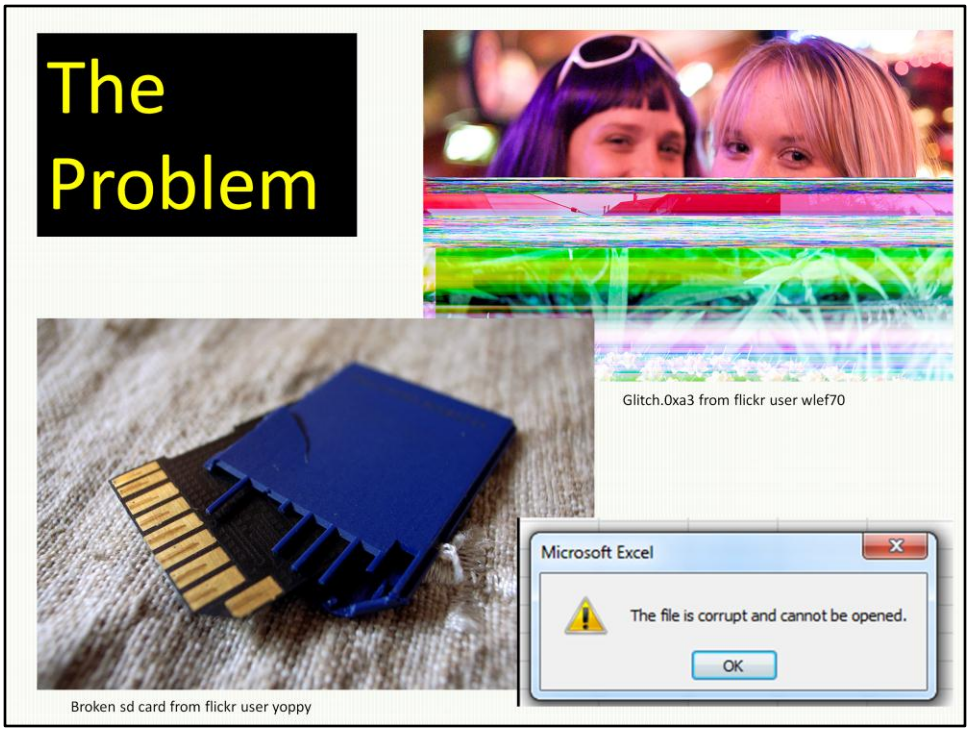
10-2014

Getting Started with Digital Preservation: Initial Steps

Max Eckard
Grand Valley State University, eckardm@gvsu.edu

Kevin Driedger
Library of Michigan

Poll: Show of hands... How many people have every personally lost something digital. Whether digital photos or music, your hard drive crashed, an on-line service went out of business, or you had unreadable flash drives or unreadable files? Has anyone ever washed their iPod?



There are a lot of things that can go wrong with our digital files--they are actually pretty fragile compared to their analog counterparts.

Complex: A lot of things had to come together, like hardware, software, operating systems. These things change all the time, and these environments can be hard to recreate.

[Where as you can still understand the contents of a torn photo, with a corrupted digital photo you get something like this, or, in the worst cases, something like this]

Technological changes: Happen so fast! Unlike cave paintings, written thousands of years ago which we can still read, if we have something from just 10 years ago that's saved on a floppy disk or was written in an older program, there's a good chance you can't open it today.

Physical threats: There always the threat of disasters, natural. Thumbdrives or external hard drives are susceptible to damage, theft or loss, putting it too close to a big magnet.

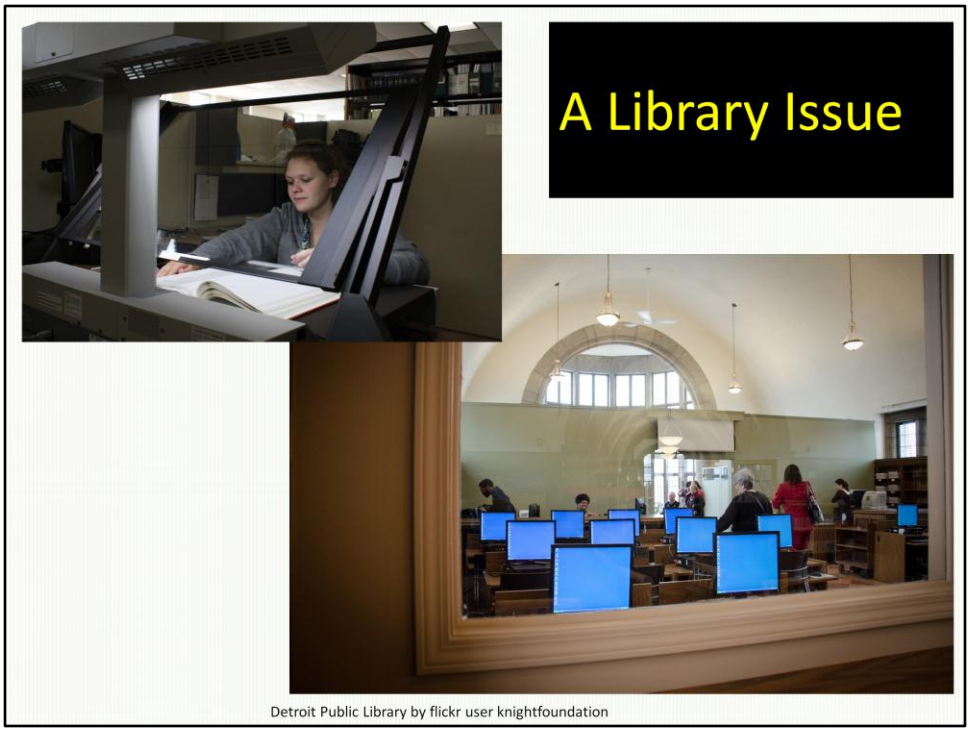
Human threats:

Malicious ones. Hackers, viruses. ILOVEYOU, Shell Shock, Stuxnet

Also not-so-malicious ones. Accidentally deleting something, or forgetting your password, or washing your iPod.

Sheer volume

Loss is significant



Kevin and I also wanted to make the quick point that problems of digital preservation, access, are really library issues.

Most basic argument is that this is what we do. Collect, describe, provide access and preserve for future generations.

Digital collections are much more than pretty images of old books, but are a vital research tool, and they will get used in ways we can't even imagine. They support our library's mission (for us, teaching and learning at GVSU) just like print materials do.

Just as our physical collections need special attention to ensure their ongoing persistence--proper environment, proper handling, proper cataloging, proper repair--likewise the preservation of our digital collections require proper environment, proper handling, proper cataloging, proper repair.

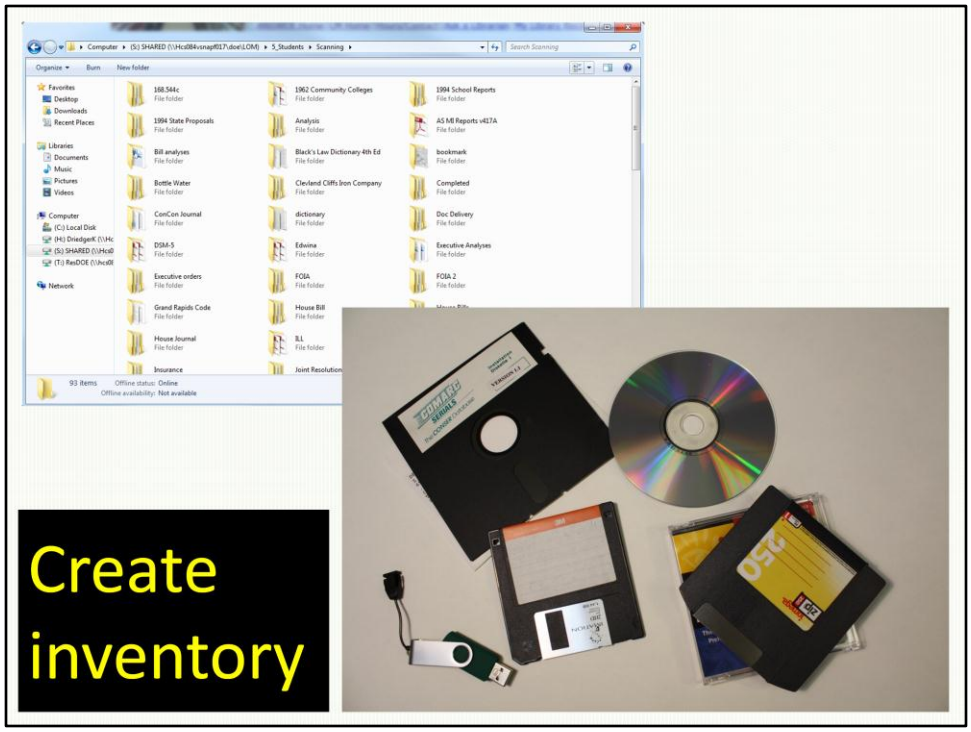
Don’t know where to start
Confusing models
New vocabulary, full of scary words (file format migrations, corruption, emulation)
Lots and lots to know
Constant change
Usually a lot to deal with (volume-wise)
Lack of expertise on staff

Which is why we created this presentation.

What you have and what you will have

Good preservation decisions are based on an understanding of the possible content to be preserved.
Identifying that content is a first step to planning for current and future preservation needs

Begin by creating an inventory.



When print books become part of a library's collections where I work, they usually go through standardized acquisitions and cataloging steps that we've been doing for decades. The result is a centralized catalog record with abundant information about that book. That is one inventory.
But Digital materials, at least in my experience, don't always find their way into our collections in the same standardized and structured way, so our record of what we have is often not as robust.

What an inventory will do is:

It will be the foundation for all future preservation activities.

It will be a growing tool. Next you will hear from Max about how you will build on it.

Two important things to remember in creating an inventory:

1) Content is more important than style.

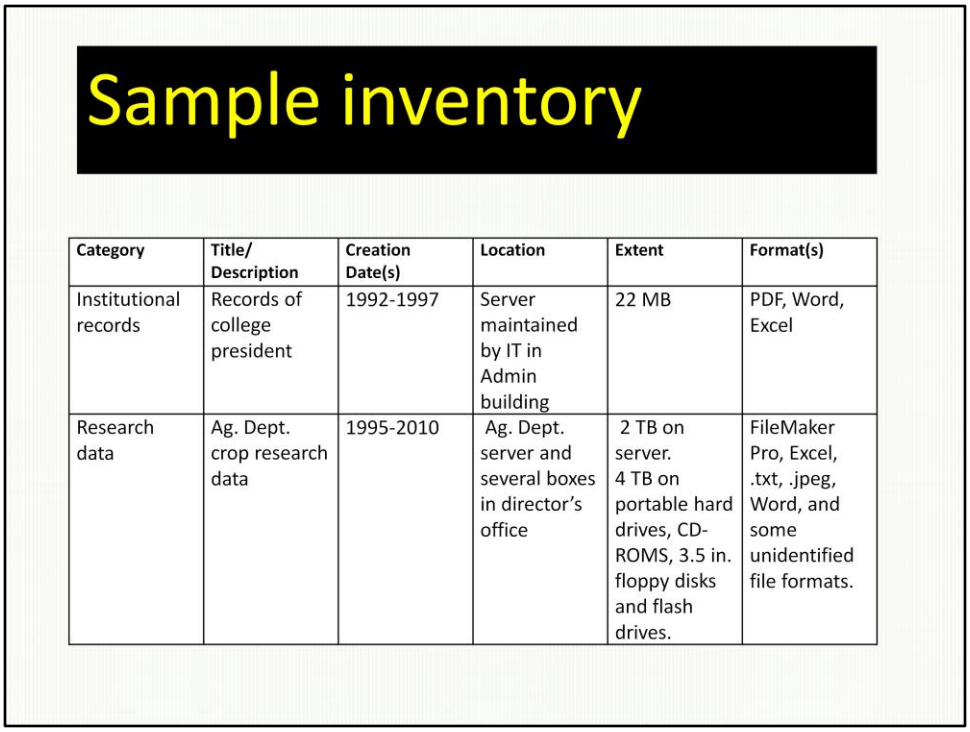
It doesn't need to be anything more elaborate than a spreadsheet. That an inventory exists is much more important than it being a really thorough inventory.

2) Be consistent, and concise. Level of detail in inventory depends on extent and nature of all collections involved. The more consistent you are now the less you will need to redo things in the future.

Some potential elements include:
Title
Content category (e.g. research data, special collections, institutional records, etc.)
Format type (e.g. images, video, audio, text, etc.)
File/Media types (e.g. .doc, .jpeg, .html and cds, magnetic tapes, hard drives)
Extent (number of files, size of collection)—bonus points if you can convert this to an estimated digital file size (or does that come later?)
Physical location
Dates (coverage, creation, inventoried)
Maybe note about condition (if it looks damaged…)

Do you have formats you can no longer access?
Yes
No

This inventory will help you plan for the future, because not only will it give you a sense of the volume and size of the collection, but also what kind of hardware and software you will need in the future.



Example of LM inventorying (or not)

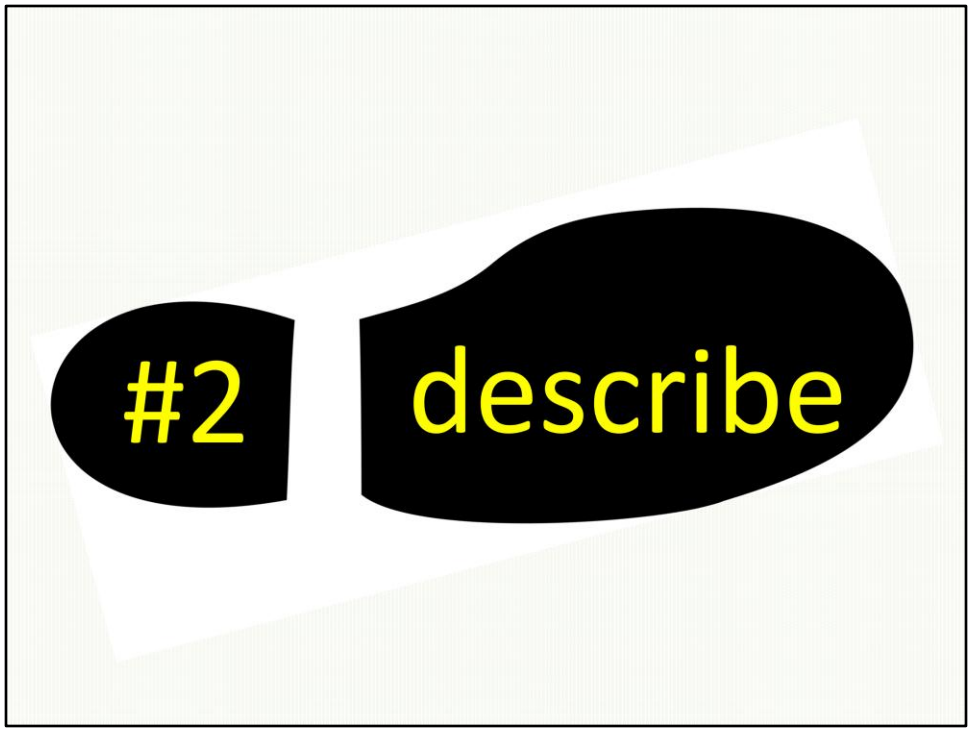
Briefly about Selection

Our libraries will not necessarily need to preserve all digital content in our collections. Will also need to prioritize collections for preservation activities. Influencing factors include:

Content – does the collection have value / fit library's scope (Probably the hardest question)

Technical – is it feasible for us to preserve the content? (Emphasize being good stewards)

Access – do we have the ability – both technical and legal – to provide access to content?

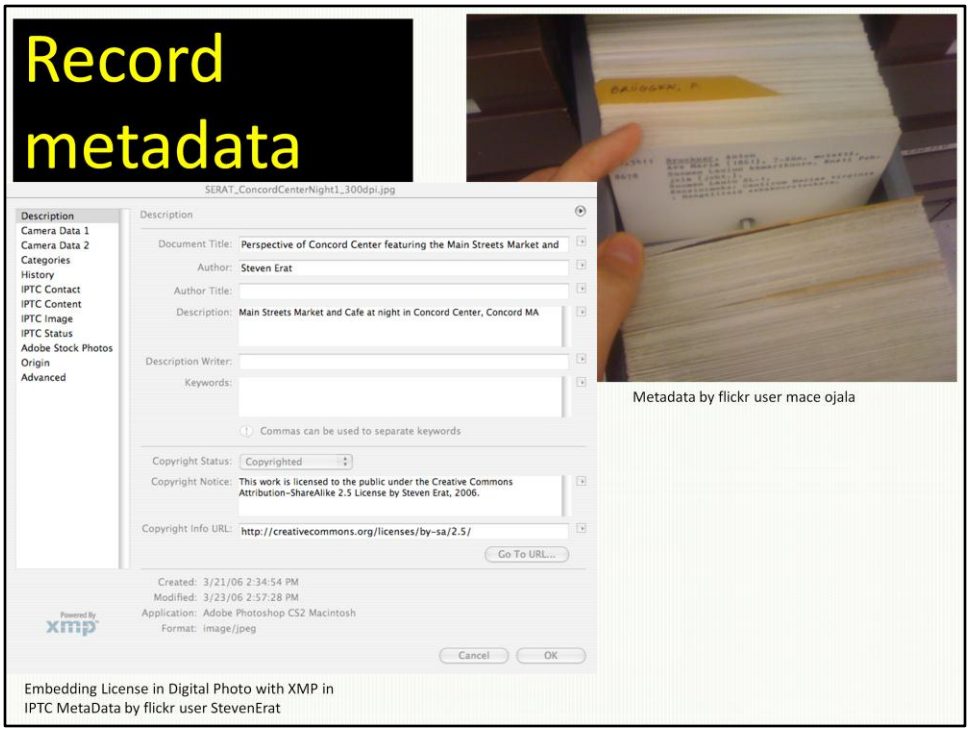


Transition:

We're doing this now with some "potentially" digital stuff. We heard a recent AVPreserve report making argument about paper vs. media. How we only have a 10-15 year window and these things need priority since paper is relatively stable. Experientially we've found this to be true. Many of the old reel-to-reel audio is developing sticky shed, and we can't get a good read anymore. So we've been making a list of what we have, what we think is on it, how much we think is on it (time-wise) so that we can convert that to an estimated size. Then we're making priorities.

Describe:

Description = Metadata



Metadata is just a fancy word for something that libraries have been doing for a long time: cataloging.

Metadata is data or information about whatever it is your describing:

- Book, like libraries have done for a long time.
- Digital Object
- Song (Pandora Radio)
- Telephone metadata

Inventory is inward-focused, administrative in nature and typically recorded at a high-level.

Metadata builds on inventory to describe individual digital objects so that they make sense to **current** and **future users**. Does this in a couple of ways.

First there's descriptive metadata. You're probably familiar with this, and it has the most obvious correlation to the types of description that libraries have always done.

Compare to book
Title
Author
Description

Publication Date
Information on spine and covers

This information tells you what it is your about to read, and provides *context* for the book. It even tells you how to read it (pages numbers tell you what order to read them in).

Metadata provides that same context about and structure for digital objects (which aren't packaged as neatly as books), telling user what a digital object is and how to use it.

It's pretty obvious why this can help current and future users.

There's also technical metadata (i.e., file format) that helps computers help users and future users (by telling the computer what program to use to open or display a digital object), and administrative metadata (location of preservation version of a file) that helps the librarian or the institution help users and future users.

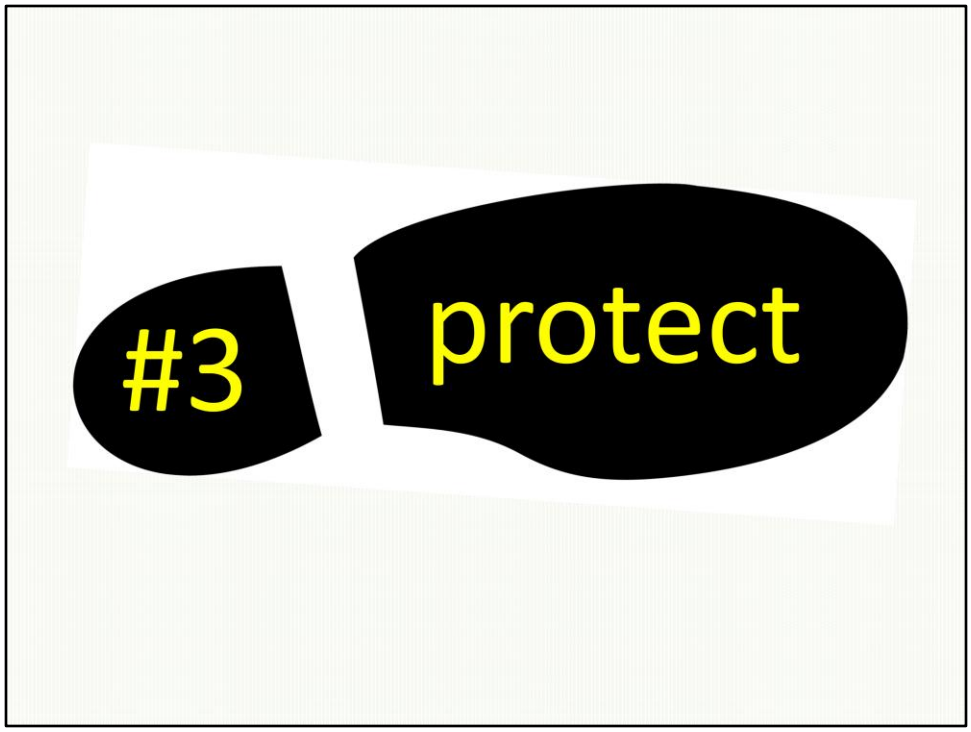
To be sure, these latter two types of metadata become more important as you get more serious about digital preservation…

But, what's really important is that you're creating metadata, even if it is descriptive metadata.

You've already got a good start in your inventory that Kevin described.

How you do it (formal or not, collection- or item-level) this is up to you, your users, and your organization.

Most importantly, you have to think about your users. You have to anticipate what kind of information will they want to know, and how they will want to use that information to search and browse through your collection.



LM pretty much everything gets full MARC record in OCLC, which becomes a Dublin Core record in CONTENTdm

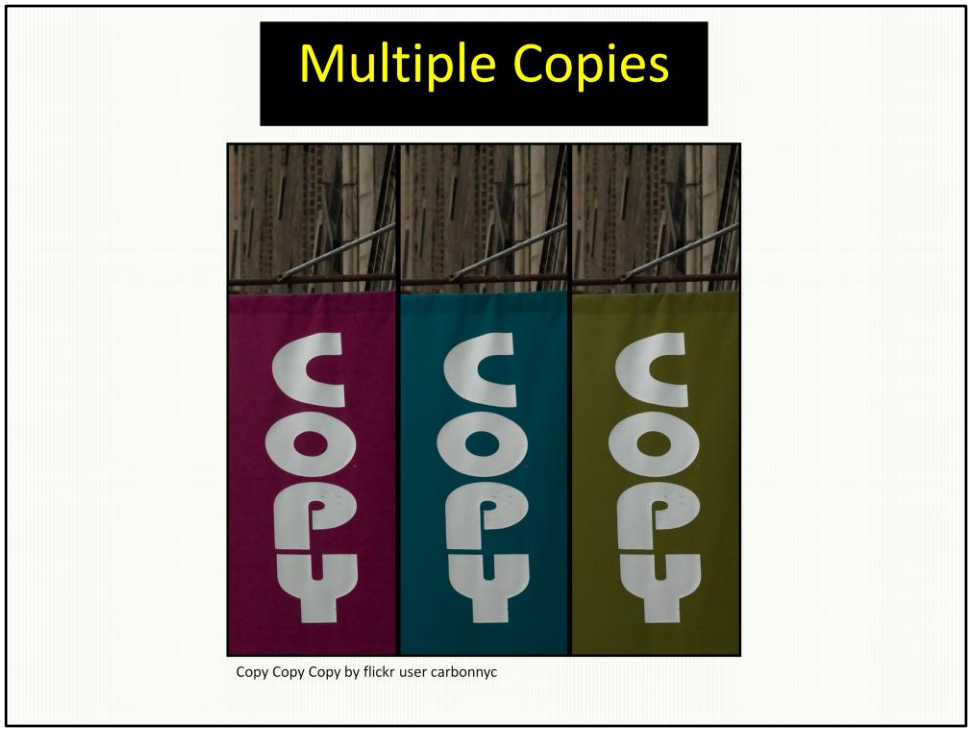
Thanks to our steps of identifying and describing we now have a good idea of what we have. Now we've come to the crux of the issue this afternoon. We want to make sure what we had yesterday we will still have tomorrow. And for that we move to the step of protect.

Why do we need to protect?

Because digital objects are really not that different than physical objects in that they are all bound to fall apart. Or, as this slide points out, Nothing is permanent. We often live in the illusion that the IT department has this all under control but hard drives crash, files becomes corrupt, file formats and media types become obsolete.

We need to protect these files to make sure they remain there, and remain unchanged. This involves basic risk assessment and management. What are the risks, what are their potential impact, what are the priorities, determine how to mitigate these risks.

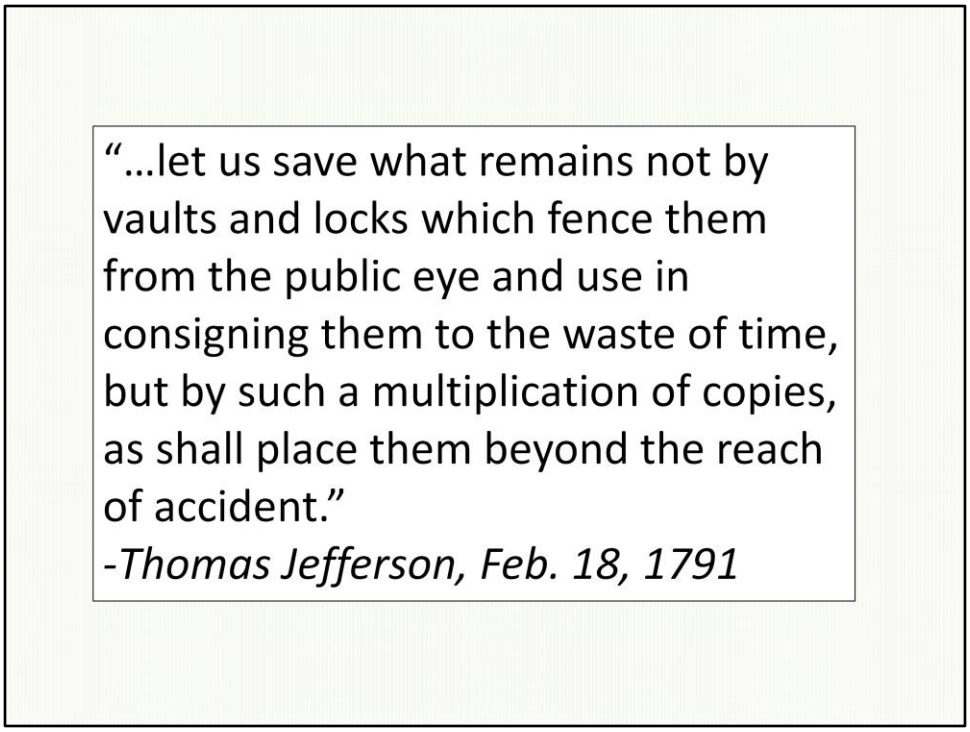
We're going to look at 3 actions you can take to help protect your digital collections.



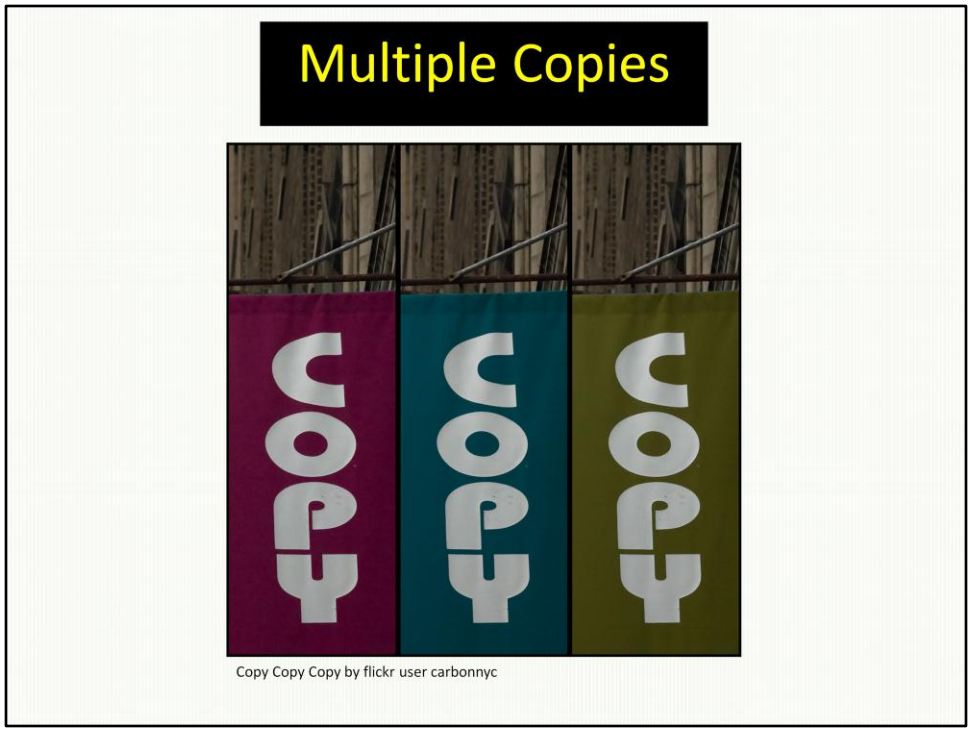
The most vulnerable object in the world of digital preservation is the unique object. A file that exists only on a single floppy disk, or a single iPod.

The ability to make multiple identical copies is one of the greatest strengths of digital material, and what most distinguishes them from physical materials (give example of photocopier vs. making copy of digital file).

So people leverage that ability. LOCKSS (Lots of Copies Keeps Stuff Safe), it's a storage program but also a strategy.



This worked for Thomas Jefferson: "…let us save what remains not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident." That's from February 18, 1791. It worked for Thomas Jefferson, it works for us.



How many copies is enough? At least 2, but more is better.

Greater preservation benefit by keeping multiple copies in multiple physical locations. Obviously the wide the geographic redundancy the better.

Finally, sometimes your IT can help with this (sometimes they can't), and actually using networked storage space can make your life a whole lot easier. A lot of time, these environments will automatically take care of these things I mentioned above.

Transition

GVSU IT network space automatically makes a copy of things in a spinning disk data center in Allendale and one in GR, and then copies to tape once a week to an off-site location.

LM copies are in OCLC's digital archive and scanned content is also kept on SOM network. Do not have multiple copies of published removable media – but exploring some options.

Security (KEVIN)

Preservation is not accomplished simply by copying files onto multiple network drives (although that is a huge and important step)

We also need to know where these materials are physically located and provide some physical protection – like don't keep server under a leaky pipe or next to an exposed window. Protect from theft.

Need to manage who has network access – may be levels of access.

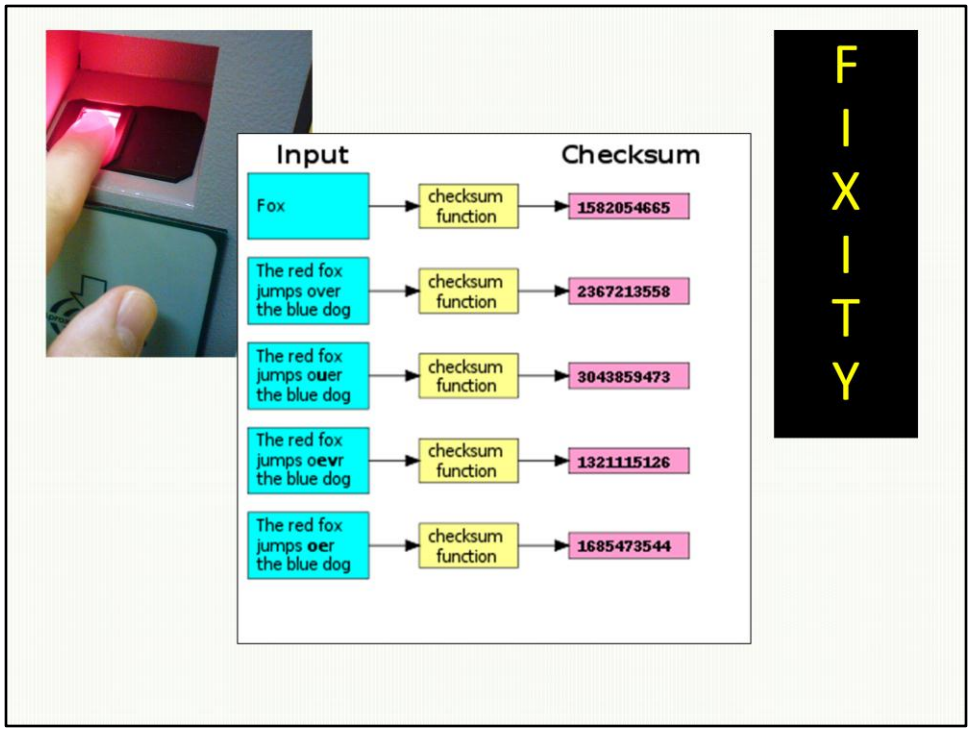
[SLIDE]

This is where passwords play a role.

Other types of protection include virus checking incoming material.

Also, if your library has a disaster plan – and you should – does it include your digital collections?

LM example – removable digital media part of public collections – security strips. Things that go into our digital archive – only a few of us have access and none of us currently knows how to edit/delete material from the archive.



Now we're getting a little more advanced.

Preservation is not accomplished simply by copying files onto multiple network drives (although that is a huge and important step), and creating policies for who has access to those drives.

Also need to protect these files to make sure they remain there, and remain unchanged.

It is the responsibility of an archive to take something from a creator, and to safeguard it. Then, in 10 years when someone requests that item, they need to be able to present it in a way that ensures authenticity, that it has not changed over time, whether that's by accident, like an individual bit or byte got corrupted or you spilled coffee on your external hard drive, or someone has gone in and changed the beneficiary of a digital will on an PDF.

So the computer world has developed something to check a digital objects fixity called a checksum.

Basically, a checksum uses a computer algorithm to create essentially a thumb print for a digital object. The algorithm is run on the underlying 1s and 0s of a digital file. If those 1s and 0s change, so does the resulting algorithm.

In practice, and this is what we do at GV, when we get a new digital object we calculate a checksum (there's computer programs available for free that do this) to take the files thumb print. We store the resulting checksum in administrative metadata. We don't display this to the user, thinking it's potentially confusing, but some folks choose to do that.

Then, at regular intervals or, in just whenever we want, we go back and audit that checksum on all the copies of that object to make sure they haven't changed. If they have changed, that's why we have multiple copies, and we can replace the changed copy with the authentic copy. It sounds hard but you can get a computer to do it pretty easily.

So that's fixity.

Digital Preservation is not one person on one path…

…but a dance with many partners

The steps of this dance – identify, describe and protect – are not difficult to start with. As you practice and get more comfortable with them you can begin to add some flourishes and some new steps.

(Mention resources and local practitioner groups as ways to keep up.)

Questions?

Max Eckard
Grand Valley State University
eckardm@gvsu.edu

Kevin Driedger
Library of Michigan
driedgerk@michigan.gov
@kevindriedger

And thanks