2014

'Good Enough' Really is Good Enough

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

'Good Enough' Really is Good Enough

Best Practices Panel

Max Eckard
Metadata & Digital Curation Librarian
Grand Valley State University

Outline

- About us
- What does digital preservation look like here?
- Compromising on best practices
- Informal audits

Start by talking a little bit about us to give some context.

Then talk about what digital preservation looks like here

How we've compromised on best practices.

And conclude by talking about a practice that we've found helpful, doing informal audits of our digital preservation efforts with other institutions.

ABOUT US

Grand Valley State University is a 4-year, public institution with about 24,408 students. We are a Master's Large institution according to Carnegie Classifications.

So, we're a comprehensive university… but we're not an R1, and we don't have the resources of an R1.

This is an interesting statistic from our President's Office that tells you a little bit about the resources that we have.

Our state appropriation per fiscal-year-equated student has actually gone down since 1988.

Our limitations

- Limited funds
  - No line for digital preservation
  - Commercial solutions are prohibitively expensive
- Most of the software we want is open source
  - GVSU and library IT departments have limited staff
  - Don't always have expertise or money for servers
  - Little to no support

Which leads me to some of our limitations here at the library. First an acknowledgement that, yes, we're standing in the Taj Mahal of libraries. But appearances can be misleading. Just like everyone else, we have **limited funds**.

Here, there's **no line for digital preservation or curation efforts** for things like student time, software or hardware. Funds are currently maxed out, which means that in order for us to get money for digital preservation, some program or service that already exists has to get less money.

So that's hard, especially since we're talking about a "they'll thank us in 10 years" kind of thing going up against other services that have a clearer, more *immediate* benefit to our patrons/users.

Having limited funds also means that **commercial solutions are almost always prohibitively expensive**. Which leads me to my second point: **Most of the software we want is open source** (Archivematica, BitCurator, etc.). In one way, this is great, since it's free. But in another way, it's hard.
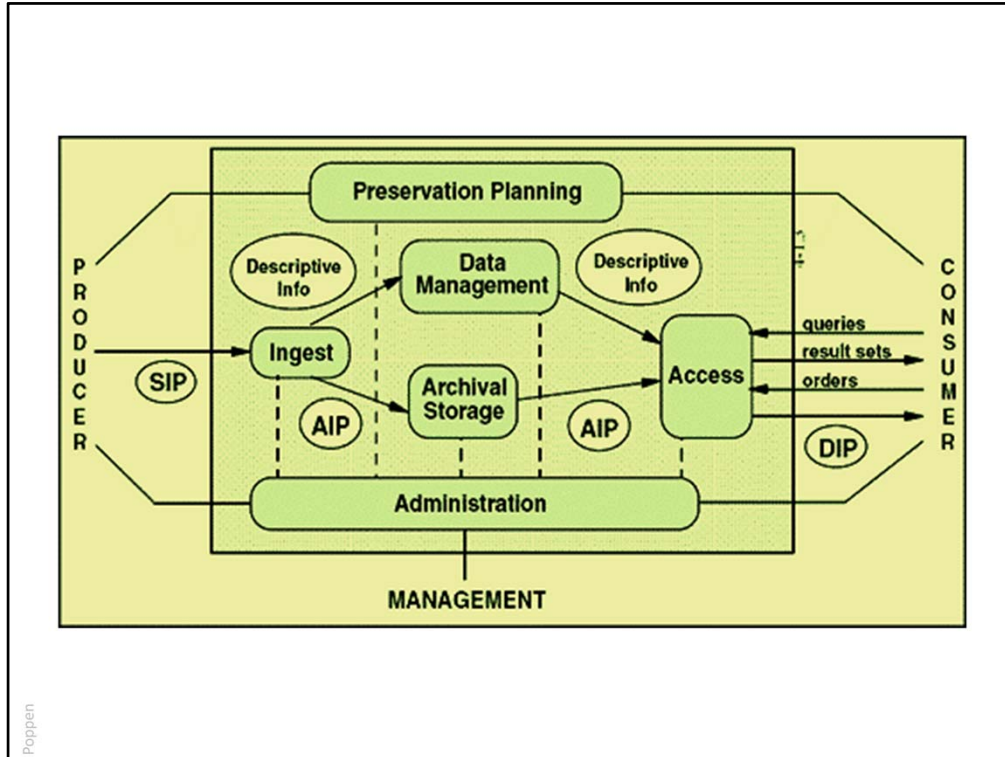
Campus IT and library IT both have limited staff. This means that we **don't always have the expertise** or **money** to support the servers that we would like for them to run.

As a result, we end up with **little to no support** for this type of software.

WHAT DOES DIGITAL PRESERVATION LOOK LIKE HERE?

So what does digital preservation look like here?

Well, it looks like this, which is exactly how it looks at a lot of other places.

I want to make an important point that when it comes to digital preservation, there are some basic principles, frameworks, standards and models, such as OAIS, that everybody follows.

In fact many of the principles, frameworks, standards and models that inform digital preservation efforts are intentionally designed to be content and system agnostic because it's their implementation that changes based on the resources that you have.

In other words, *what* you do doesn't change, but *how* you do it does.

COMPROMISING ON BEST PRACTICES

OK, how we've compromised on best practices. Or, ways that we've at least tried to compromise *intelligently* on best practices.

Archival Storage

Metadata

File formats

Our role

Really I think we've done this in four ways, probably more.

Archival Storage, with a capital "A" and "S" (I'm using OAIS-speak)
Metadata
File formats
Our role

The first three are more technical in nature, and the last one is more higher-level.

Archival storage

Metadata

File formats

Our role

So, Archival Storage.

We keep some of our **data on heterogeneous storage media**.

Specifically, I'm talking about the fact that when it comes to digital video, we keep a one copy of SIPs and AIPS on an **external hard drive** here in the library, and one copy on an external hard drive in the Special Collections & University Archives (and one copy on the cloud, but I'll get into that).

I know this isn't ideal for a number of reasons, not least of which because external hard drives are susceptible to damage, theft and loss, but also because their disconnectedness makes it hard to do the things that characterize Archival Storage, things like syncing and fixity checking, which for us has become a very manual process.

So, here's the story. Before we started collecting video, we used networked, backed up, geographically redundant Archival Storage, provided by our campus IT for *all* of our Archival Storage needs. And in fact we still use this space for *some* of our storage needs.

As it turns out, though, uncompressed digital video takes up a lot of space.

Our campus IT is very generous, but since we started collecting video, it's been hard for IT to keep up with our storage needs. Video has had to become an exception to our rules when it comes to Archival Storage.

This continues to be an issue because we are doing more and more video projects, and

because in any case we project that our storage needs in general are only going to continue to increase.

Archival Storage

- Cloud Computing
  - Storage (not the service) component of Archival Storage
  - Geographic distribution
  - Cloud Computing Guidelines

- Still looking to improve

Some more about **Archival Storage.**

Cloud Computing. I don't personally think that utilizing the cloud is a compromise on best practice, but I know there are some people who might, so I'll concede that yes, there are some disadvantages to not having total control over your content. However, there are some benefits.

One, it helps us out with the **storage component (not so much the service part, the syncing and all of that) of Archival Storage**. We use the cloud as one of our storage locations, and the particular service provider we use makes it really easy to do fixity checks with content we have on local servers. So that's good.

We also get some **geographic distribution** that we wouldn't be able to have otherwise. Our local network storage provided by campus IT gives us some redundancy between Grand Rapids and Allendale, but using the cloud gives us the ability to have a storage location in a geographic area with a completely different disaster-threat: Oregon.

Lastly, I just wanted to mention that what we're doing falls completely within our IT **cloud computing guidelines**. We're not putting any sensitive research data on here, or university business data.

All of this being said about Archival Stroage, this is also probably the first area we're looking

to improve as we ramp up our infrastructure in this area.

Archival storage

Metadata

File formats

Our role

Metadata

- We're not using XML
- But we are being structured…
  - Tab-delimited TXT files
- …and consistent.
  - Well-documented guidelines
  - Using subsets of Dublin Core, PREMIS

Metadata, not really a lot to say here, just feel like I have to mention it because it's in my job title.

We're not using XML, or we're at least not doing a lot of coding in XML.

But we are being **structured and consistent**, which IMHO is what is really important.

For structure, we're using **tab-delimited text files**. Spreadsheets, essentially, which will make any future conversion to XML easier.

For consistency, we have well-documented **guidelines**.

I did also want to note that we're also not going really crazy with metadata. We use kind of a common sense core of Dublin Core for descriptive metadata and PREMIS for preservation/administrative metadtata that, again, IMHO, is a happy medium between too little and too much description.

Archival storage

Metadata

File formats

Our role

So, file formats. This is particularly important once you start collecting born-digital content.

File Formats

- Some history…
- Our number one priority: saving (and describing!) the bits
- Data Management:
  - Normalization for access
  - Normalization for preservation based on a number of factors
    - Basic vs. full-level preservation

This is kind of an interesting **story**. Being a recent[ish] graduate and new[ish] professional, without a lot of experience, I came into this job very nerdy about file format normalization and migration. I had the expectation that this was going to be part of our workflow for every single digital object.

It didn't take us long (some of us longer than others) for us to realize that giving this kind of Cadillac treatment to everything causes a huge bottleneck, especially because the whole operation is, again, pretty manual for us.

Now, we've kind of changed our tune. Right now, **our number one priority is saving (and describing) the bits as they are deposited**, getting the 1s and 0s and their associated metadata into Archival Storage. We think of the other file format related services as value added, but not necessities.

So now, this is kind of what **Data Management** (again, that's a capital "D" and "M") looks like for us.

We always **normalize for access**, and in fact, especially for audio and video we've taken some steps in recent months to make this easier for us.

Normalization for preservation, though, happens more on a case-by-case basis, and we've made a distinction between **basic vs. full-level preservation efforts**. Basic is more of a "save the bits" approach and with full we do more for normalization and migration.

In order for something to qualify as needing full-level preservation, we have to ask ourselves some questions. Some are relatively easy questions about the original format (including whether it's common, proprietary or compressed), but some are much harder, and try to determine the future value of a particular object.

Archival storage

Metadata

File formats

Our role

So, the higher-level one. Our role in all of this.

Our role

- Sometimes it's more educational than technical
- Be honest!
- Research data management example…

- Being a responsible archive

In this area, best practice would probably be to have a digital repository for all types of digital content created by GVSU, but this simply isn't realistic. **Sometimes our role has to be more educational than technical**, for example, informing people on best practices.

We also have to **be honest**, knowing what we can and cannot do and not making promises we can't keep.

I think the most clear cut example of this in our case is with **research data**. We simply don't have the resources for a full-fledged data repository that will meet all the needs, throughout the data lifecycle, of everyone doing research here on campus.

And while we some infrastructure in place in our Special Collections & University Archives to support digital preservation for certain types of digital objects (this is what I've been describing), this isn't a service we offer yet to scholarly and creative content like research data, the kind of stuff that typically goes into our IR.

And so our focus is more educational than technical. We focus our efforts on getting the word out about best practices for research data management, collaborating with other campus stakeholders who are involved in this process and getting into the research project cycle as early as possible, including the planning stages, even before any digital objects get created.

And we have to be honest with people. We have to tell them that while

ScholarWorks@GVSU, our IR, can help them meet some basic data sharing requirements, but if you're looking for long-term curation or preservation of your data we'll have to hook you up with someone else.

Sometimes, taking these approaches can feel disappointing. But actually, I don't think they're necessarily bad things. They're part of what it means to be a responsible archive.

INFORMAL AUDITS

Lastly, I just wanted to talk quickly about informal audits. This is something I think I first learned about at the DigCCurr Professional Institute in Chapel Hill, but it's really a good practice no matter your size or resources, or how far along you are into your digital preservation journey.

What are they?

A way of getting an idea of your repository's "trustworthiness" without going through a formal audit or certification process

Lowercase "T"

Why do them?

- Doing it together is better
  - Commiseration
  - Reassurance that you're on the right track
  - New ideas!
- Less scary than TRAC or TDR
- We've got some good models…

Yes, the idea of kind of "baring it all" to your colleagues at other institutions can be a little intimidating, but there are some good reasons to do informal audits.

For one, **doing all this together is better**. It really is, if for no other reason than it gives you some people to **commiserate** with. But, what is more likely to happen is that you'll get some **reassurance that you're on the right track**, and somebody could give you a **new idea** or potentially point something out to you that can make you're life easier.

Also, doing an informal audit is less scary than doing an actual audit, like Trusted Repositories Audit & Certification (TRAC), (that would be the capital "T") or getting a Data Seal of Approval. In fact, you can count on one hand the number of North American repositories who have actually gone through the process for getting a Data Seal of Approval and passed.

And, finally, there are some good models to use so you don't have to start from scratch…
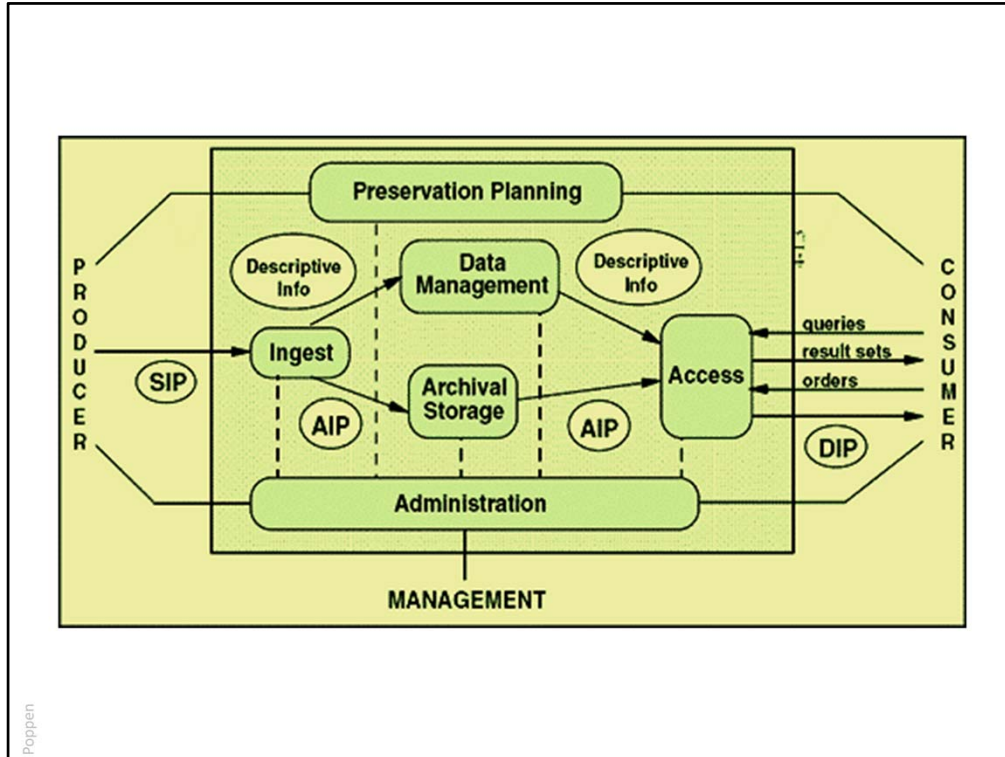
There is good old OAIS. I used this once on an informal audit with Brian Wilson at the Henry Ford Museum and Archives.

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

And there's also the Library of Congress National Digital Stewardship Alliance Levels of Digital Preservation, which I used more recently with these guys. Actually I think it was Lance's idea originally.

It's LC's happy medium between two extremes of a spectrum, with their personal digital archiving materials on the basic end and something like TRAC on the other.

It doesn't cover things like policies, staffing or organizational support, but one of the things I like about it is that it's overall structure is progressive with basic on the left and more complicated on the right. It's good for seeing where you are comparatively.

THANK YOU!