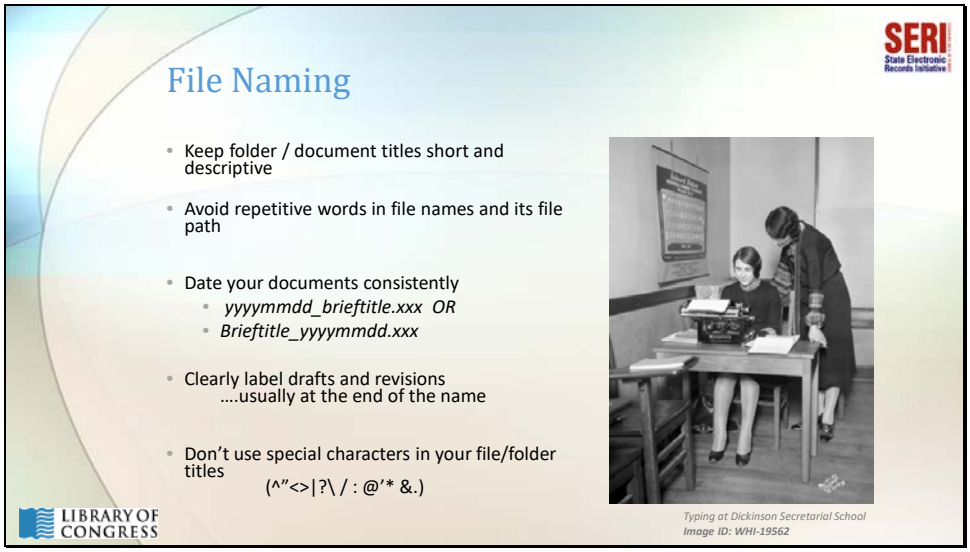
Standards – Need a baseline so that everyone knows how to name items as well as how NOT to name them, OR where and how items will be stored

Finding – were the minutes saved as April Minutes, 04 minutes, Board minutes, recent minutes, etc

Accidental Overwriting – for example: photos from a digital camera, meeting minutes/agendas

Generally speaking, avoid special characters in file names. While your system may accept them now, there is no guarantee these characters would move to a new system. These characters are often used for certain tasks by certain programs. Using them could lead to file loss or file errors.

Slide 6



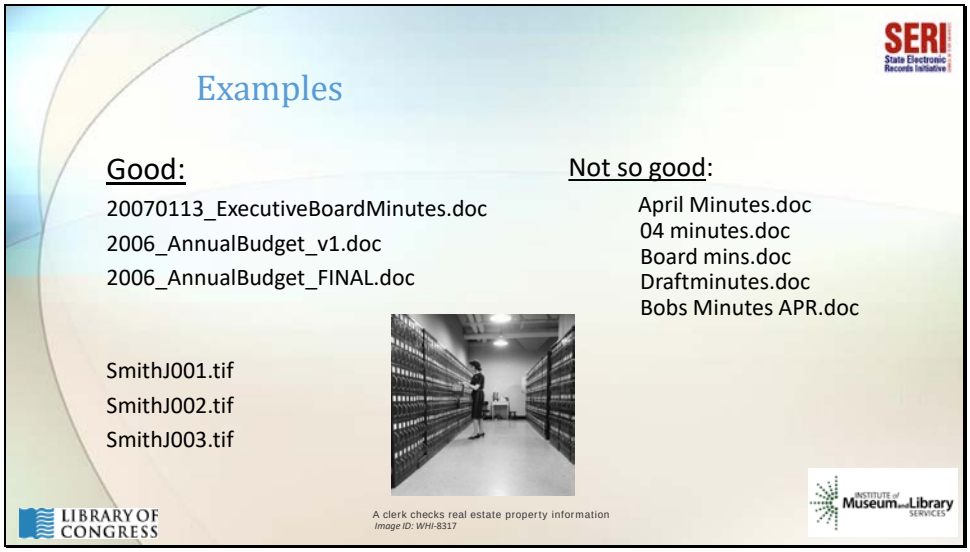
Keep folders and document titles short and descriptive –
If my record is a file name with 167 characters, while really descriptive, it is too hard to work with. Can't read the entire title in a file list and can't copy it if it's buried in several layers of folders.

We tend to name things in ways that make sense to us at the time, but this is not handy for long term preservation. You need to name things in a way that will make sense 20 years from now.

Has anyone inherited files from previous employees or projects – do they make any sense?

Think of these examples of folder titles: "My stuff" "Important" "To Read"

Slide 7



Here are some good and bad examples.

Good – Board Minutes + Series of images by photographer John Smith:

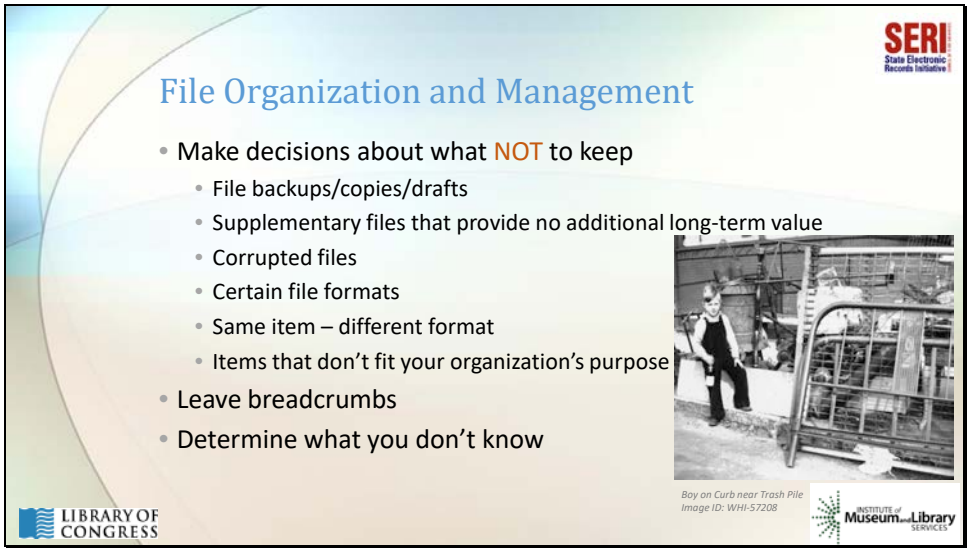
Slide 8

As you are creating your inventory, you are likely to discover a lot of really simple places where you can clean up the files you are reviewing.
Co-locate – It's OK to move things around if it makes sense to do so.

Don't Bury – If you have several layers to hunt through, it can be really hard to find anything – Shallow is better. It also makes the file path shorter (fewer characters)

Purge – Unless there is a <u>really</u> good business reason for keeping them.

Slide 9

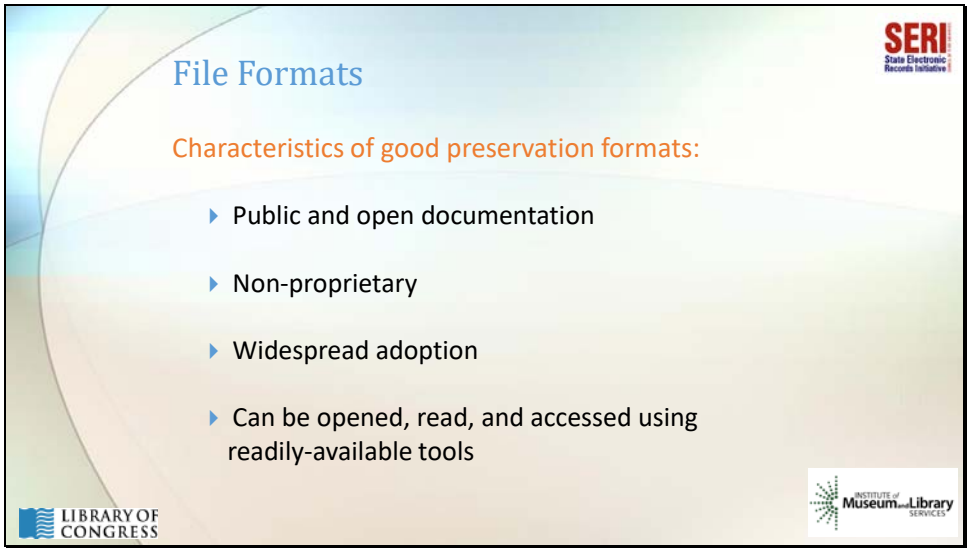


Some things you should not keep:
File backups – for example: Speeches had multiple drafts: final + copies in several different font sizes
Supplementary files – folder of images that were used in a power point.
Files you can't open – Corrupted
Formats – you may receive Word and pdf – May not want to keep both.

Breadcrumbs – it's OK to leave "sticky notes" (AKA "READ ME") files in folders. They can give a brief description of contents, retention schedule, any naming conventions used.

Determine what you don't know – unknown file formats, files on old media (floppies), password protected, etc.

Slide 10



Here's what you should look for when you are thinking about preservation formats – or even just thinking about file storage or file creation.

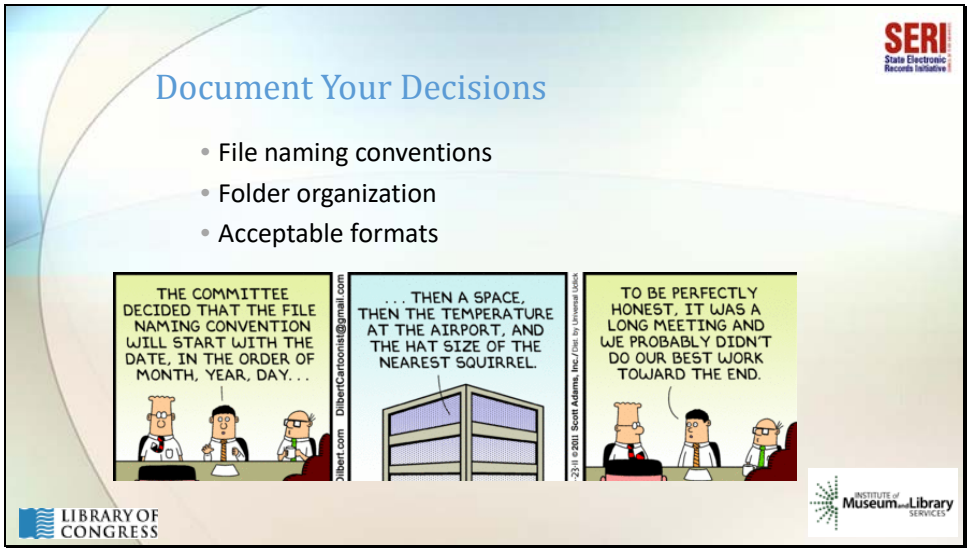
Public and open documentation - Ensures future access

Non-proprietary

Widespread adoption - Lessens risk of sudden obsolescence

Can be opened, read, and accessed using readily-available tools

Slide 11



Once you've decided how you want to handle file naming issues and have made file management decisions – Create documents standardizing your file naming conventions, folder organization and acceptable formats.

Write down what you know. It doesn't have to be long…..

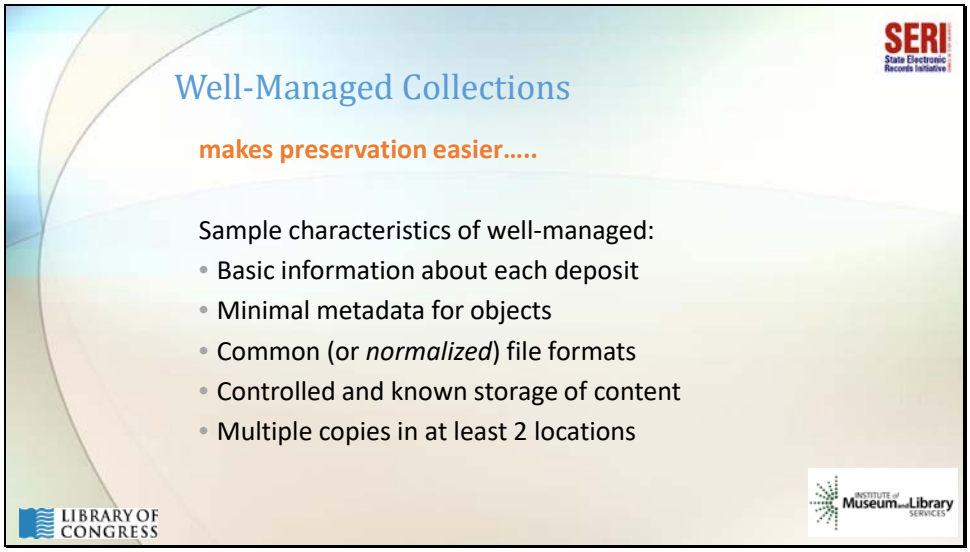
You can distribute it in your organization – post it on an intranet, place it in a procedures manual

WHY –

- You will not be the only keeper of the information. (If you aren't here to ask, how will they know?)
- It will help others who may be helping you with the inventory
- You can hand it out to organizations or departments you receive information from
    - Tell donors, "In order to better manage our files, we will accept these file types and formats, they will be named this way. Do not give us password protected documents"

You don't have to organize and fix everything, but you do need to give other people the tools to help you.

Slide 12



Once you have a handle on naming and what you will or will not accept, you can broaden your scope and start looking at your content at the collection level.

This is a list of things you want to be true because without them, you may have a really hard time implementing a system to manage your digital content over time. It's not the system's fault (necessarily), the digital content hasn't been prepared well enough for preservation.

Regarding Basic information – Your inventory should answer most of this for you. This would be information like date received, donor name, or collection title

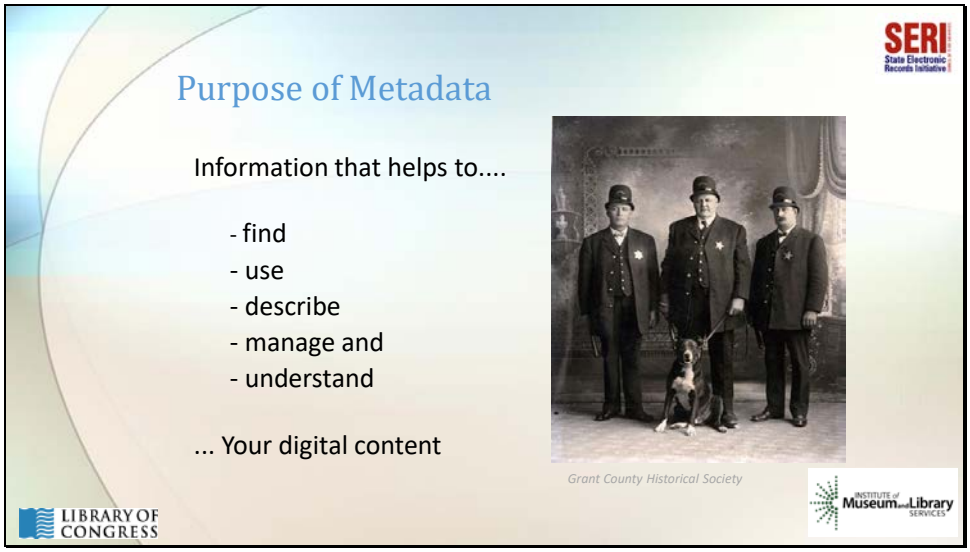
Minimal Metadata for objects – This is entirely up to your organization. What information do you feel is necessary for preservation and access? There are standards. For example, PREMIS (which stands for Preservation Metadata Implementation Strategies).

Normalized - Are you using common (or normalized) file formats?
TIFF or PDF are very common file formats and will be around for a very long time.

Controlled and known storage of content - If you know where your copies are, and who has access to those files, that makes the content or collection easier to manage

So Lets take a look at METADATA

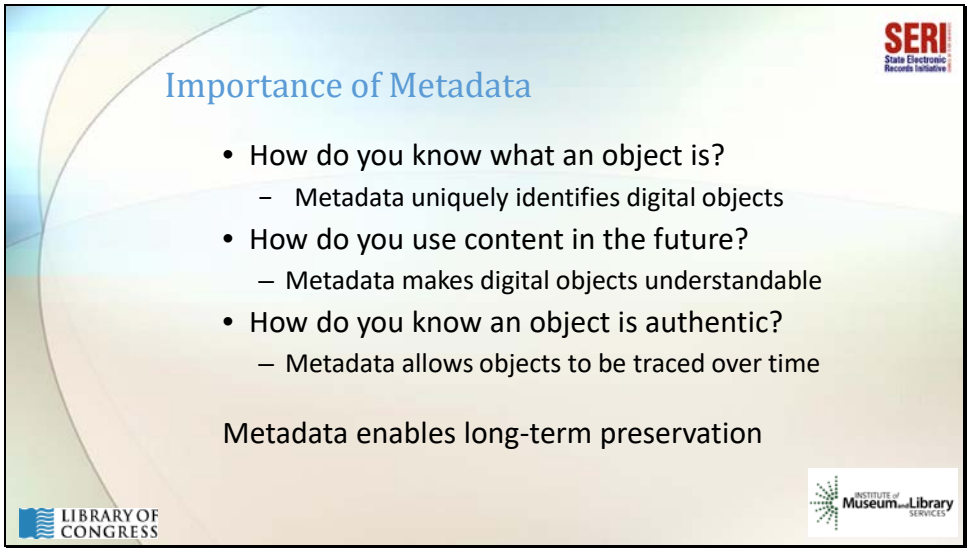
Slide 13



Metadata is defined as "data about data".

This is not a new concept – you run into metadata everyday with physical objects – where the data is about a “thing”

Slide 14

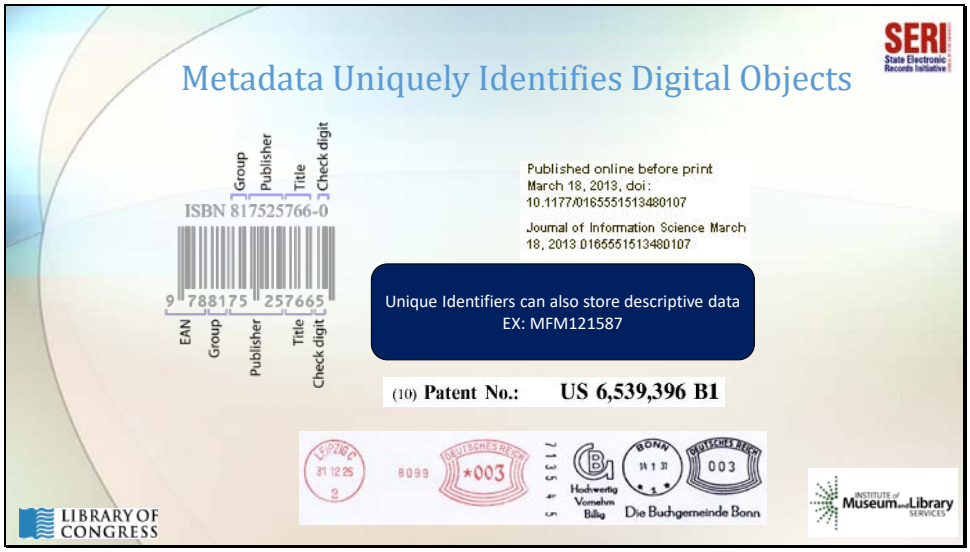


Why is all of this important?

Metadata is essential for preservation. The archival community has yet to identify a single, very standardized definition of preservation metadata; however, it should include all the information needed to manage, find and use digital content over time – but what that exactly means is open to discussion.

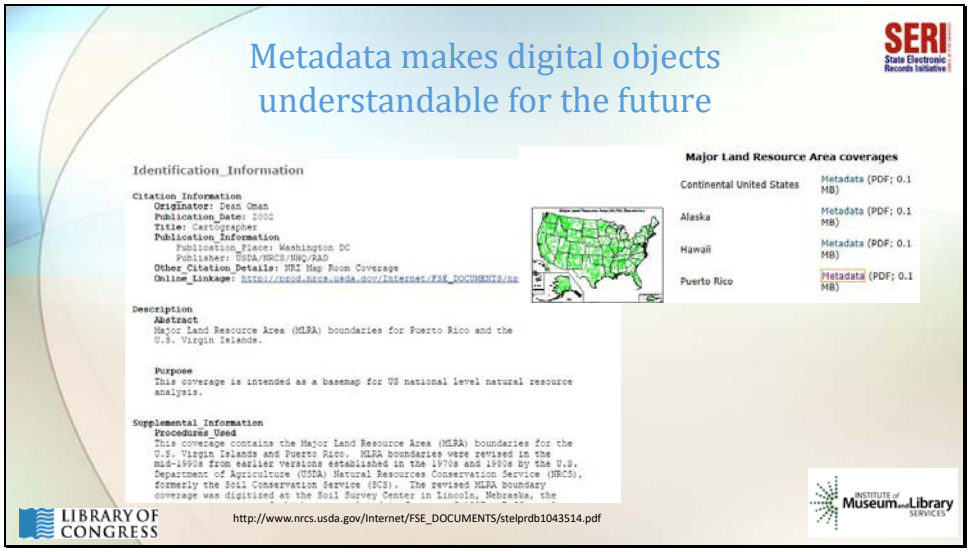
In depth discussion of metadata is a more advanced topic than is intended for this module presented to a novice audience. You should be aware that metadata is essential for preservation of digital content and there are some basic steps to gain control of digital content using metadata. You might want to read about PREMIS, but that is an advanced topic.

Slide 15



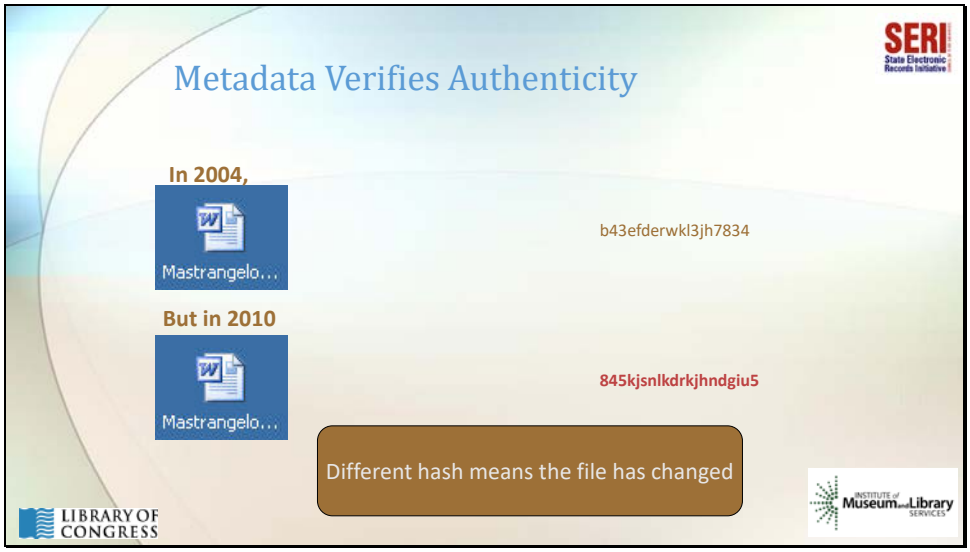
Simply put….metadata uniquely identifies digital objects.

Slide 16



This is from the USDA Conservation Service - metadata for Puerto Rico's Major Land Resource Area coverages.
It also contains 6 pages of information about Revisions, References, Related data, Access and Use constraints, cross references, etc

Slide 17

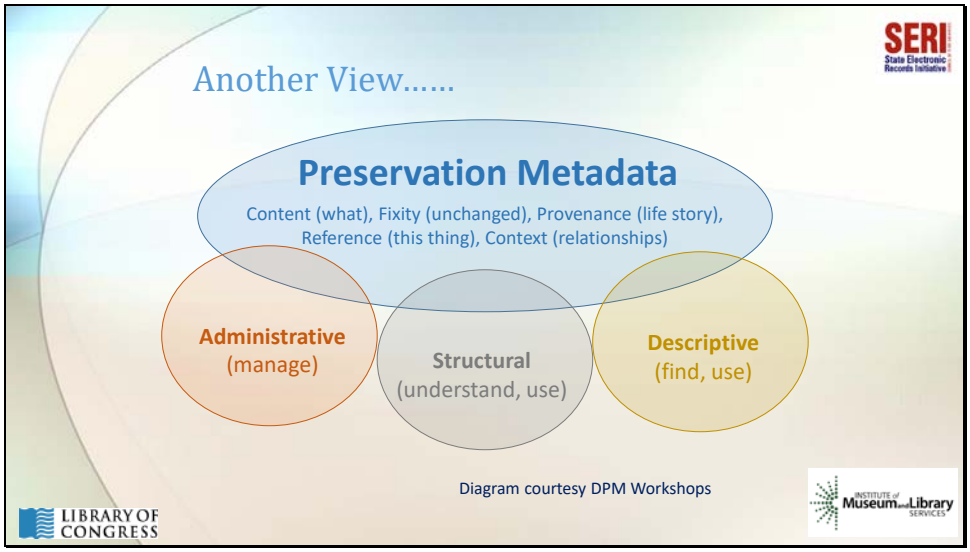


Beyond a file's name and icon, how do you know a file remains the same over time?

Hash functions use one-way encryption to turn files into strings of characters that uniquely identify the bitstream. If a file is unchanged, it will encrypt to the same hash.

We will be talking about this in more detail in the next module.

Slide 18



Administrative, Structural, and Descriptive are the most common categories of metadata defined for digital collections. These metadata categories were initially defined by the 1996 *Preserving Digital Information* Report then incorporated into OAIS.

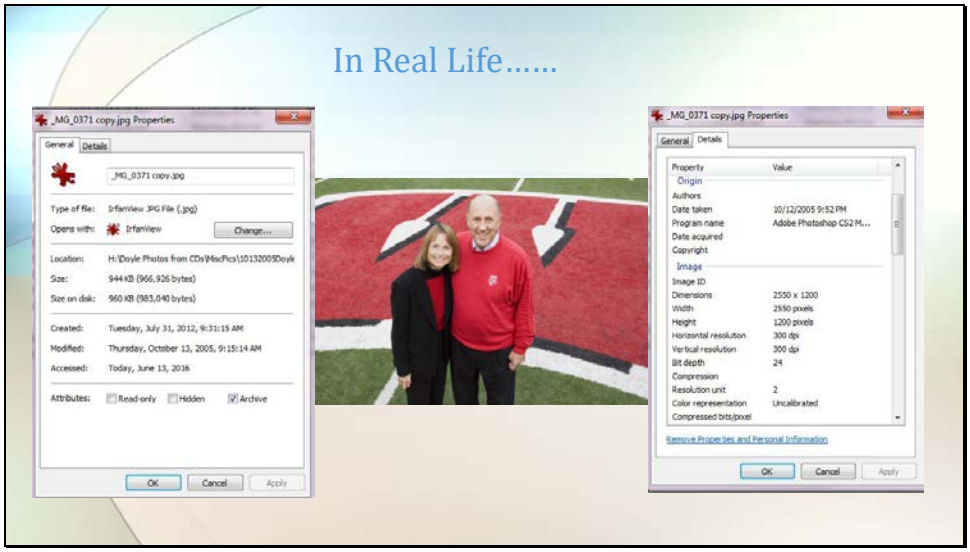
Some people would argue that preservation metadata is a subcomponent of administrative, but however you look at it, these are all types of metadata that are needed to make your object accessible and preservable.

Descriptive metadata is mostly the bibliographic information about an object such as title, author, subjects, keywords, publisher

Structural metadata provides a description of how the components of the object are organized; (single object, multi-page object that needs to go in this order, a.tif is the master, a.jpg is the access file)

Administrative metadata is the technical information you need to manage your content (where is it, format, size, rights information), Compression (used for optimized storage and delivery of digital object), Extent of master file (pixel dimensions, pagination, play time), Creation hardware (scanner or digital camera name and model), Operating system, Creation software.

Slide 19



In these examples you can see some technical (administrative) metadata within the property files of an image which should look pretty familiar to you.

General properties are on the left and show a lot of the administrative data….. file type, location, size, created, modified, accessed

The second image shows more detailed properties and you can see that you can scroll down to find more information on the image. Most of this is machine generated metadata and provides details about when the picture was taken, created, its location, size, etc.

Descriptive metadata —Description about the photo name, and subject could be manually added in the details tab if you wished.

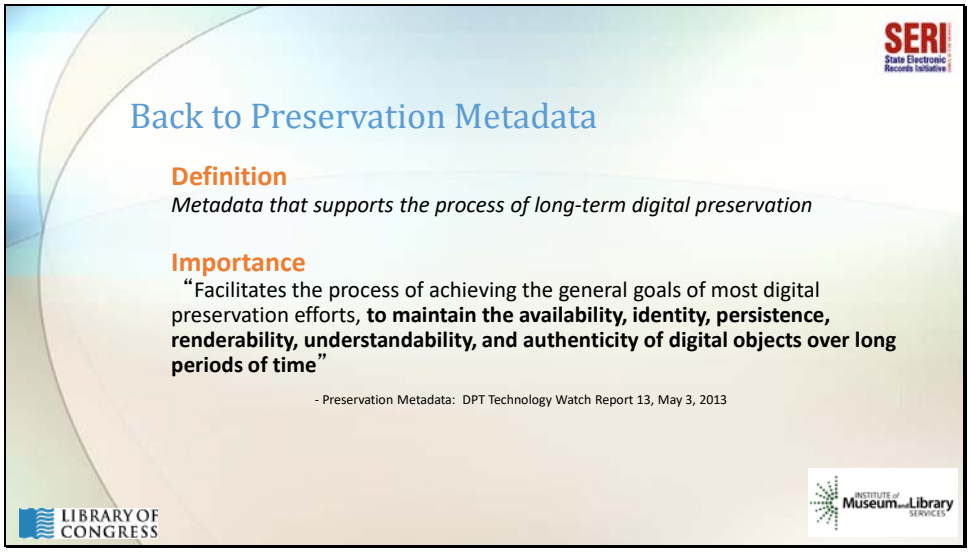
Structural– None – this is a single image file (structural metadata is for complex objects – like a book that has been digitized and has many pieces that make up the whole object)

There is technical metadata stored (embedded) within this file, but not everything I need to be able to preserve this over time is here.
No information about how this file was created (scanner model, for example)
No checksum (you have to generate that and store it somewhere)

Slide 20

These archival metadata goals came from the 1996 Preserving Digital Information Report by the predecessor of CLIR (Council for Library and Information Resources). These goals are specific to archives, and are not the only goals for metadata.

Slide 21



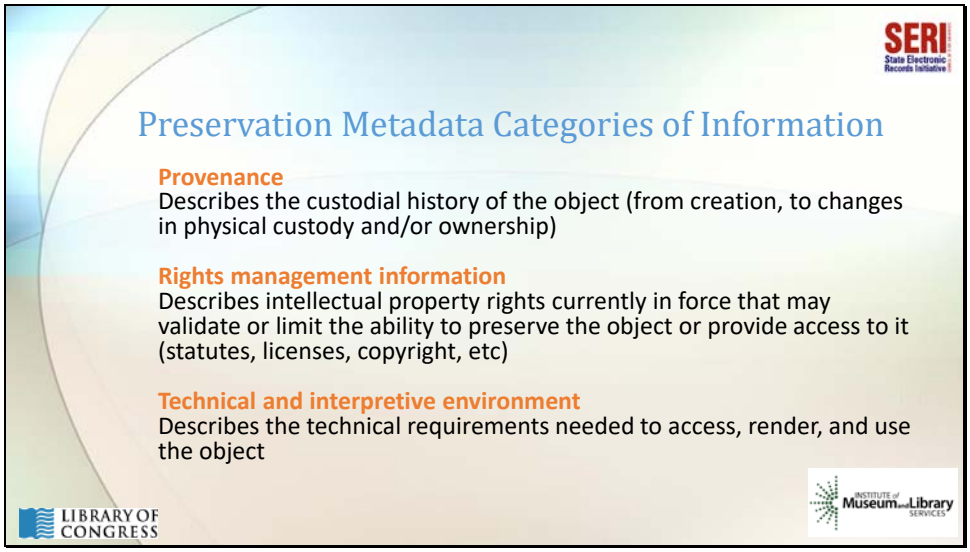
But … Preservation metadata does not necessarily fit neatly into the descriptive, structural, or administrative metadata types. Preservation can actually extend into all three.

The scope of preservation metadata is best understood not so much on the basis of the detailed function of the metadata – i.e., to describe, to structure, to administer-but instead on the process, or larger purpose that the metadata is intended to support.

Preservation Metadata Definition: Metadata that supports the process of long-term digital preservation

Importance: facilitate the process of achieving general goals of most digital preservation efforts (to maintain the availability of the records).

Slide 22



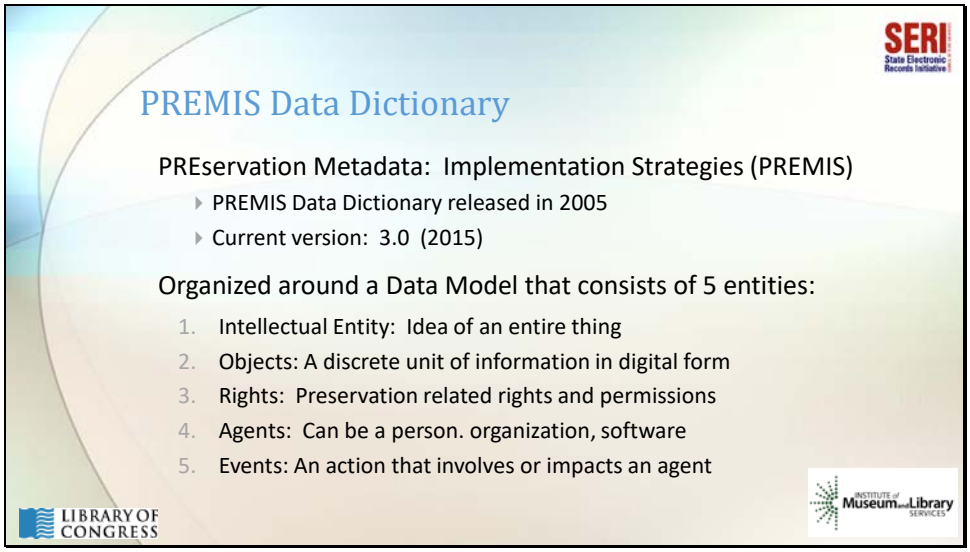
Preservation metadata builds an informational frame of reference around a preserved digital object that generally includes the following categories of information:

Provenance: Related to this would be information that serves to establish and validate the object's authenticity (that the object is what it says it is and has not been altered in any undocumented way)

Rights management information: such as copyright, and the technical and interpretive environment associated with the object (information that describes the technical requirements needed to access, render, and use the object).

When thinking about what type of preservation metadata to collect, it may help to ask yourself – Does this information directly support the long-term digital preservation process? Does this piece of information explain anything about the object's provenance or a preservation activity performed on the object, rights associated with the object, or the technical/interpretative environment needed to render and use the object?

Slide 23



In practice, the scope of preservation metadata has been agreed upon and centers around the de facto standard, the PREMIS Data Dictionary.

PREMIS was released in 2005 for the first time.
In 2015 it was updated to version 3.0.

Primary uses of PREMIS are for:
- Repository design
- Repository evaluation
- Exchange of archival information packages among preservation repositories

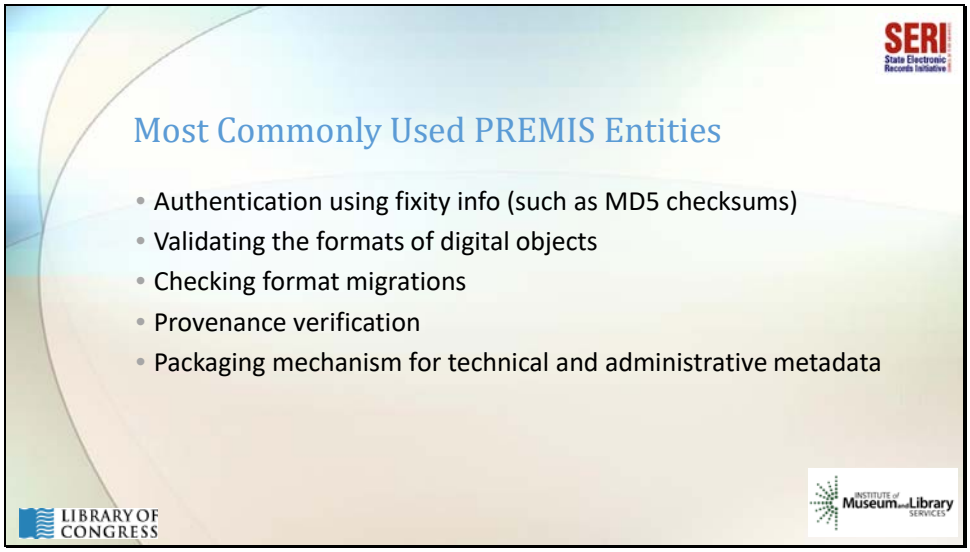
PREMIS data dictionary is organized around a Data Model that consists of 5 entities associated with the digital preservation process.

Intellectual Entities: Idea of an entire thing
Objects: a discrete unit of information in digital form
Rights: Preservation related rights and permissions
Agents: Can be a person, organization, or software
Events: an action that involves or impacts an agent

OAIS is a conceptual framework describing the environment, functional components, and information object associated with a system responsible for the long-term preservation of

digital materials. The OAIS information model served as the foundation for, or at least informed, the development of most preservation metadata initiatives.

Slide 24



Common usages for PREMIS include:

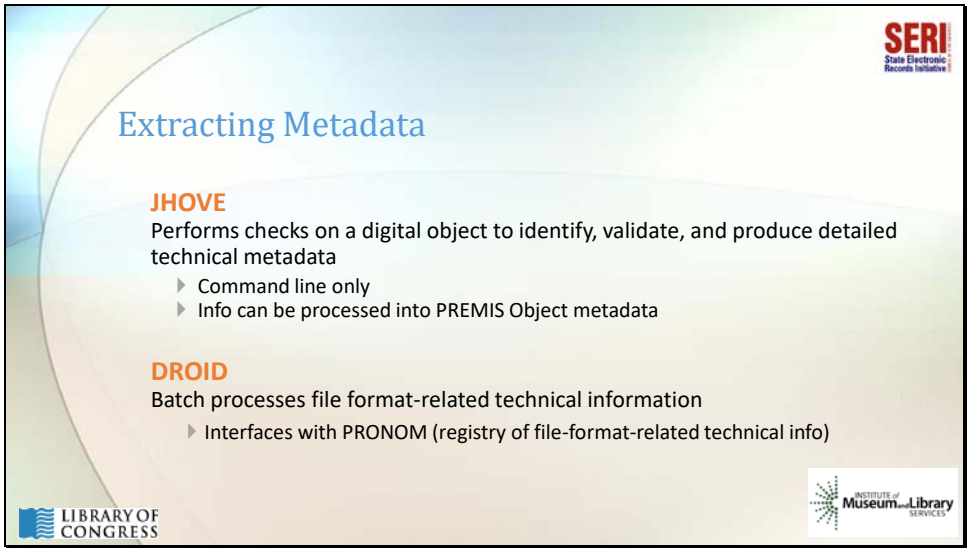
authentication using fixity information (such as MD5 checksums)

validating the formats of digital objects

checking format migrations (including recording conversions to new formats)

provenance verification (particularly using the Event entity to provide an 'audit trail' for an object)
packaging mechanism for technical and administrative metadata (as an alternative to, for instance, METS).

Slide 25



There are a couple of key tools that can be used to extract metadata from digital objects.

JHOVE (or JSTOR/Harvard Object Validation Environment) carries out a number of checks on a digital object to identify, validate, and produce detailed technical metadata from it. It produces an extensive list of information on the object itself, which can be processed into PREMIS Object metadata. For a TIFF file, for instance, approximately 40 information components are reported around the Basic digital object information, image information, capture information, and change history.

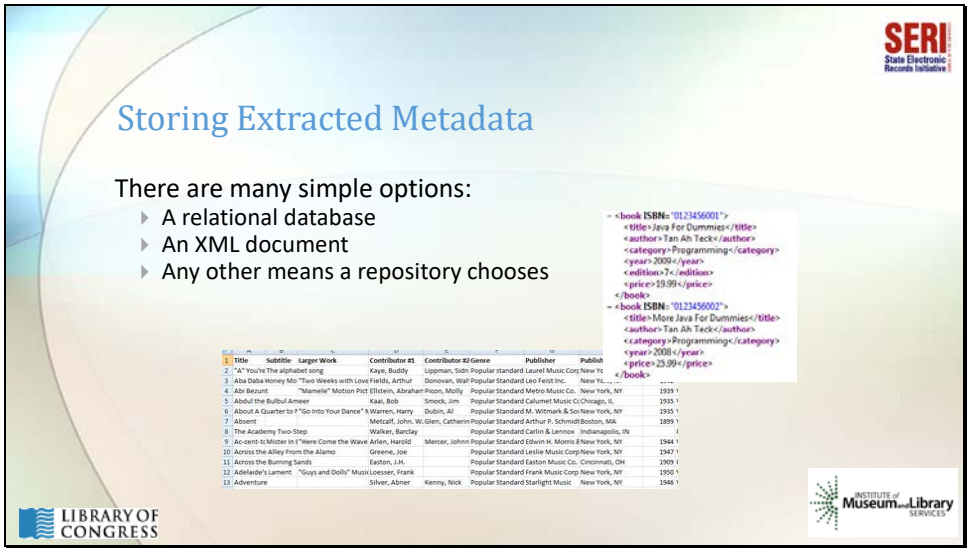
DROID stands for Digital Record Object Identification (Developed by the UK National Archives)
It is a multipurpose tool that can assist with many areas in digital preservation.
DROID is used for file format identification purposes. Useful in determining what file formats are in the collection and what files might be at risk when thinking about long-term digital preservation.

It will tell you the file format (even if the extension is missing), and it will tell you what version was used to create that file format. In addition, it enables you to:

- understand what different formats information is stored in, and how much space they take up
- understand what sort of information your organization is creating by examining the mime-types

So, there are tools that can help you gather preservation type metadata about the objects you are storing.

Slide 26

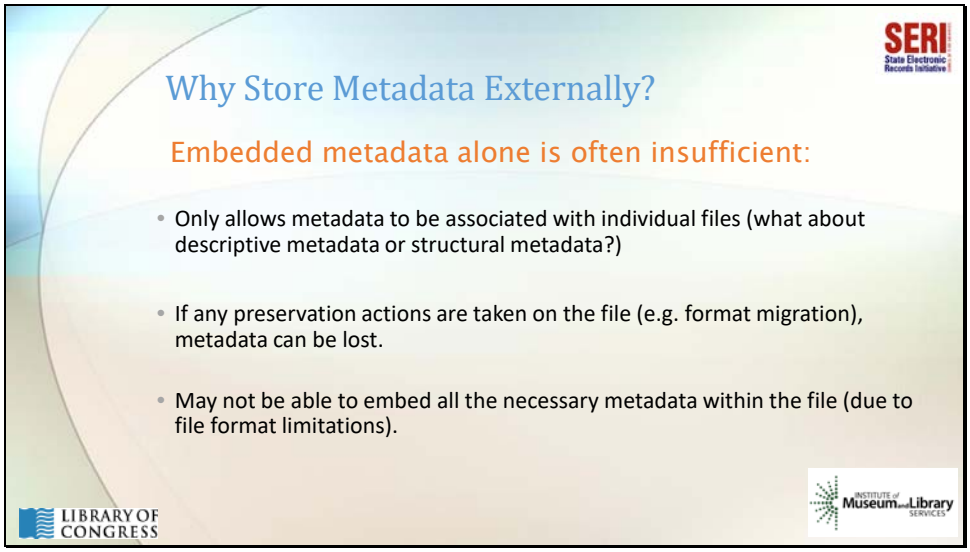


Storing PREMIS (or any type of preservation metadata) is designed to be implementation agnostic. The metadata could be stored in a relational database, an XML document, or by any other means a repository chooses.

Some archives store that metadata in a tab delimited text file along with the master copies of the image on a server. A digital preservation system or service will store this preservation metadata along with the digital files in its database.

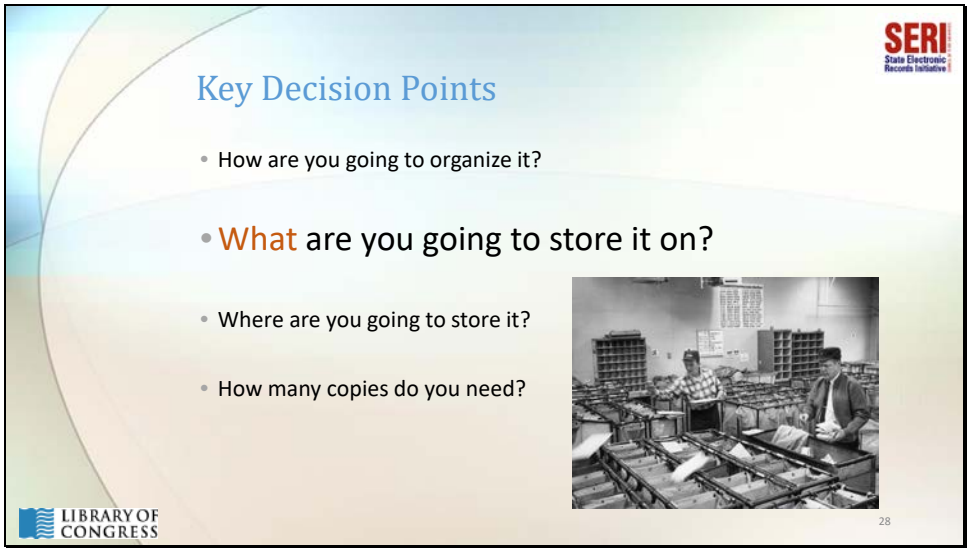
EXTRA
The survey by the Implementation Strategies Subgroup showed that repositories have implemented several different architectures for storing metadata. Most commonly, metadata is stored in relational database tables. It is also common to store metadata as XML documents in an XML database, or as XML documents stored with the content data files. Other methods include proprietary flat file formats and object-oriented databases. Most respondents were using two or more of these methods. (For more information, see the Implementation Survey Report2.)
Page 23: http://www.loc.gov/standards/premis/v2/premis-report-2-2.pdf

Slide 27



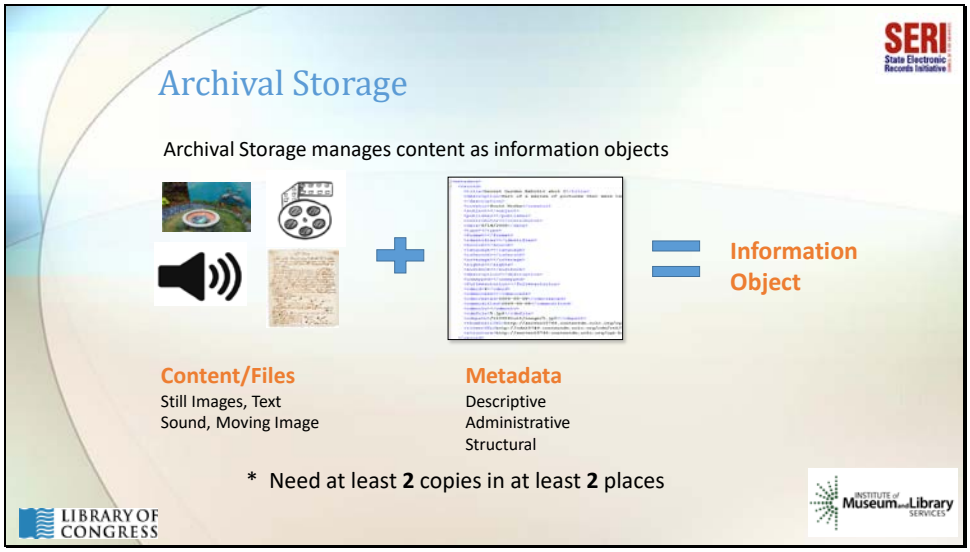
Technical metadata is often stored externally in a repository. Embedded metadata is usually not enough, hard to access, and easy to lose.

Slide 28



So now we've talked about organizing things with naming conventions and metadata…and moving on to the WHAT question

Slide 29



Archival storage manages content as information objects. It is the digital files themselves (Which can be any format - images, text, sound, video, maps) + the metadata (requiring some identification and description) which = the information object (NOTE: This is from the OAIS definition of information object).

Simply managing well-formed files in association with metadata to manage and use it, is a big step towards good practice. I'll talk a little bit more about what I mean by well-formed files a bit later on in the presentation.

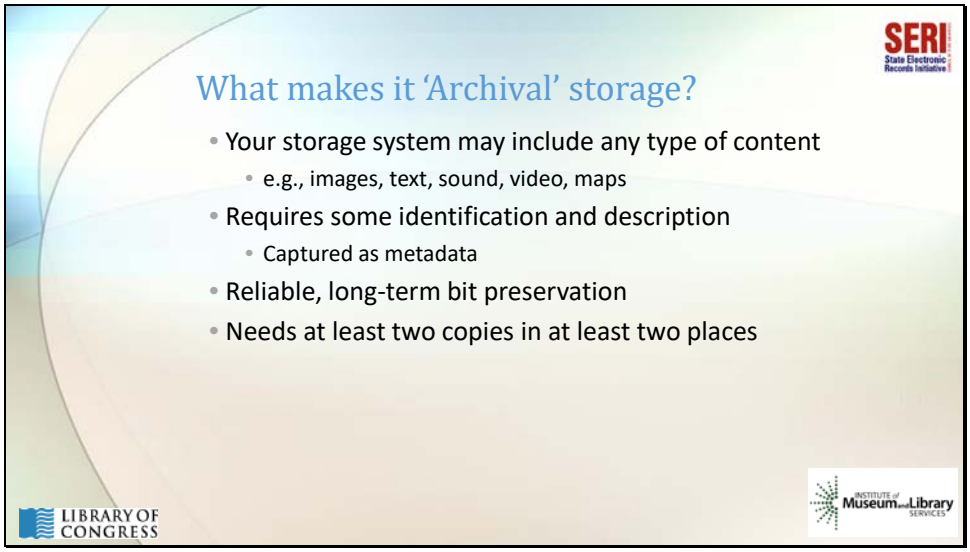
You should strive for at least 2 copies in at least 2 places. The 2 places should be geographically separated in some way.

Slide 30

Archival storage does not equal system backups. System backups are intended to bring up a system in case of a failure, including all of the files in it when it fails. Preservation wants to individually care for files containing content over time, across generations of technology, so system backups are not suited to that.

Common practice for preservation is to store content in a format that is as least software-dependent as possible – not compressed, not encrypted, preferably in ASCII or XML for text, for example.

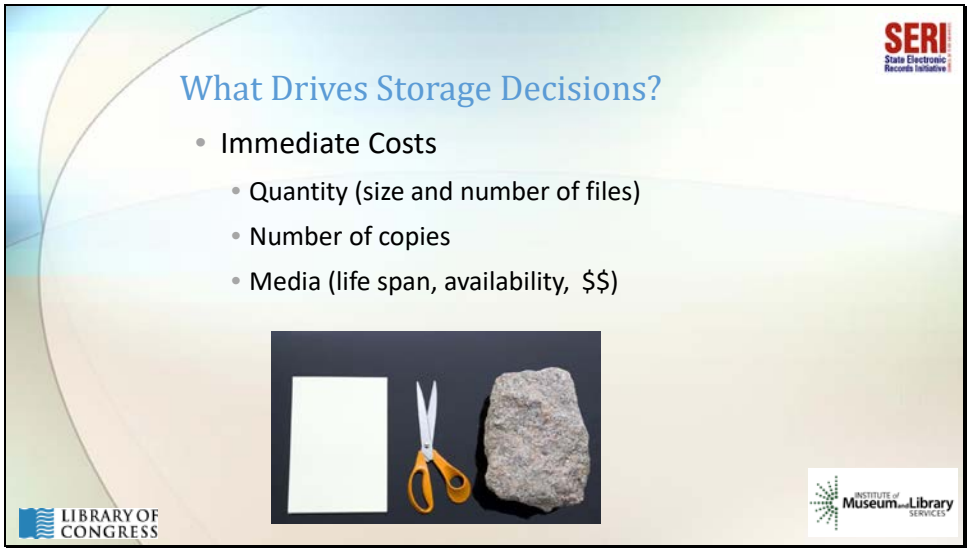
Slide 31



Some collections will contain only or mostly one type of content and will have specialized metadata and storage strategies for those objects.

However, you may use a general approach, like hard drives and Dublin Core, which are "agnostic" or "free" from format and can describe anything.

Slide 32



What drives storage decisions?

Quantity – The options you consider will vary depending on how much you have to store. How many files do you have and how big are they? Are they videos…images…documents? What if your video files are too large to store many copies?

Number of copies – how many copies of your content do you have? 2, 3, more? Again – the size and number of files will play heavily into this

Media – different types of storage devices have different lifespans. How often will you have to migrate to new hardware?
CDs – on average 5 years
Gold CDs - more
Magnetic Tape – could last 30 years, but it's very sensitive to heat, magnetic fields and dust.
If you are leaning toward a server - Is the company producing the hardware you are using to run the storage media still around?

You also need to determine where you **don't** want to store it and migrate it off those devices accordingly, such as USB drives, old media, etc.

Slide 33

Regarding Resources, how about

Expertise – Do you have the staff expertise available to run a server – whether it's through your departmental staff or IT support?

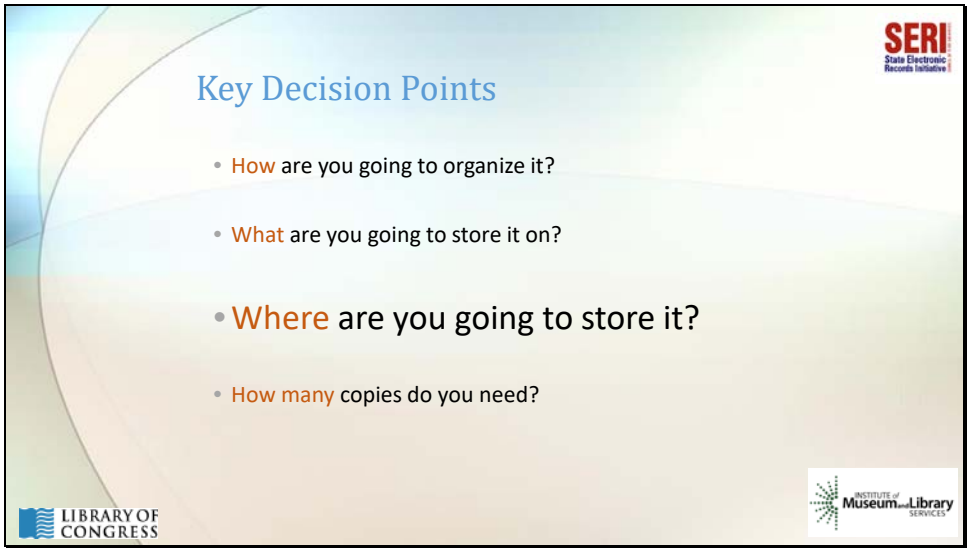
Services – are you going to manage your content locally or host it externally and with whom?

Partners – There are now hosted services and collaborative groups (such as MetaArchive) that organizations can join to meet their preservation obligations. Are you going to host your materials with a partner as opposed to a formal cloud service and if so what are the costs associated with that?

Institutional - Are there any legal restrictions to storing your materials outside your boundaries?

Basically - you need to know what you have, understand the resources and options available to you, and make the best decision you can – knowing that this isn't the "final" decision. Technology will continue to change and with it your possible options and requirements. You just need to get it somewhere for now.

Slide 34



Among key decision points is:

WHERE you are going to store it?

Slide 35

You can store it locally. You can store it with institutional partners. Or you can use large commercial options.

Slide 36

Online - Online storage is good for data that needs to be constantly accessed and updated, however it's less secure being more open to corruption, and cybersecurity threats.

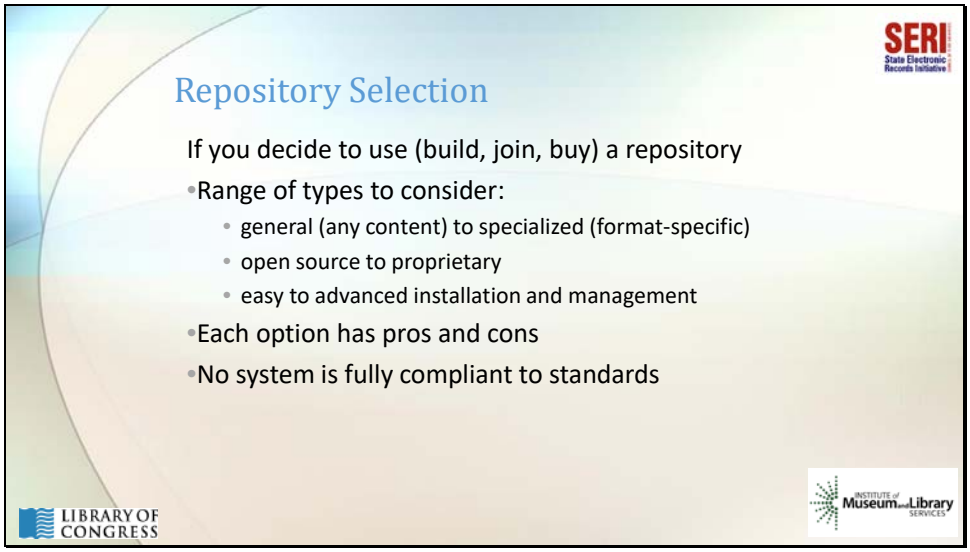
Offline or near-line storage utilizes tape or hard drive media and a robotic system that retrieves requested data. Data is not online, but can be retrieved online through a comparatively lengthy process.

The community has moved away from offline storage as a good option for preservation (though many organizations have a set of high resolution images stored on gold CDs as an artifact from when offline storage was good practice because that's what was possible technologically and affordable.

M- Disc – is a write once optical disc technology introduced in 2009 and available as both DVD and Blu-ray discs. These are much more durable than traditional DVDs – better able to handle environmental fluctuations like heat/cold, humidity, light etc.

There are different disc options with different storage capabilities, ranging from about 5GB to 100 GB. They claim to last 1000 years, but we won't be around long enough to prove/disprove it The Discs tend to be more expensive than archival gold CDs – BUT you also may not have to migrate all your data every few years like you would traditional discs. The greatest risk is technological obsolescence and not having DVD drives available.

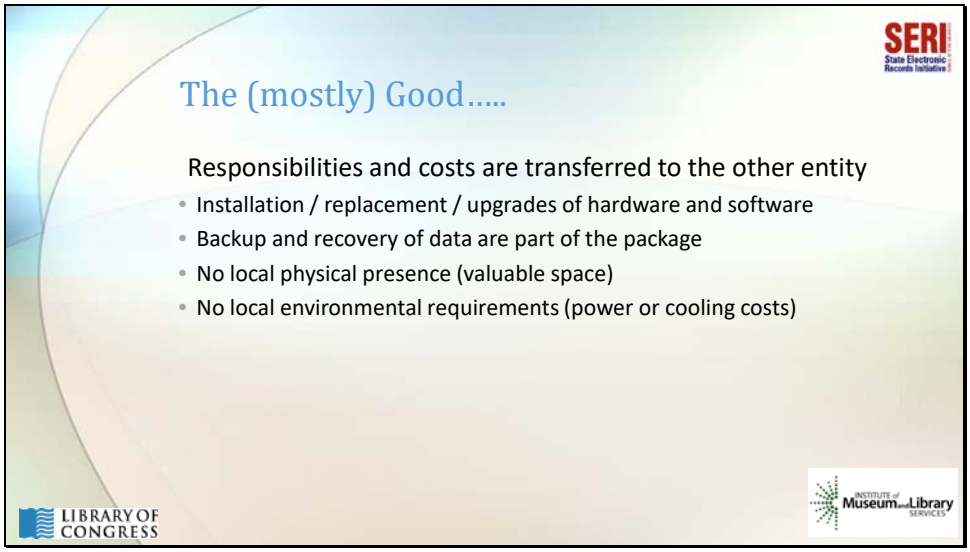
Slide 37



If an organization decides to implement repository software to manage its digital content, they should be intentional about it. Figure out your requirements and if the option you found will work for 60-80% of what you want to achieve. There are lists of requirements that can help with selection – TRAC may be overwhelming but works for this, and there are others.

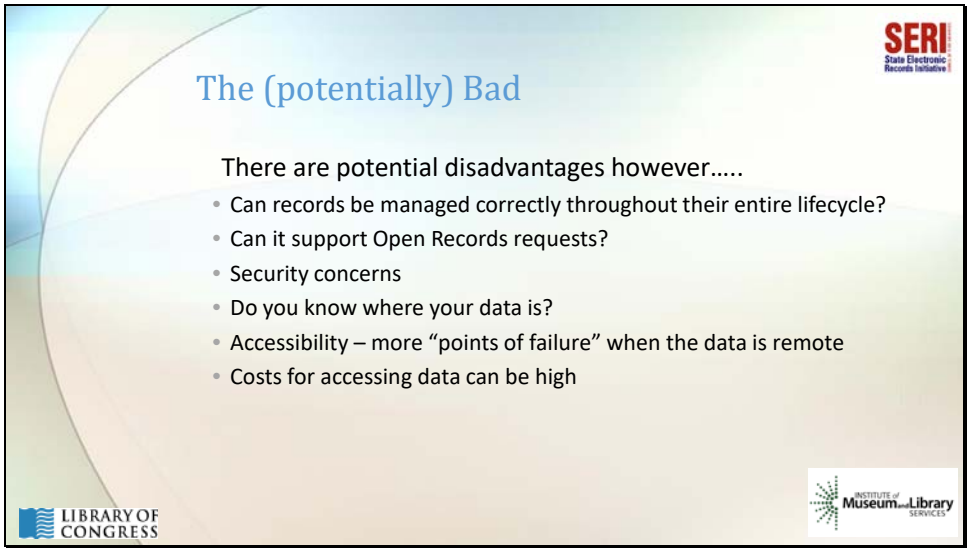
Compliant to standards means digital preservation standards, most often referring to OAIS. OAIS is an advanced and large topic, the specifics of which should be covered in another workshop. Most systems can take in content well enough (Ingest in OAIS) and get it back out (Access in OAIS). They may try to cover the full range of policies and procedures (Administration in OAIS) through an implementation of Data Management (the portion of OAIS that maintains comprehensive information about objects, collections, and the repository and provides reports about them). Another area that systems oversimplify is Preservation Planning – there is often not sufficient support for managing content over time (for example, managing content before and after a migration from one file type to another).

Slide 38



If you choose the Cloud or Partner Storage, the cost of buying/maintaining/upgrading hardware move to someone else. Technical staff are transferred to the service provider.

Slide 39



There are also negative consequences or potential negative consequences:

[lifecycle]
Does it have archive capabilities?
Can it maintain "restricted access" on appropriate records?
Does "delete" actually MEAN "delete"? Can the contractor delete or purge electronic records in accordance with approved retention schedules?

[security]
How many people have access to your records via the network?
How many people have access to the servers your records live on?
Does your contractor work with subcontractors (who you don't know?)

[where]
Does the data reside in this state, country?

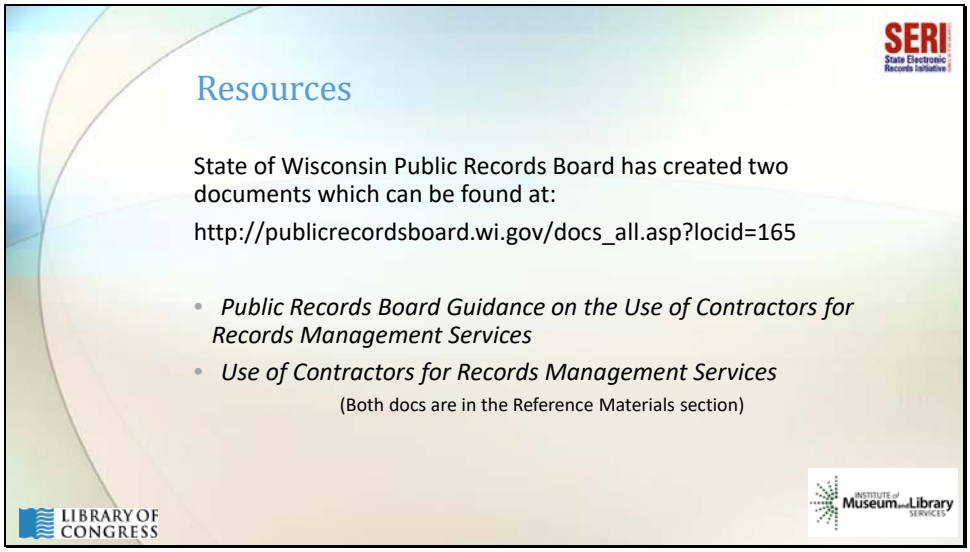
[accessibility]
Network availability across large distances can be a problem, such as service outages, power outages, severed cables, unspecified network outages.

If there IS an outage – What is the minimal acceptable time for getting content back up again?

[costs]

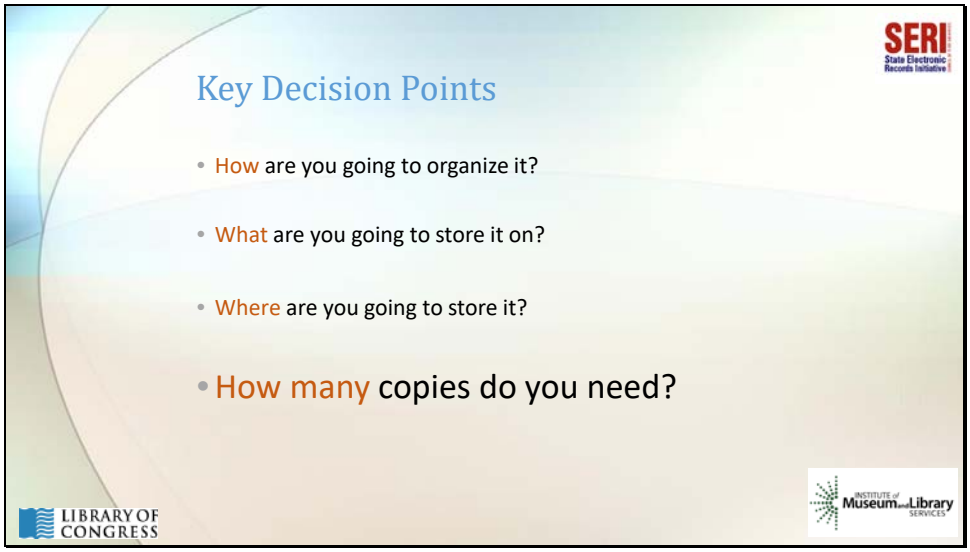
Sometimes the storage fees for holding the data are fairly low – the charges are different (and usually higher) for each time you access your data.

Slide 40



Here are some resources that may be of interest.

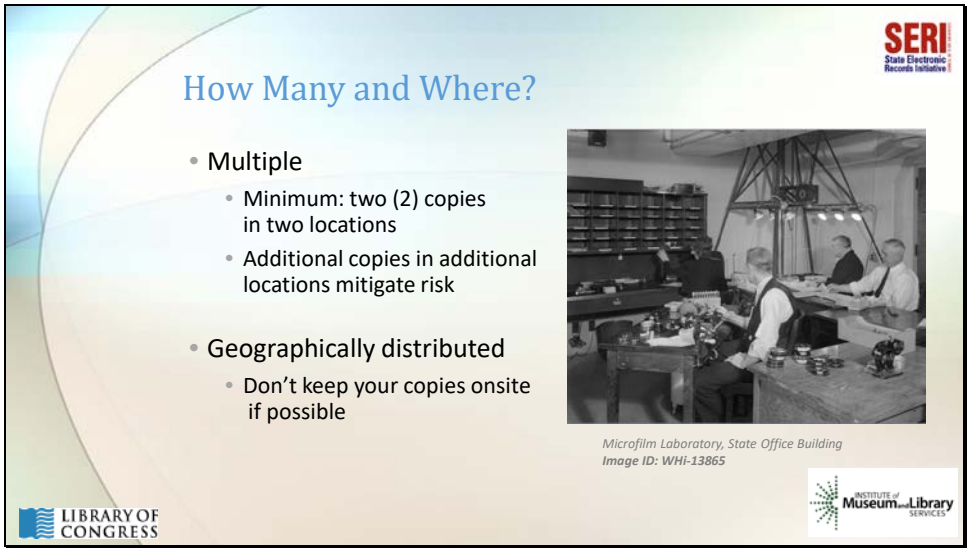
Slide 41



Finally,

HOW MANY COPIES are you going to make?

Slide 42

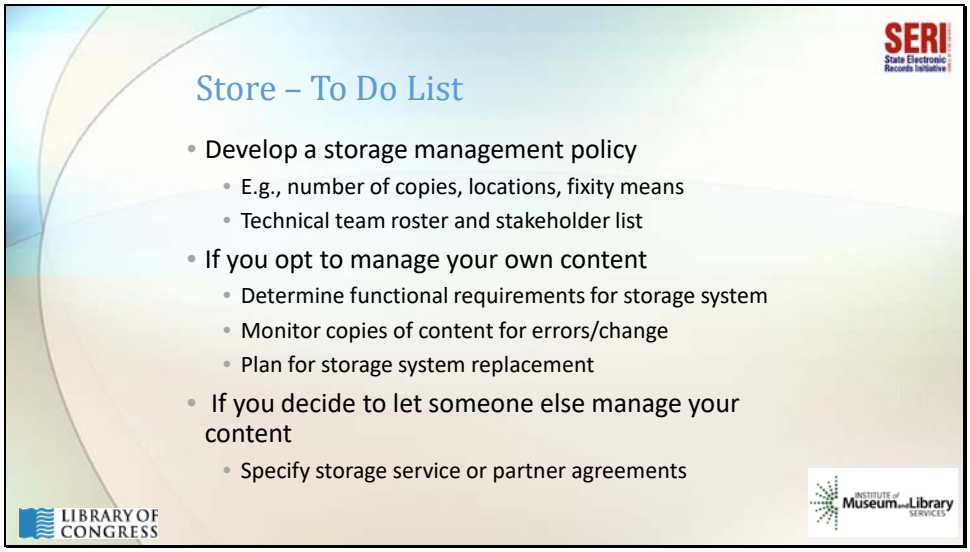


Three copies is a happy medium if you are able.

You need to decide how many copies you think are sufficient to preserve the content in your care.

It may depend on what you are preserving – If it is born digital and you have the only copy, you may want to do several copies on different media just to make sure. If you digitized something and you still have the original…maybe you just have a couple of copies in a couple of locations.

Slide 43



So here's what you need to do:

Develop a storage management policy.
Know what you are getting into if you manage your own content.
Know what you are getting into if you let someone else manage your content.

Slide 44

This completes module 3, Store. If you are using these modules in order, the next one is module 4, Protect. For additional resources on electronic records preservation and management, please visit the State Electronic Records Initiative webpage. This link is on your screen.