6-14-2023

Will our future selves thank us? An examination of born-digital curation practices at UKL

Megan M. Mummey
University of Kentucky, megan.mummey@uky.edu

Sarah Dorpinghaus
University of Kentucky

Ruth Bryan
University of Kentucky, ruth.bryan@uky.edu

Emily Collier
University of Kentucky, ebcollier@uky.edu

Will our future selves thank us?

An examination of born-digital curation practices at UKL

Best Practices Exchange
June 2023

Speakers

- Sarah Dorpinghaus, Director of Digital Strategies and Technology
- Megan Mummey, Director of Manuscript Collections
- Ruth E. Bryan, CA, University Archivist
- Emily B. Collier, Wildcat Histories Technology Consultant




You will be hearing from us in the following order – First you will hear from Sarah Dorpinghaus about shifting digital preservation infrastructure, then myself on implementing born digital appraisal, then Ruth Bryan on the acquisition of university publications, and then Emily Collier on web preservation. It may seem like we are all talking about disparate subjects, but each presentation will build on each other to form an in-depth case study of how we have been attempting to wrangle the beast that is working with born digital materials.

So if you know me – you know I have a tendency to say flip things (because I'm a youngest child so I'm always trying to get a laugh) like "that's future Megan's problem". But I've been an archivist for enough time now that I when run across problems, I get angry, and say "who did this!?"…it was "past Megan". So this panel came together upon the realization that we are all trying to not do this. We are struggling with various pain points, like time, expertise, understaffing, and trying our best to plan for the future in the constantly changing landscape around digital stewardship.

UK Institutional Context

Margaret I. King Library

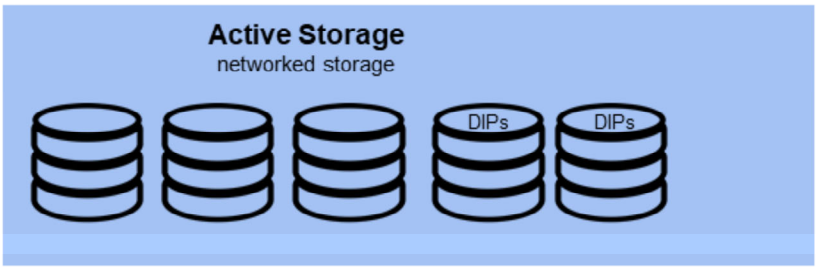
- UK is a research 1, land grant institution in Commonwealth of Kentucky
- UK Libraries uses Archive-it, Bit Curator, ArchivesSpace, Webrecorder
- UK Libraries' digital preservation repository and digital library are both home-grown systems
- UK Libraries has engaged in archival born digital work since 2015 and web archiving since 2018
- Digital preservation efforts are focused on primary source materials
- 1.3 FTE (100%) of time working with born digital archives, 8 FTE partial (5-15%), and on average 0-3 student employees

Before we dive in – we wanted to share a little of UK's institutional context. We are a research one, land grant institution in the commonwealth of Kentucky. We use Archive-it, Bit Curator, Archivesspace, and Webrecorder. Our digital preservation repository and digital library are home-grown systems. UK has engaged in born digital work since 2015 and web archiving since 2018. UK's digital preservation efforts are focused on primary source materials. We have around 1.3 FTE working with born digital archives 100% of the time, 8 FE who spend a partial amount of their time, and then on average 0-3 student employees depending on the ongoing projects. The 100% full time employee is extremely new – new as of June 1.

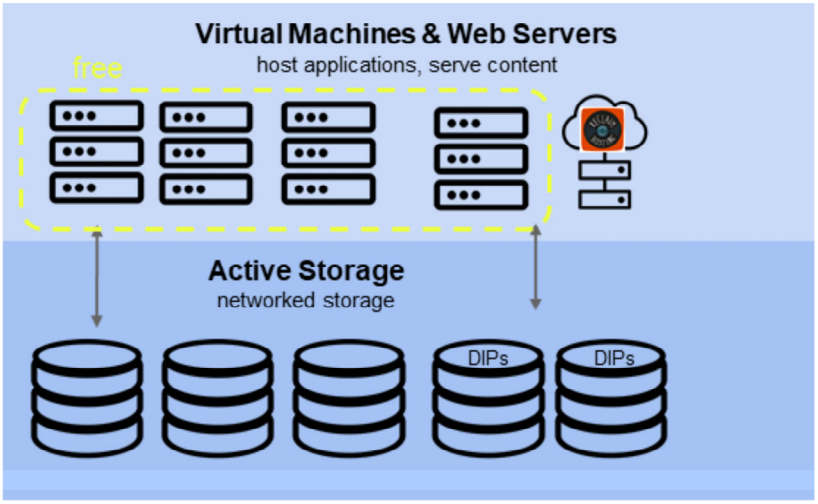
UKL Digital Preservation Infrastructure

A changing landscape, 2021-2023

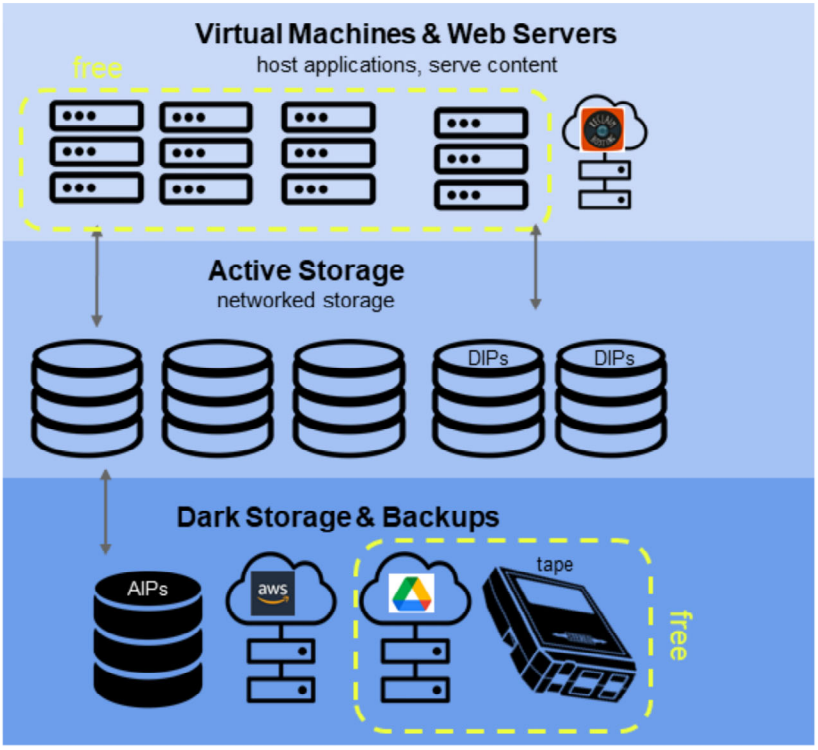
Start by walking through UKL’s digital preservation infrastructure - as of 2021



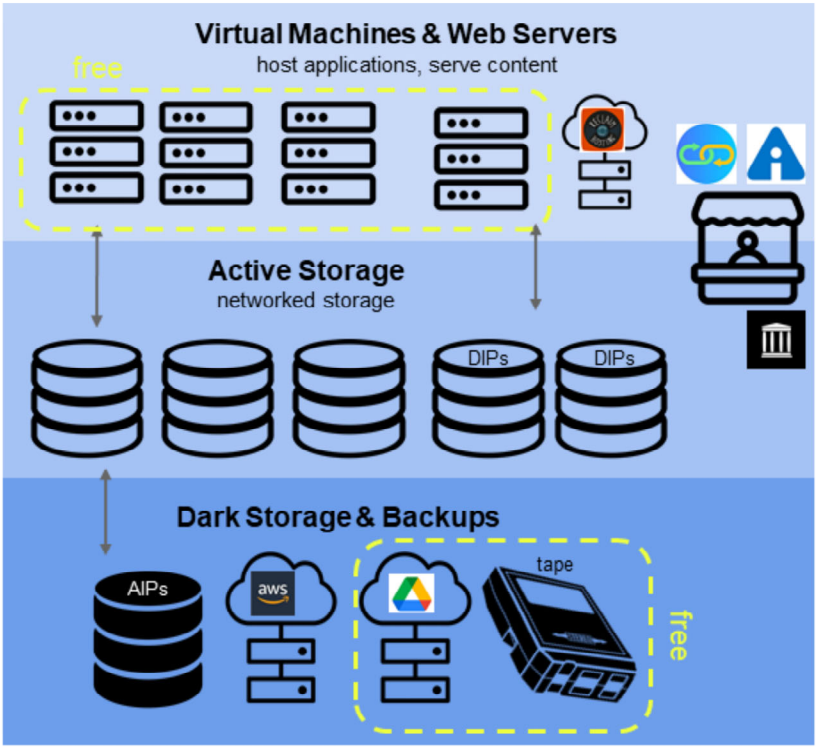
Active storage: holds working copies, scratch space for digital processing, includes our DIP store or access copies



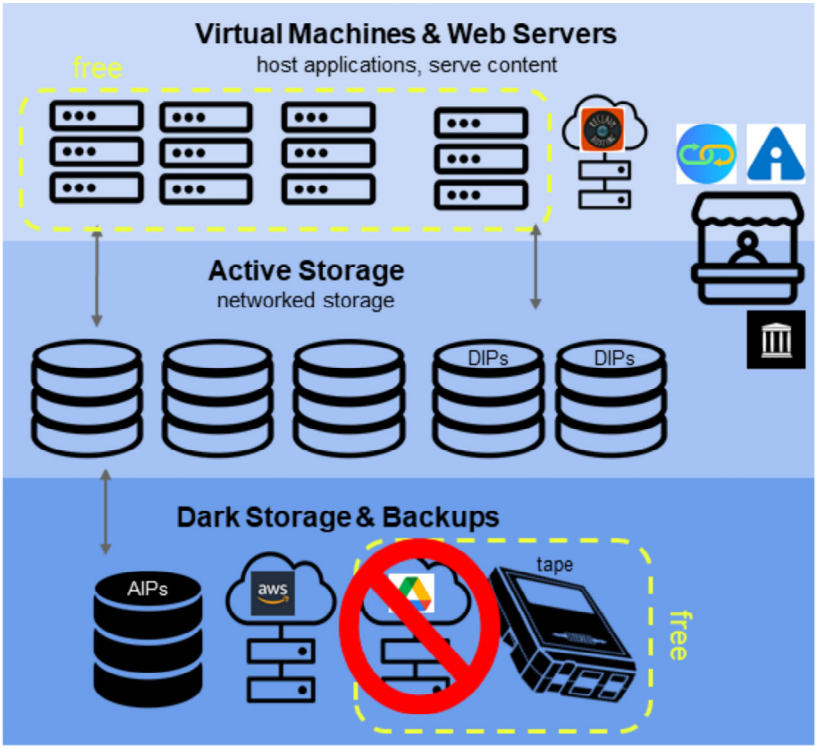
VMs & WS: host applications like ASpace and our digital libraries as well as tools that do some automated processing. Mostly UK-IT provided (for free), but we also have some vendor VMs, mostly Reclaim.



Dark storage and backups: AIP store on campus storage, AWS Glacier, GD (unlimited storage through campus contract) and campus tape -- GD and tape free! Glacier, $1/TB/yr

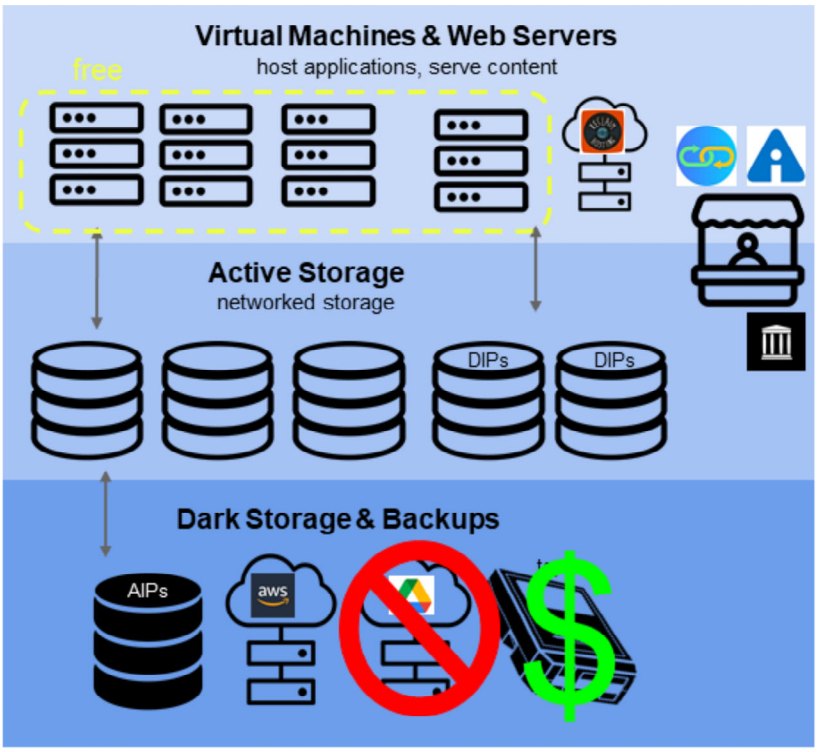


Vendor supplied tools relevant to this discussion: Archive-It and Internet Archive, Webrecorder



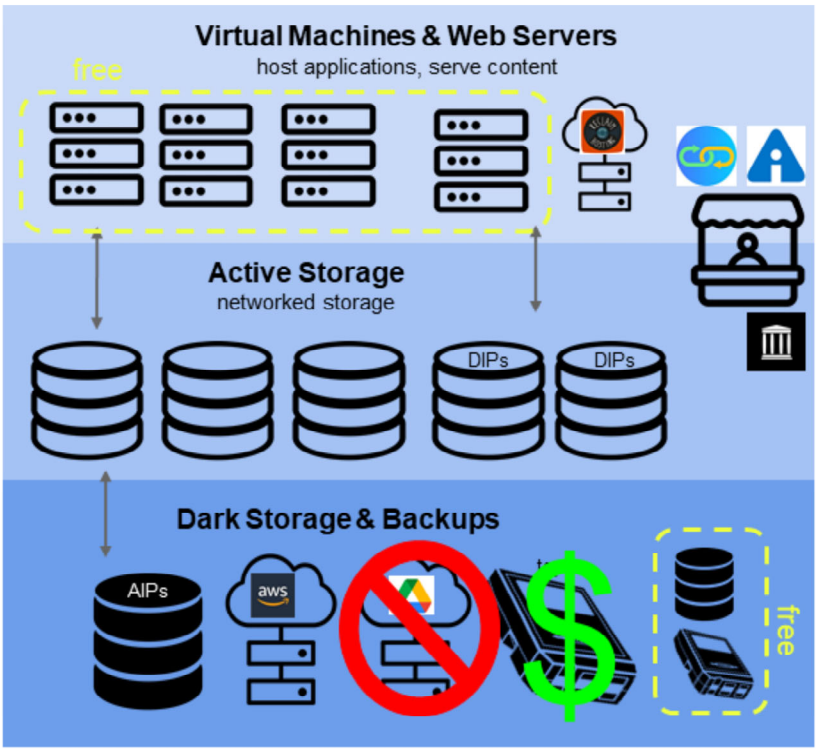
Change 1 (late 2021): Unlimited storage contract with Google would end by December 2022.

a. Unlimited to 15 GB / year
b. Removed a backup location



Change 2 (April 2022): Learned that we would be charged for for campus-supplied tape storage

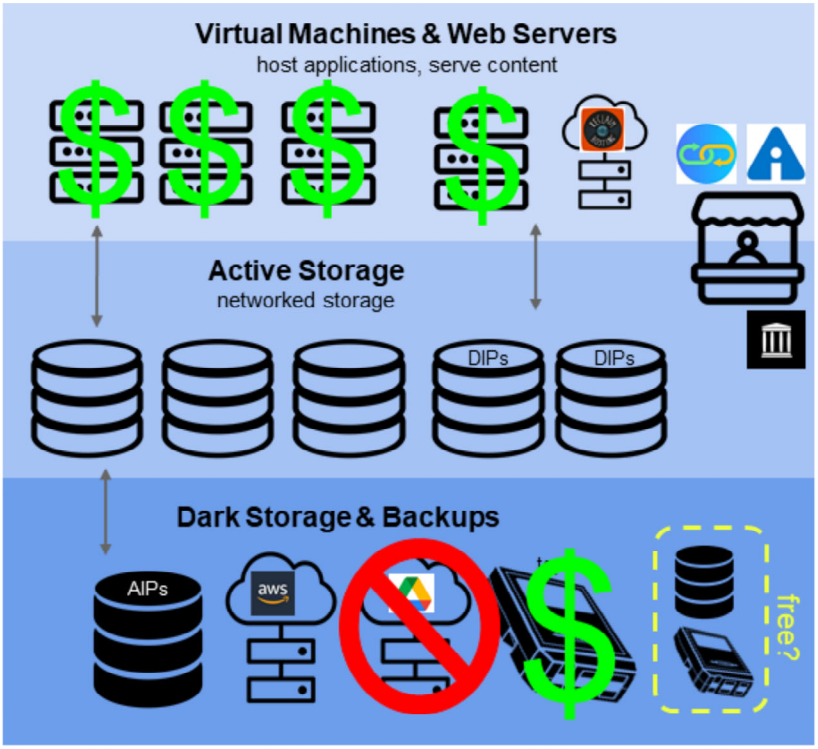
a. $70 / TB
b. Increases our costs by roughly 35% of our annual bill



Change 3 (December 2022): Learned about UK's new data protection services

i. Daily backups via automated client on 3 different types of storage media in 3 different regions
ii. All data is encrypted
iii. Fixity testing at least quarterly
iv. Tracking for obsolescence of storage and media

Meets the "sustain your content" level (top level!) for Storage category of NDSAs levels of preservation

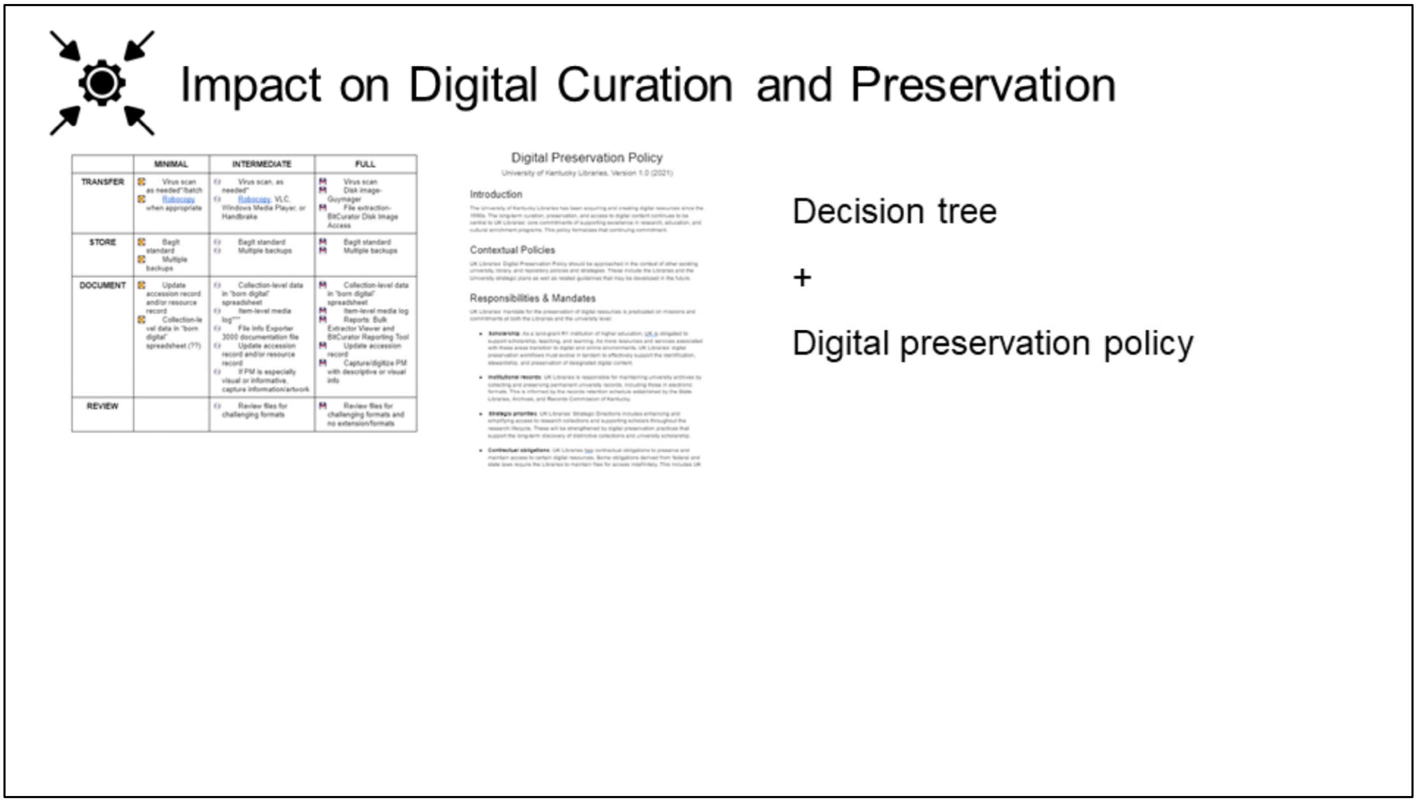


Change 4 (early 2023): New pricing structure announced for VMs and web servers

i. Pros: 5 year seamless migrations, support from IT, backups as previously described
ii. Cons: doubles our infrastructure cost

These changes within a year and a half have had and will continue to have major implications for our resource allocation, workflows, and how we approach our work. We've responded to each of these changes individually to address the top concerns or opportunities, and have yet to do a comprehensive review and restructuring of our practices to account for the new variables.

That aside, I want to share some of our immediate responses.

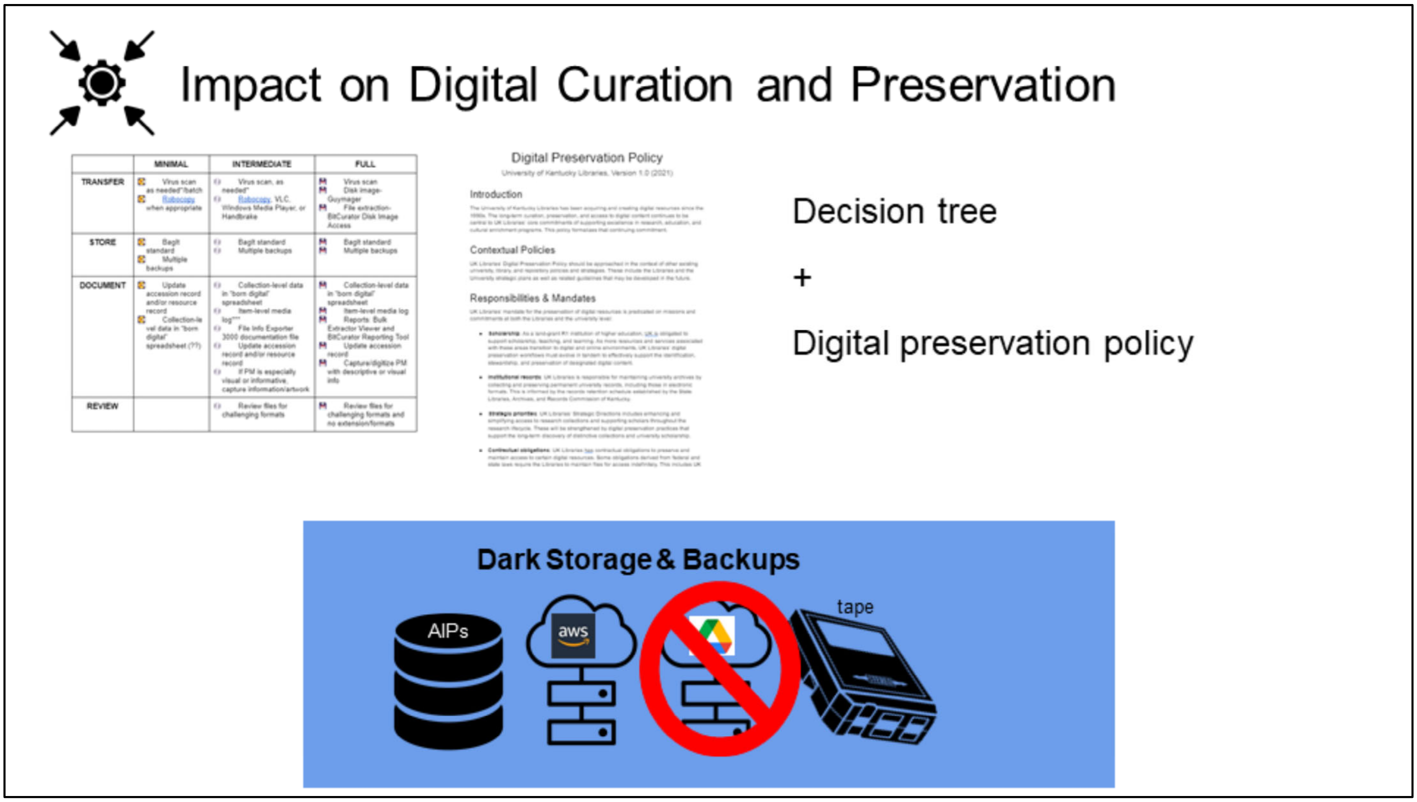


We were fortunate to have two key documents to help respond and decide the most responsible path forward:

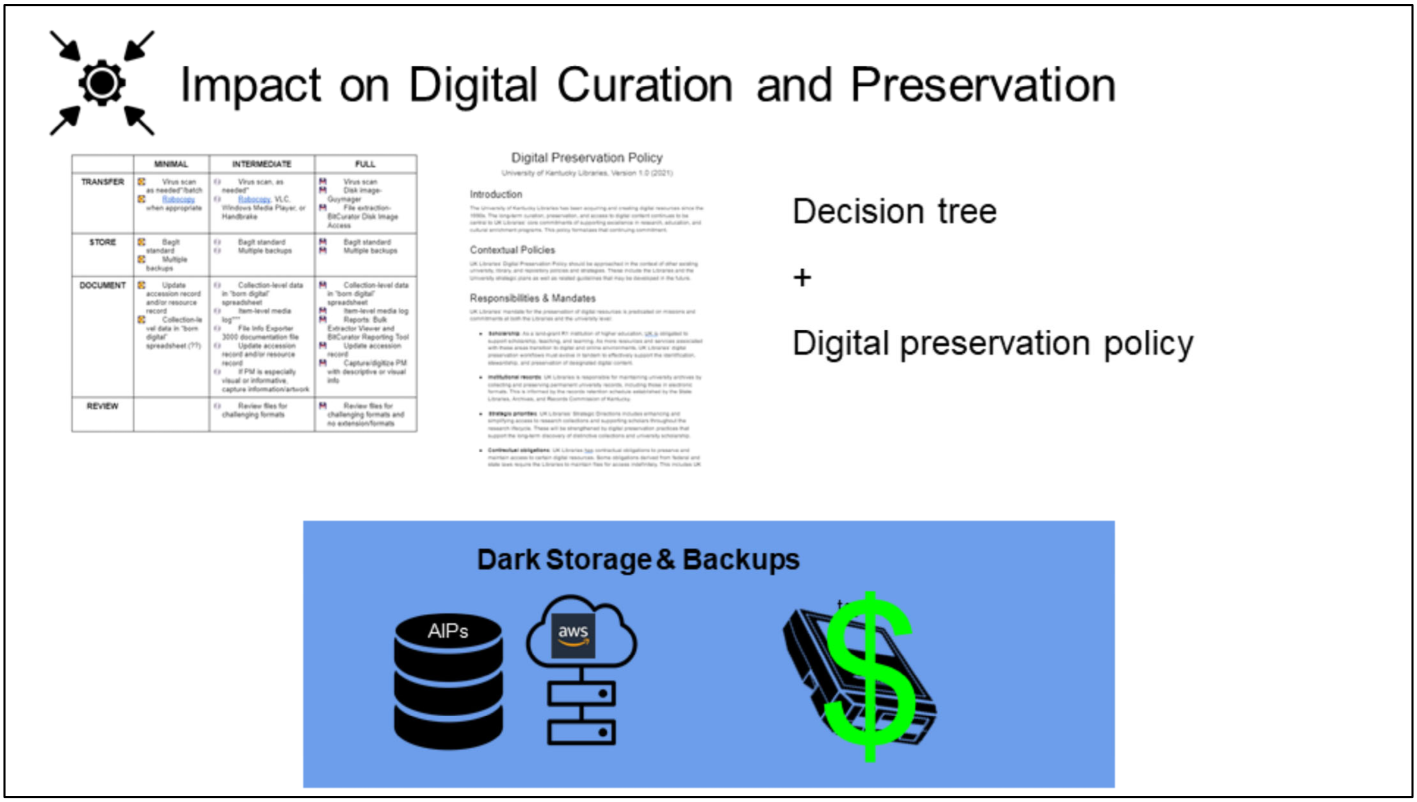
Decision tree *to help make appraisal decisions for our born-digital resources. Megan will discuss this further.*

Digital preservation policy *(approved in July 2021) that*

- *Identifies content that is out of scope for digital preservation*
- *Digital assets that do qualify and divided into 4 tiers with increasing levels of preservation*
- *Articulates an institutional commitment to digital preservation (we'll see how this holds up when we get the final bill from IT for our VMs)*



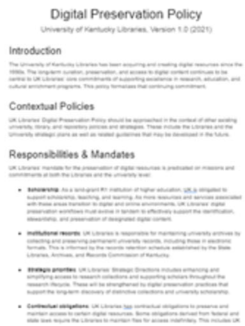
Thankfully, the sunsetting of unlimited Google Drive storage did not have a devastating impact on us-- we simply did a review of the content and cleared it. Our various backup locations and tracking made this relatively easy.



For the tape storage, we did some bulk purges according to our digital preservation policy. Born-digital content remained along with at-risk digitized content (mostly AV and photos). This reduced our footprint and bill by about ⅓.

Impact on Digital Curation and Preservation

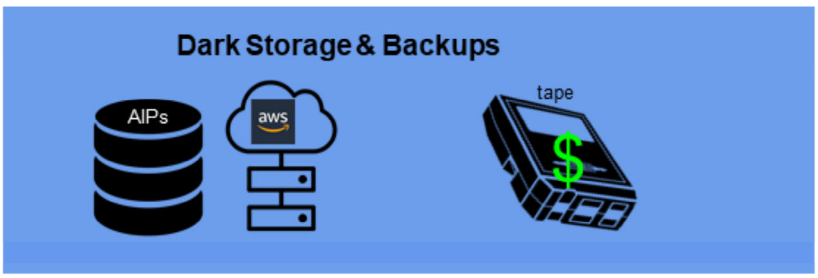
	MINIMAL	INTERMEDIATE	FULL
TRANSFER	Virus scan as needed*/batch Robocopy when appropriate	Virus scan, as needed* Robocopy, VLC, Windows Media Player, or Handbrake	Virus scan Disk image-Guymager File extraction-BitCurator Disk Image Access
STORE	BagIt standard Multiple backups	BagIt standard Multiple backups	BagIt standard Multiple backups
DOCUMENT	Update accession record and/or resource record Collection-level data in "born digital" spreadsheet (??)	Collection-level data in "born digital" spreadsheet Item-level media log*** File Info Exporter 3000 documentation file Update accession record and/or resource record If PM is especially visual or informative, capture information/artwork	Collection-level data in "born digital" spreadsheet Item-level media log Reports: Bulk Extractor Viewer and BitCurator Reporting Tool Update accession record Capture/digitize PM with descriptive or visual info
REVIEW		Review files for challenging formats	Review files for challenging formats and no extension/formats



Decision tree

+

Digital preservation policy

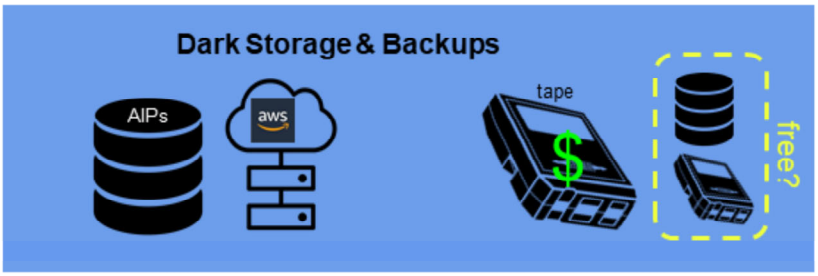


Impact on Digital Curation and Preservation

Digital preservation cost estimate

1 TB = approx $200 / year

(up from $152 / year)



Also, we now have updated cost estimates for preserving born digital content.

1 TB = $220/yr to preserve
30 TB photo collection = $6600 / year to preserve

Impact on Digital Curation and Preservation

Funding challenges

How does this impact appraisal?

Haven't had a chance to see how the increased cost of our VMs impacts our work, but I can imagine that this means less funds for digital preservation infrastructure

How do these prices impact our appraisal decisions? What capacity do we have for appropriate stewardship of collections? Megan will explore this further

Takeaways

Be flexible and nimble.

*Build systems and workflows that are **flexible and nimble.** Some things are truly out of your control, but the diversity of our infrastructure allowed us to respond quickly and weather these changes*.

Takeaways

Be flexible and nimble.

Build and maintain relationships with IT.

It is important to build and maintain relationships with IT- Lib IT and Campus IT

- *a. They will know about potential infrastructure options and solutions than you may not be privy to.*
- *b. How?*
  - *i. Try to identify the right people (may take a few attempts) Tell them about your work/needs. Touch base with them- keep them in the loop when considering utilizing vendor tools, cloud solutions, etc.*
  - *ii. Communication is especially important for your Lib IT. Even if they don't have capacity to take on management aspects, getting their perspective and recommendations is valuable. It's also important for them to know about the digital preservation and curation work that you're doing*

Takeaways

Be flexible and nimble.

Build and maintain relationships with IT.

Storage costs increase over time.

Storage costs will increase overtime. Collection acquisition rate outpaces any long-term decrease of storage costs. Robust infrastructure is expensive. Drive it home to administration that digital preservation will only require more financial resources as time goes on. Plan on 5-8% increase each year and adjust your budget accordingly

Takeaways

Be flexible and nimble.

Build and maintain relationships with IT.

Storage costs increase over time.

What other digital assets should be preserved?

Consider digital preservation needs beyond collections - backup CMS databases, code, etc.

All icons from the Noun Project https://thenounproject.com/

Hello again. Just a reminder that I'm Megan Mummey and for the next 10 minutes, I'll be talking about implementing earlier and more aggressive digital appraisal. The management of born digital workflows, policies and procedures only recently came under my management (about a year ago). So I am not a digital archivist, I am an all-around generalist who is an administrator as well as hands on archivist. To be pithy for a second, I consider myself a "fair-weather digital archivist. I do not always work with born digital records. When I do – because it is not my specialty I want to rely on established workflows. If there is not one, I say - well that is future Megan's problem. Our original program was very much focused on "grabbing all the bits" and we will figure out what to do with them later when they are processed (which realistically could be years down the road). Recently, I have become more aware of the environmental impact of digital preservation. This forced me to come to terms with being "lazy" with my born digital archives. And maybe "lazy" is not the correct phrasing. "Under-resourced" and "under-staffed" are better ways to frame it. During this presentation I am going to describe a little more institutional context around the processing of born digital records, present a case study on the acquisition and appraisal of a large collection of born digital records, and the resulting policy/procedure changes.

Institutional context

- 400+ TB of digital material (born digital and digital surrogates)
- ArchivesSpace used as collection management system
- Homegrown digital library and digital preservation repository
- Dedicated born digital workstation
  - BitCurator
  - CD/DVD, zip, and floppy disk drives
  - Specific software

	MINIMAL	INTERMEDIATE	FULL
TRANSFER	Virus scan as needed*/batch Robocopy when appropriate	Virus scan, as needed* Robocopy, VLC, Windows Media Player, or Handbrake	Virus scan Disk image- Guymager File extraction- BitCurator Disk Image Access
STORE	BagIt standard Multiple backups	BagIt standard Multiple backups	BagIt standard Multiple backups
DOCUMENT	Update accession record and/or resource record Collection-level data in "born digital" spreadsheet (??)	Collection-level data in "born digital" spreadsheet Item-level media log*** File Info Exporter 3000 documentation file Update accession record and/or resource record If PM is especially visual or informative, capture information/artwork	Collection-level data in "born digital" spreadsheet Item-level media log Reports: Bulk Extractor Viewer and BitCurator Reporting Tool Update accession record Capture/digitize PM with descriptive or visual info
REVIEW		Review files for challenging formats	Review files for challenging formats and no extension/formats Reformat copy of files with Quick View Plus if needed Search for PII and other sensitive information at a general level

So here is our institutional context. And I am just going to leave this up here for a quick second – since we covered quite a bit of it earlier. But I'll just point out a few relevant details. We have 400+ TB of digital material (which includes both born digital material and digital surrogates) We also have a dedicated born digital workstation that runs BitCurator, has CD/DVD, zip, and floppy disc drives. The program has been up and running since around 2015/2016. This is an illustration of our born digital migration decision tree. So, for migration and accessioning we have three tiers, full, intermediate, and minimal. After ingest we keep two copies of the files – the unorganized and essentially un-appraised files and then when they are processed – a copy of the processed files. So we have well developed workflows for migration and back up. We are less well developed in the areas of appraisal and description.

Environmental Impact of Digital Preservation

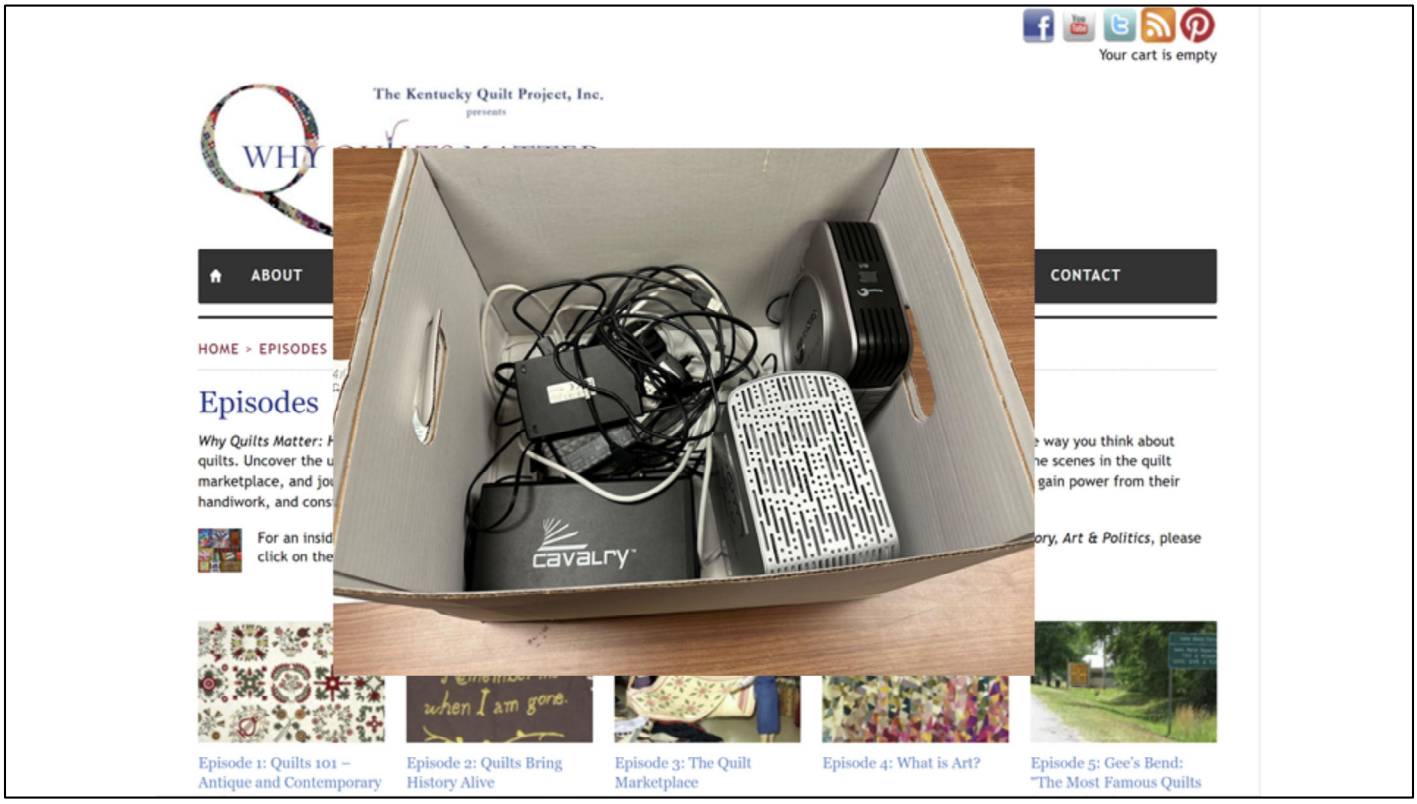
"When [the challenges of digital preservation] are confronted in an environment where staff time is scarcer than digital storage, it can be tempting to appraise digital content in a cursory manner."

Pendergrass, Keith L., Walker Sampson, Tim Walsh, and Laura Alagna. "Toward Environmentally Sustainable Digital Preservation." *The American Archivist* 82, no. 1 (2019): 165–206. https://www.jstor.org/stable/48659833.

Before we dig into the case study, I would like to share the idea that the work I am doing is rooted in. That Information Communication Technology (ICT) used by Cultural Heritage Organizations for things such as digital preservation and access has a negative environmental impact. I am not going to go into the arguments for why archivists should be aware of their environmental impact or why climate change is an issue or how big the impact of Cultural Heritage Organizations on the environment is. There is already plenty of scholarship on these issues. I am here today to share that reading this article on the screen and others like it made me aware that the decisions I make every day have an impact – and maybe there are things I can do to lessen that impact. I have highlighted a quote from the article here that comes from the digital appraisal section of the article. "When [the challenges of born digital preservation] are confronted in an environment where staff time is scarcer than digital storage, it can be tempting to appraise digital content in a cursory manner." That quote really resonates with me, and it makes me deeply uncomfortable, because this is what I have been directly doing. I, like I am sure many of you, also have very little time. And I wouldd just like to confront my privilege here – I'm at an R1 and I've never really worried about storage – though Sarah just outlined why I'm going to have to start worrying! As a "fair weather digital archivist" I tend to do things the easy way and follow a procedure that someone else has written and not critically examine what it is I am doing. I am having to retrain myself to confront the difficult tasks and decision points in my job – rather than kick that can down the road for someone else to deal with – because

kicking that can down the road has consequences.

In the Fall of 2022, myself and the director of the Nunn Center for Oral History, Doug Boyd – worked with a donor organization, the Kentucky Quilt Project, to bring in their records relating to a series of documentaries on quilts – Why Quilts Matter. The Nunn Center has a large collection of oral histories relating to quilting and this donation would bring in the original interviews done for the documentaries as well as the documentaries themselves. So everything I am going to talk about now has nothing to do with the management of the oral histories – those are being preserved and worked with very ably by my colleagues in the Nunn Center for Oral History.



We ended up with 4 record storage cartons, 6 hard drives, and 1 digital file transfer. The donor was very concerned that we get ALL the files, so she gave us everything she could find. The digital files added up to 3.5 TB. This is where normally Megan the “fair weather digital archivist” would have followed our documentation to a T and dutifully transferred the hard drives most likely using the minimal option on our migration decision tree, bagged everything, and backed it up in multiple places. However – sometimes following procedure to the letter – puts blinders on you. I am busy – we are all busy. And had I not paid attention to what I was preserving – I would have made a big mistake blindly preserving all of that data. So I decided to do some intensive appraisal.

Appraisal questions to ask

- What is the collection's archival value?
- Which records can we provide access to and approve the use of?
- Which records are essential to documenting this work/subject/person?
- Realistically – which records will patrons use?
- Do these files contain proprietary formats?
- Do we need to keep drafts of the documentaries?
- What at the bare minimum (given our capacity) should we preserve?

Here are the questions I asked myself during the appraisal of these records.

These questions led me to the answer that – these records have archival value because they document a documentary made in Kentucky on art – art traditionally done by women. We collect Kentucky documentary collections. So it fits in with an established collecting area. But that there were major problems with many of the files foremost among them the proprietary formats and use issues.

Hard Drives/filetransfers	Size	No. files	Contents	Decision
HDD3 (seadisk)	58.5 GB	15,815	Image archive and image working files	Don't keep image archive except for photographs clearly provided by Shelly
HDD4 (new volume)	58.5 GB	15,816	Image archive and image working files	Don't keep image archive except for photographs clearly provided by Shelly
HDD2 (edit 12 backup)	1.76 TB	4,071	Image archive, final cut pro files	final cut pro files can't use - non destructive footage - proprietary can't use - old final cut pro only useful if you have old final cut
HDD1 (edit 13 backup)	1.59 TB	12,872	Cache files, DVD encodes, and "george" footage	Keep b-roll footage if any can be found mostly photographs - don't need cache files
file transfer	1.49 GB	4,448	Shelly's working files, website files, along with image archive	Don't need to keep website files
HDD5 (Seagate)	119 GB	83,149	Image archive	Don't keep image archive except for photographs clearly provided by Shelly
HDD6 (Seagate)	367 GB	55	Episodes 1-9 WQM, transferred from HDCAM Videotape masters, mp4 and MOV files	Keep - highest quality versions
Received 6 Harddrives and 1 file transfer	**3.5 TB**	**136,226**		

This spreadsheet shows you what was on each hard drive and in each file transfer. I would like to take this moment to say on this slide that I've not yet finished the work with this collection – but these are the appraisal decisions for the collection to be carried out. We got one exact duplicate hard drive of the image archive. But we also got five different copies of the "image archive" – containing all the photographs used in the documentary as well as those considered for use. These copies are either only slightly different or WAY different from each other. We got the original final cut pro files, the producers working files, copies of all of the commercials produced for documentaries, the cache files, the DVD encodes, as well as the high-res copies of the documentaries themselves.

Hard Drives/filetransfers	Size	No. files	Contents	Decision
HDD3 (seadisk)	58.5 GB	15,815	Image archive and image working files	Don't keep image archive except for photographs clearly provided by Shelly
HDD4 (new volume)	58.5 GB	15,816	Image archive and image working files	Don't keep image archive except for photographs clearly provided by Shelly
HDD2 (edit 12 backup)	1.76 TB	4,071	Image archive, final cut pro files	final cut pro files can't use - non destructive footage - proprietary can't use - old final cut pro only useful if you have old final cut
HDD1 (edit 13 backup)	1.59 TB	12,872	Cache files, DVD encodes, and "george" footage	Keep b-roll footage if any can be found mostly photographs - don't need cache files
file transfer	1.49 GB	4,448	Shelly's working files, website files, along with image archive	Don't need to keep website files
HDD5 (Seagate)	119 GB	83,149	Image archive	Don't keep image archive except for photographs clearly provided by Shelly
HDD6 (Seagate)	367 GB	55	Episodes 1-9 WQM, transferred from HDCAM Videotape masters, mp4 and MOV files	Keep - highest quality versions
Received 6 Harddrives and 1 file transfer	**3.5 TB**	**136,226**		

So out of these things – how much can we as an institution actually provide access to and grant use to? – the answer was very little. We decided to only keep photographs from the donor's private collection of quilts as those are the only images owned by the donor. And just an aside that is where the quilt on the title slide came from! The rest of the images were licensed from other cultural heritage institutions, and then individual artists. And, no, that license did not include archiving the documentary records. We are keeping the lists of photographs considered which were generated by the documentary creators. As well as the image guides produced for each documentary – which list the image, the owning institution.

Hard Drives/filetransfers	Size	No. files	Contents	Decision
HDD3 (seadisk)	58.5 GB	15,815	Image archive and image working files	Don't keep image archive except for photograph clearly provided by Shelly
HDD4 (new volume)	58.5 GB	15,816	Image archive and image working files	Don't keep image archive except for photograph clearly provided by Shelly
HDD2 (edit 12 backup)	1.76 TB	4,071	Image archive, final cut pro files	final cut pro files can't use - non destructive footage - proprietary can't use - old final cut pro only useful if you have old final cut
HDD1 (edit 13 backup)	1.59 TB	12,872	Cache files, DVD encodes, and "george" footage	Keep b-roll footage if any can be found mostly photographs - don't need cache files
file transfer	1.49 GB	4,448	Shelly's working files, website files, along with image archive	Don't need to keep website files
HDD5 (Seagate)	119 GB	83,149	Image archive	Don't keep image archive except for photograph clearly provided by Shelly
HDD6 (Seagate)	367 GB	55	Episodes 1-9 WQM, transferred from HDCAM Videotape masters, mp4 and MOV files	Keep - highest quality versions
Received 6 Harddrives and 1 file transfer	**3.5 TB**	**136,226**		

Next on the chopping block are the final cut pro files. These are not just files in a proprietary format, but files in a 2011 version of a proprietary format. Also – the reasons we are preserving this documentary have nothing to do with the making of the documentary. They did nothing unusual or unique – we are keeping it because we have a collecting focus on quilts and Kentucky produced documentaries. These film editing files we cannot look at or use.

Hard Drives/filetransfers	Size	No. files	Contents	Decision
HDD3 (seadisk)	58.5 GB	15,815	Image archive and image working files	Don't keep image archive except for photographs clearly provided by Shelly
HDD4 (new volume)	58.5 GB	15,816	Image archive and image working files	Don't keep image archive except for photographs clearly provided by Shelly
HDD2 (edit 12 backup)	1.76 TB	4,071	Image archive, final cut pro files	final cut pro files can't use - non destructive footage - proprietary can't use - old final cut pro only useful if you have old final cut
HDD1 (edit 13 backup)	1.59 TB	12,872	Cache files, DVD encodes, and "george" footage	Keep b-roll footage if any can be found mostly photographs - don't need cache files
file transfer	1.49 GB	4,448	Shelly's working files, website files, along with image archive	Don't need to keep website files
HDD5 (Seagate)	119 GB	83,149	Image archive	Don't keep image archive except for photographs clearly provided by Shelly
HDD6 (Seagate)	367 GB	55	Episodes 1-9 WQM, transferred from HDCAM Videotape masters, mp4 and MOV files	Keep - highest quality versions
Received 6 Harddrives and 1 file transfer	**368.5 GB**	**circa 4,000**		

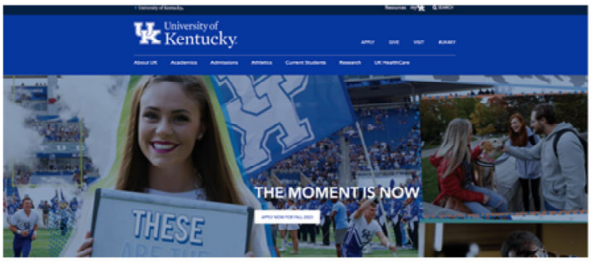
Then we have cache files, dvd encodes, and footage from “George”. We also need none of these things. Originally, the “George” footage was identified by the donor as very important b-roll for the documentaries. And upon closer examination what was the b-roll? Another copy of that image archive. Nothing we can use. Additionally, we plan to preserve the DVDs using handbrake and then discard the DVD encodes. This will give us the look and feel of the DVD without having all the disparate files that are not combined. We will be keeping the master files for the documentaries and the producer’s working files (scripts and other planning documents) minus the majority of the image archive and the website files. We will be preserving the website with our institutional copy of archive-it which is a better route because it preserves the functionality of the website. This brings us down to a much more manageable estimated 368 GB, taking us down 3 terabytes.

Takeaways

- Added in a more rigorous appraisal section to our documentation
- Saving everything is a trap. Do not do it! It's bad for archives and it's bad for the environment.
- Reappraise your born digital collections
- You do not always need fancy tools for digital appraisal

https://tinyurl.com/5n7kbk98

This experience led me to build out an appraisal section on our born digital migration decision tree. Here's a qrcode to our document – remember though this is a working document, so it is not very pretty and "official" looking. I included the appraisal section on our migration decision tree – because these things should go hand in hand. I want to make the following explicit in our documentation for those who engage in born digital work at UK, because there are quite a few, that they are encouraged to NOT SAVE EVERYTHING. Saving everything is a trap. It's bad for archives and it's bad for the environment. This experience has also encouraged me to embrace the concept of reappraisal – there are collections that I am now thinking about – that need to have their born digital records reappraised, because we saved too much. Also – you don't always need a fancy tool to tell you if you have duplicates in a collection. They can be helpful if you want to go down to the bits, but ask the question "do I need to take this down to the bits?". Sometimes taking the time to look more deeply works well too. So to wrap up – Don't be like I used to be – don't kick the can down the road and just "grab everything" because it might be useful some day. In the end I would say 90% of the digital records in this Why Quilt's Matter? Collection do not in fact matter and are not essential to preserve.



Suddenly, Everything's Online!

What Do We Do Now?

Ruth E. Bryan, CA
University Archivist
University of Kentucky Libraries
Special Collections Research Center

We're going to pivot now to look at some of our challenges and strategies for online document preservation and description, specifically for university records. "Suddenly, everything's online! What do we do now?" This is an updated version of a presentation I gave at SAA in August 2022.

In 2018, Sarah and I were successful in advocating for the UK Libraries to purchase a subscription to Archive-It. We were also able to hire Emily as our first web archives intern and began selecting university websites and crawling university seeds. At that time, I thought of web archiving as just one of many acquisition and preservation methods for university records and allied documents. We also thought that we would be able to do appraisal, and set up crawls for all types of web content, including a yearly crawl for the main uky.edu seed, and after that one push of work, it would all be fine. We established that fine = 75% of the website or web page is captured, so our quality assurance threshold is good enough rather than perfect.

SUDDENLY! (The Realization)

=

2019

Situation:

- Key records and papers
- Are distributed online only
- Not transferred to Archives
- Likely to be lost
- Web archiving is more central than before
- Preservation is complex
- Is thus resource-heavy
- Do we need to re-align effort?

But, by the next year, the realization washed over me that most key university records are being distributed or published online only and not routinely transferred to the archives the way they were in the past. The COVID pandemic accelerated this trend. So, web archiving is actually more important or more central than I had thought, because without proactively acquiring these documents, they are likely to be lost because of the ephemerality the web.

Yet, as we also found, web archiving is technologically complex. It's not as easy as setting up the crawls, doing some quality assurance, and they're good forever, which requires more resources than many other formats.

SUDDENLY! (The Realization)

2019-2021

- 2019-present: University history researcher
  - No regulations, 2005 onward
  - No Senate Council, 2006 onward

Here are some specific examples. Starting in 2019, a researcher investigating university administrative history requested records we didn't have in print or digital format, but were online, including superseded university regulations from 2005 onward, and University Senate Council meeting minutes from 2006 onward.

SUDDENLY! (The Realization)

2019-2021

- 2020-2021: COVID-19 web archives and "In This Together" documentation project
  - Social media and videos in blogs not captured using Archive-It

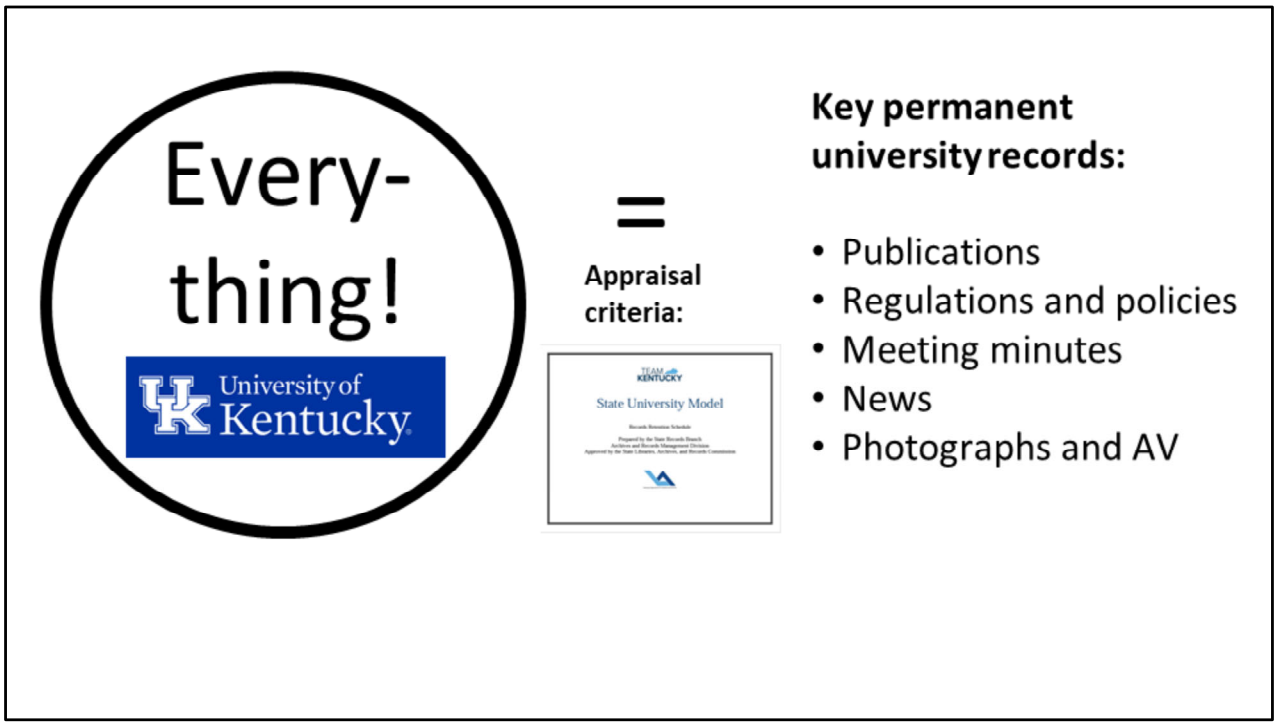
Further, for our COVID-19 documentation projects, we couldn't capture the donated social media accounts using Archive-It.

SUDDENLY! (The Realization)

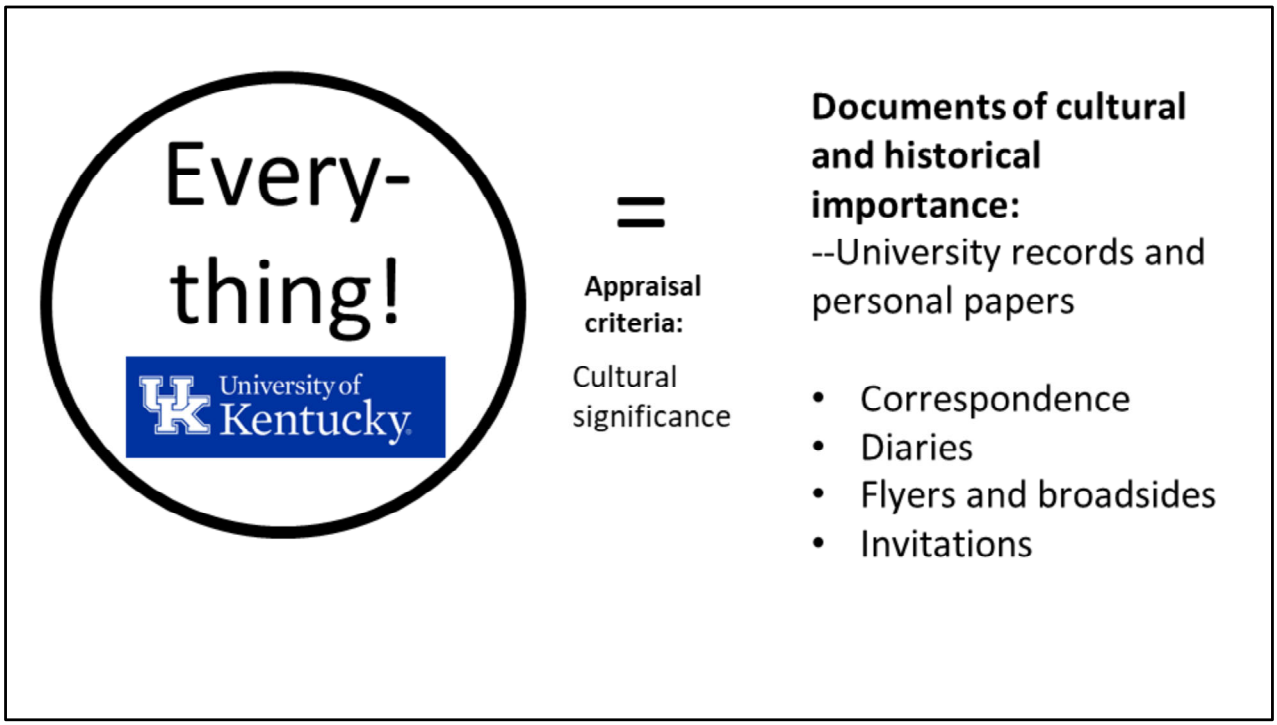
2019-2021

- 2019-present: Student and employee protest and activism
  - Latino Student Union
  - Black Student Union
  - Students for Black Lives Matter
  - United Campus Workers of KY

And in 2019, Daniela Gamez Salgado, the then-President of the Latino Student Union, donated some of the LSU records to the archives, mainly social media. We would also like to assist in preserving the contemporary voices of other social justice and activist student and university organizations, such as those listed here, most of which is also conveyed through social media or through complex websites that are difficult to crawl to meet our 75% threshold.



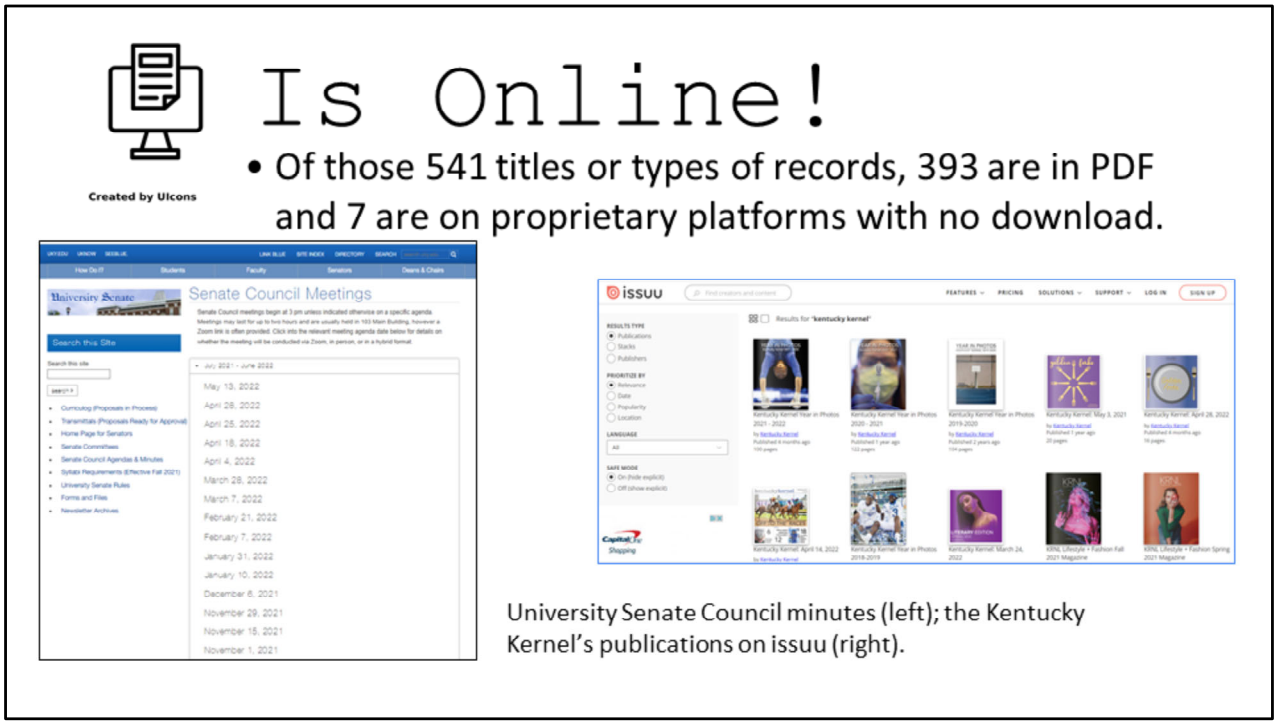
The “Everything” in the presentation title refers to key, permanent University of Kentucky records as outlined in the State University Model Records Retention Schedule. These permanent records provide documentation of the university’s decisions and actions, finances, and planning. Formats include publications; regulations and policies; minutes of university decision-making bodies; press releases, newscasts, and newsletters; and photographs and audiovisual recordings.



“Everything” also includes other documents of cultural and historical importance that the records schedule considers non-permanent, but that are crucial to documenting the experiences and activities of university units and individuals. They often provide a counterweight to the official or public stance or story of the university. Social media posts, blogs, podcasts, and video blogs are today’s correspondence, diaries, flyers, broadsides, and invitations.

In a review of the UK website that student Alex Reaugh started in 2021 and that I continued in 2022, we found 541 titles or record types across 58 university units. Since most of these titles or types of records are produced serially, there are many more individual titles.

In 2018 for our Archive-It subscription request, I counted 70 social media accounts for high-level administration, colleges, student government, athletics, and fraternity and sorority councils. This is obviously the tip of the iceberg, as it doesn't include student organizations, for example.

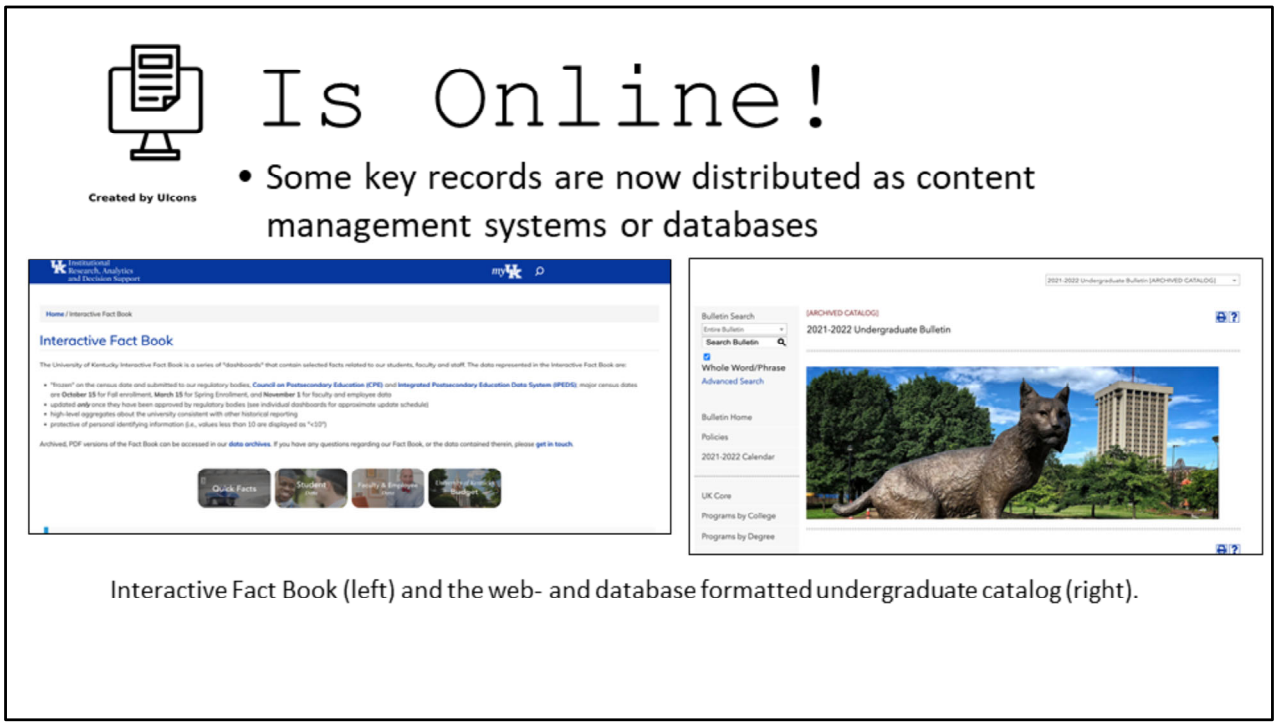


The ways in which these key university records and other documents of historical value are distributed online varies considerably. For example, of those 541 titles or types of records, 393 are PDFs and can be downloaded or easily crawled, like the University Senate Council meeting minutes. On the other hand, not many but some important publications are distributed on proprietary platforms with no download option like the Kentucky Kernel "Year in Photos," which is now the only yearbook.

online journal by UIcons from Noun Project

Of those 541 titles or types of records we found in our survey, 122 are websites or blogs. UKnow, the university's press releases, is an example. We crawl UKnow monthly as a stand-alone seed, and this content seems to be captured well. However, embedded YouTube videos are often not captured, such as the one on the "Kentucky Can" capital campaign area of the Office of Philanthropy home page.

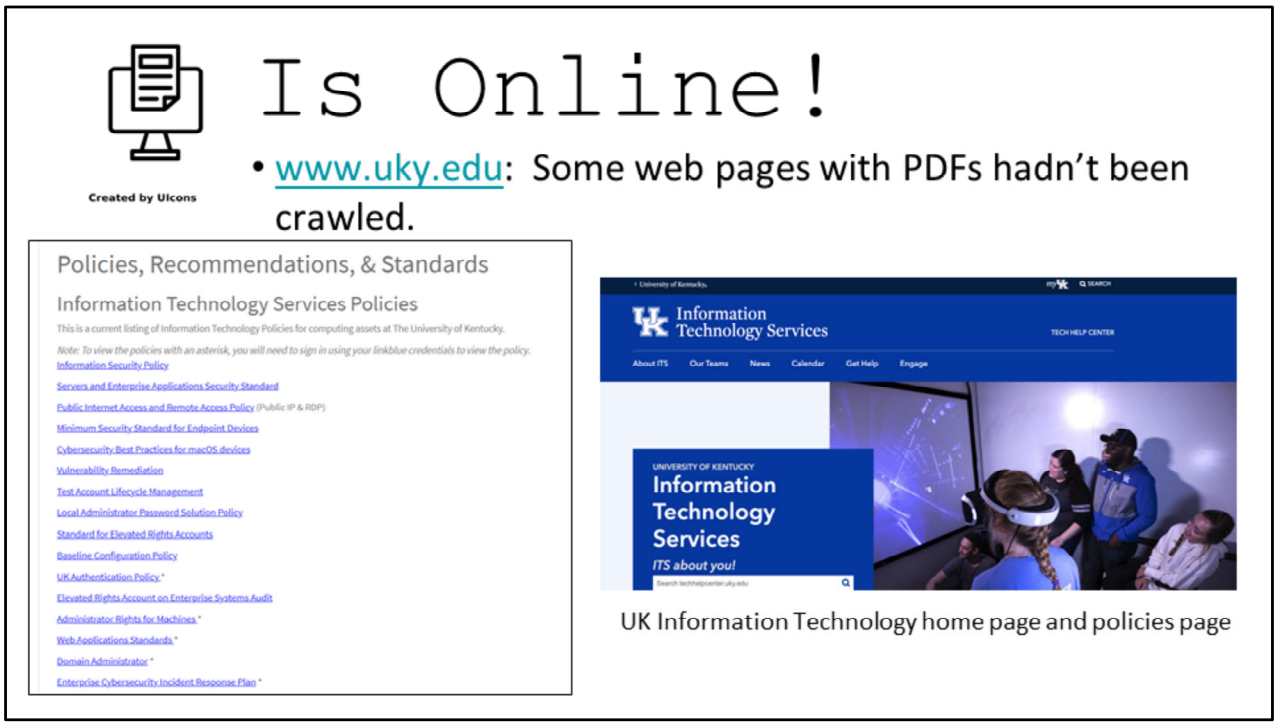
online journal by UIcons from Noun Project



Interactive Fact Book (left) and the web- and database formatted undergraduate catalog (right).

And, of the 541 titles or types of records, a few of those that used to be print publications are now distributed as content management systems or databases. This allows users to interact with the data, but makes capturing the information by web crawling impossible. Two examples of this are the UK Fact Book and the Undergraduate course catalog.

online journal by UIcons from Noun Project

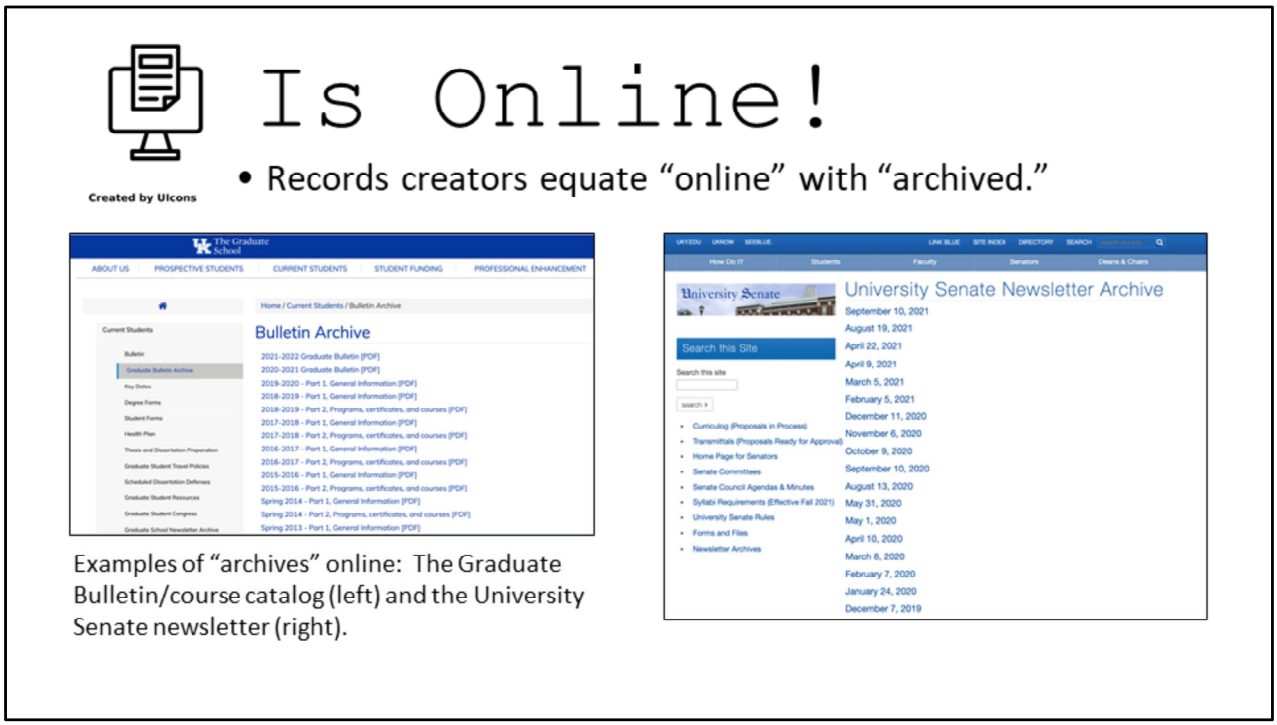


As I mentioned, when we first scoped the uky.edu website for crawling, we thought it would be fine to crawl that main seed yearly. I thought most content was being captured until I reviewed it in 2022. For example, the Information Technology web pages hadn't been crawled, including all the university's IT policies and standards. Clearly, there was something happening with the crawl, the scope, or changes to URLs over time.

online journal by UIcons from Noun Project

The dynamic Web 2.0 functions of social media platforms are difficult to capture and preserve, including infinite scrolling as demonstrated by this Latino Student Union Twitter account.

online journal by UIcons from Noun Project



And, currently, records creators believe that putting documents online is the same as archiving them, so there's no need to send a copy to the archives. The archivist must now proactively search for and acquire these records.

online journal by UIcons from Noun Project

What Now?

- Are web-based documents too expensive to acquire?
  - Only acquire the easy ones?
  - Or only from the highest administrative levels?
- Give up idea of acquiring comprehensively
- De-center the records retention schedule as an appraisal tool
  - Records created by powerful units will most likely survive without intervention
  - Re-align preservation and description priorities

Boxes everywhere!

What Now?

Are web-based documents too expensive to acquire and maintain, even if their content is important? The acquisitioning of physical materials and born-digital records hasn't slowed down, and the legacy collections that haven't been adequately preserved and inventoried are still in the backlog. Should I focus acquisition resources on those documents that cost less to acquire, like the PDFs that can be downloaded and static web pages? Or only on the records from the highest university administrative levels? For now (i.e. 2022 to 2023), I've decided not to limit web archiving for a couple of reasons:

Even with print-format records, in the 20th century when the university was smaller, the archives never acquired permanent university records comprehensively as outlined in the records schedule. Even with all the problems I found, preserving online university records has the potential to be more comprehensive and less subjective than how we acquired university records in the past.

Another key thing to consider is that the University of Kentucky is a predominantly white institution with a long history of racism and prejudice against people of color in the community and on campus. The records schedule is a risk management tool that serves the interests of the university and privileges the records of its most powerful units and people. It's probable that the records they create will be more likely to survive without

archival intervention. De-centering the records retention schedule as an appraisal tool would allow me to focus my scarce resources on those documents created by underrepresented people and organizations, whose voices and actions are crucial to preserve, in order to diversify the archival record and to provide a countermeasure to the “official” viewpoint of the university elite. These documents are predominantly social media accounts, which are the most difficult.

What Now?

- Acknowledge: collection management takes longer
  - Constant research
  - Constant change
  - Item-level work

Time Trials Excerpts (minimum effort in hours/stage)

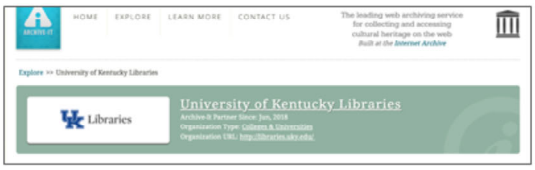
	Analog/born-digital accessions not on media	Social media crawls
Donor/creator communication + paperwork/appraisal	2-3	2-3
Write acc. record	.5-3	.5-3
File download/folder directory/upload file/bag	.16	
Set up crawl		1
Research issues		1-4
Totals	2.66-3	5.8-15.3

Based in these appraisal criteria, I think it's important to focus on web-based documents. So, I need to acknowledge that acquiring and managing online records and papers just takes longer.

This table is an excerpt from time trials for various acquisition and collection management stages for a variety of formats. Analog and born-digital accessions not on physical media take the shortest time and social media crawls take the longest. I can share the full table with those that are interested.

What Now?

- Re-allocate or find additional resources and partnerships
  - Re-assign student budget
  - Step back from funding physical or born-digital collections
  - Grant funding (Project STAND "Wildcat Histories")
  - Serendipitous relationship building: Public Relations!
- Information and advocacy

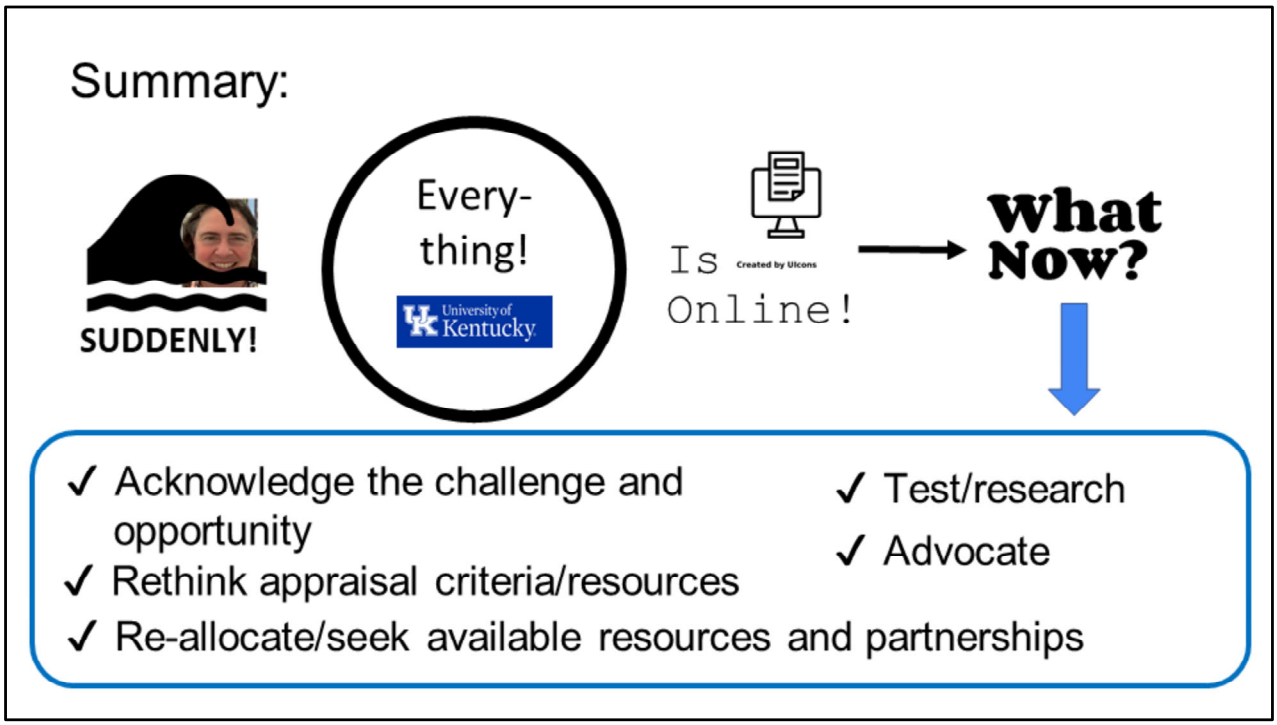


UK Libraries Archive-It banner

So, since acquiring documents distributed online takes longer, I need to find and/or allocate additional resources and get creative about partnering with others. For 2022-2023, I have been able to request that most of my student budget be converted to continue to employ Emily as a part-time web archiving specialist. This means stepping back from other formats and collections for now. We were also small mini-grant from Project STAND to work with the Latino Student Union on their social media accounts, which is what Emily will be talking about.

Because of our web archiving work, a web developer in the UK public relations office got in touch with me and is willing to help with preserving web sites! We just started this partnership.

Researching and writing this presentation provided an opportunity to gather information and approaches to realign my own effort, at least for this year, which I have immediate control over, and to share information with others.



So, to recap, starting in 2019, I suddenly realized that key university documents are now being distributed online only. Managing these web-based documents is complex and requires additional resources that many analog and born-digital records. Given that I have scarce time and money, what do I do now?

First, I acknowledge the technological and resource challenge of online formats, and the opportunity their acquisition provides for a wider, stronger presence of voices and content in the historical record.

Second, I rethink appraisal criteria, moving away from the university records schedule, prioritizing web-based documents, and more carefully quantifying the resources required for collection management.

Third, based on my appraisal, I re-allocate the resources I already have access to, at least for this year, and I seek out or respond to additional resources and relationships.

Fourth, I continue to test and research to refine resource requirements and appraisal criteria.

Fifth, I use the research, testing, thinking, and practice to advocate for more support.

Even a small step means preserving key records, but your collection policy, institutional context, and existing resources will determine what “key records” are for you! Your future colleagues will thank you!

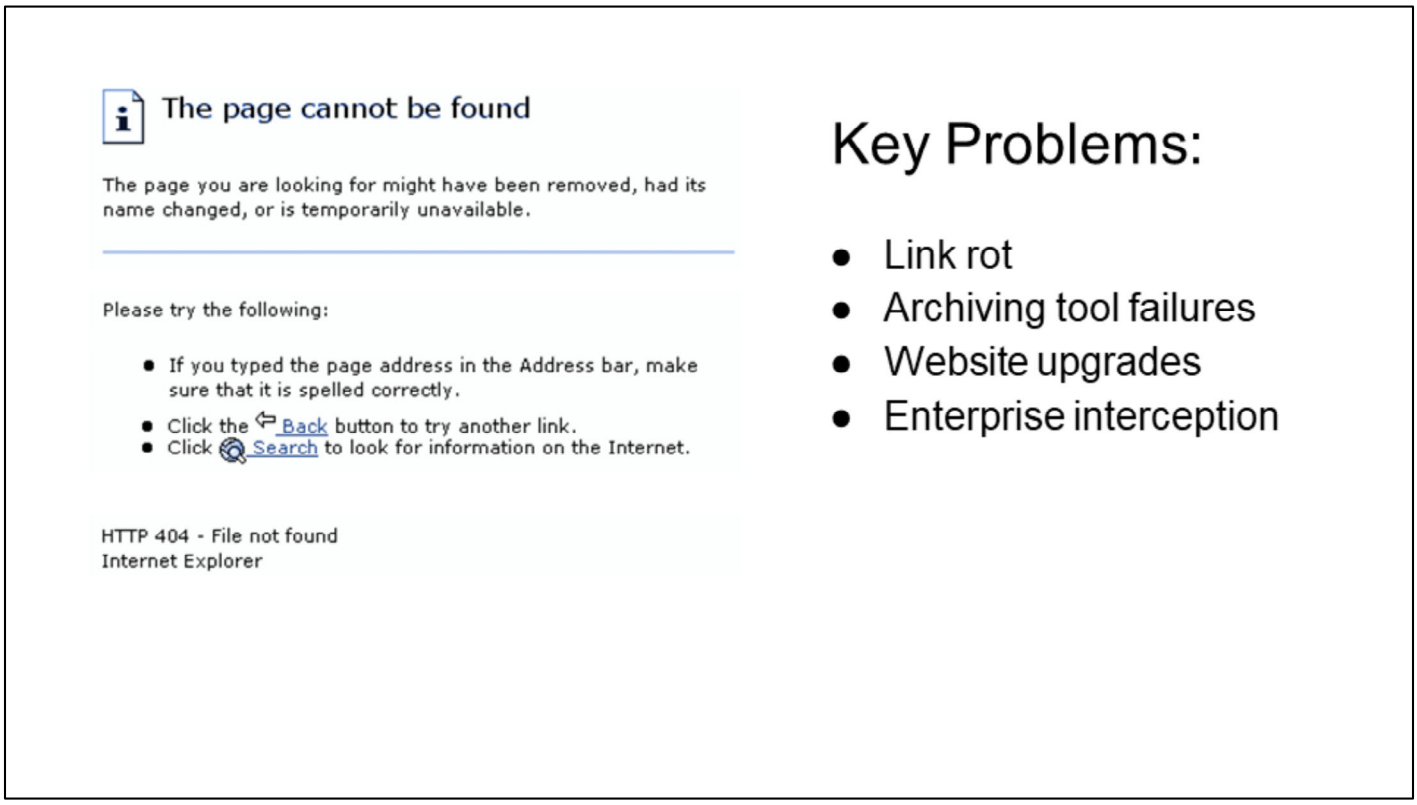
online journal by UIcons from Noun Project

Wildcat Histories:
Preserving Activist UK Student Organization's Legacies

Emily Collier
Web Archiving Technology Consultant
Special Collections Research Center
University of Kentucky

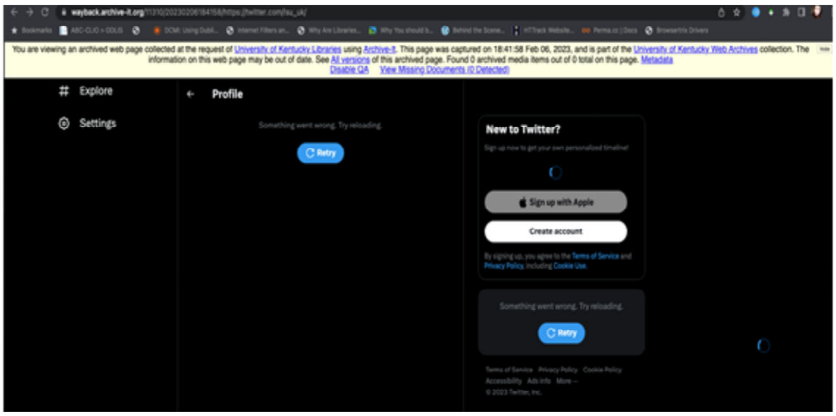
Seeds belonging to the Latino Student Union on UK Libraries' public Archive-It page

So Ruth has given a wonderful overview of the beginnings of our web archiving program and I will be diving a little deeper into our partnership with the Latino Student Union and our efforts to archive the cultural heritage found in social media, a much more tricky and volatile type of resource to capture and preserve.



So some of us who have worked with web archiving know many of the key problems when facing any website, including link rot, archiving tool failure, upgrades to sites like the inclusion of dynamic or interactive content, and of course when website proprietors make changes to their sites or install things like permissions. SItes can be changed, moved, taken down, become hidden behind paywalls or logins, and couple this with the struggle of web archiving tools to capture dynamic scripts, and you can really end up in the weeds.

Social Media: Our Notorious Evader



A failed Twitter crawl with Archive-It

WHY?

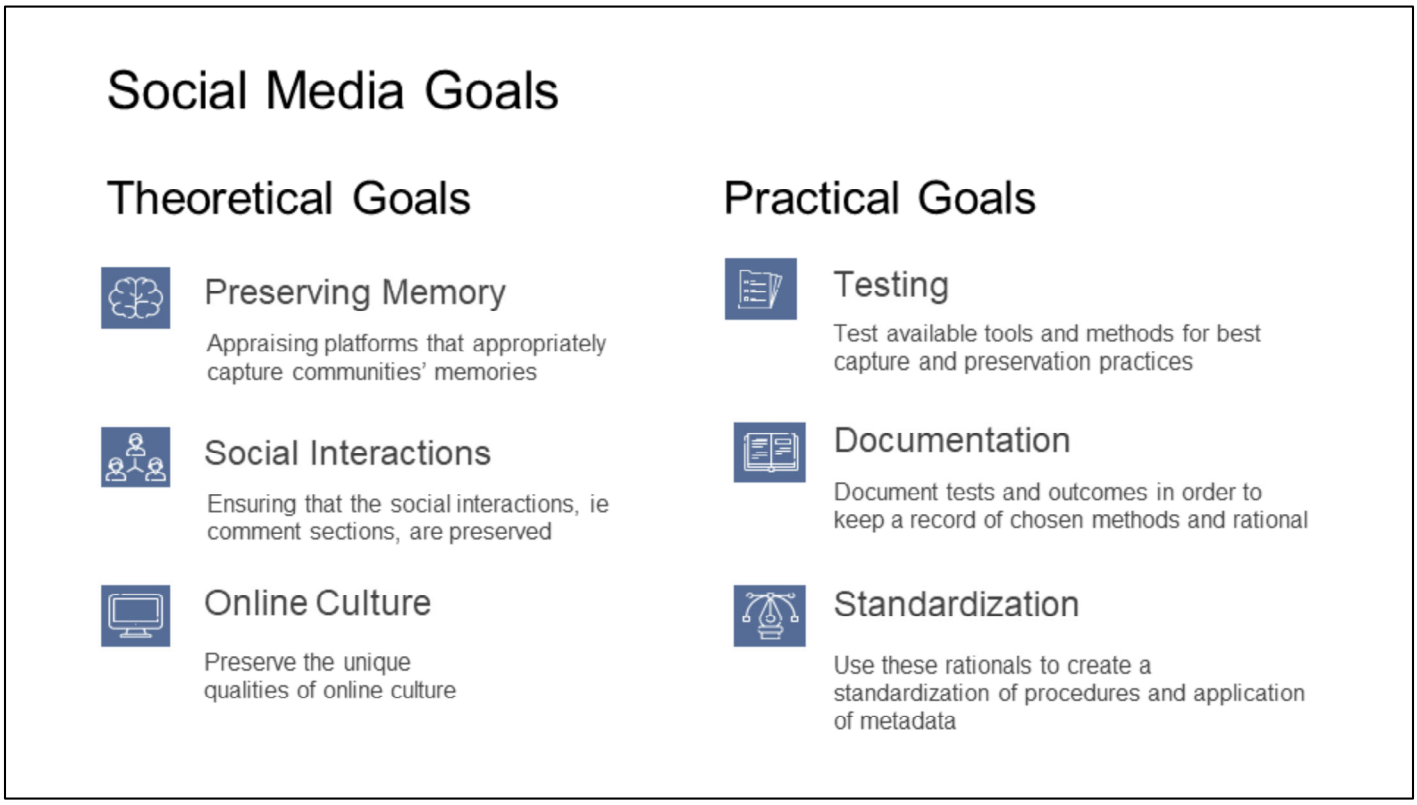
- Complex, interactive scripts
- Crawler traps/infinite links
- Near-constant updates create an arms race with preservation technology
- Crawler blockers and permissions obstacles

And It has been an absolute challenge attempting to capture social media. It is just notoriously difficult. Most of the scripts are complex and interactive, which is just the nature of any web 2.0 platform. Social media also often contains crawler traps, such as infinite links, meaning we have to be more careful about our scoping practices or else we end up with a lot of undesired content. Social media sites are also more subject to enterprise interception.These sites go through constant updates and changes to formatting and interactive features and it really is an arms race for many web archiving tool developers to ensure their tools are able to work through those updates. What's more, many of these sites specifically have crawler blockers included in their scripts and also prevent content from being viewed without being logged in. Archive-It has had quite a hard time with this and Twitter (and social media in general) so one of the exciting parts of our student organization partnership is the opportunity to research other options.

Wildcat Histories: A Flagship

- A Project STAND (**ST**udent **A**ctivism **N**ow **D**ocumented) mini-grant
- IMLS and Mellon Foundation funding
- April 2022-Aug 2023
- Partnered with the Latino Student Union
- Use the partnership as a pilot for building procedures on archiving student organizations' online content, specifically social media

So here we are with our Wildcat Histories project. Project STAND has been around for about 5 years and attracted Ruth's attention due to its focus on ethical documentation of student activism in marginalized or underrepresented communities. As she already had a working relationship with the Latino Student Union and a small collection of their materials, she approached them for a partnership. We received a $14,000 grant through Project STAND funded through IMLS and the Mellon Foundation for work to be completed from April 2022 through August of this year. The goal was to use Wildcat Histories as a flagship for building successful procedures for archiving student organization's online content, specifically social media.



So here I have outlined the project goals for my part. The theoretical goals include Preserving memory, right. Understanding and appraising those social media sites to ensure we are appropriately capturing the voices of the LSU members. It's easy to grab extra content when web archiving so this first step ensures I am capturing the voices without grabbing links that go too far out of context. It is also important for us to preserve social interactions on these sites, so the comment section being a really great place for this. It gives us context. It gives us a better understanding of the opinions of community members and it also allows us to see trends in thoughts. All of this culminates into preserving the online culture of a group, which is distinct. AND it really gives us a unique perspective into the functions of a group. Social media allows for pictures and videos and conversations that you just don't get with printed meeting minutes or flyers. This leads me to our practical goals. My role is to test the current tools available in order to find the ones that work the best for capturing those theoretical goals. I'm also keeping documentation of these tests and their outcomes as this allows me to a chance to decide which methods and practices are the best. The rationals I make will then go into creating a standardization of procedures and best practices for capturing social media sites.

Student Archiving Goals

- Empowerment: Return control to student organizations
- Community: Build positive and productive relationships between students and Special Collections
- Sustainability: Create solutions to integrate archiving into student organization policy

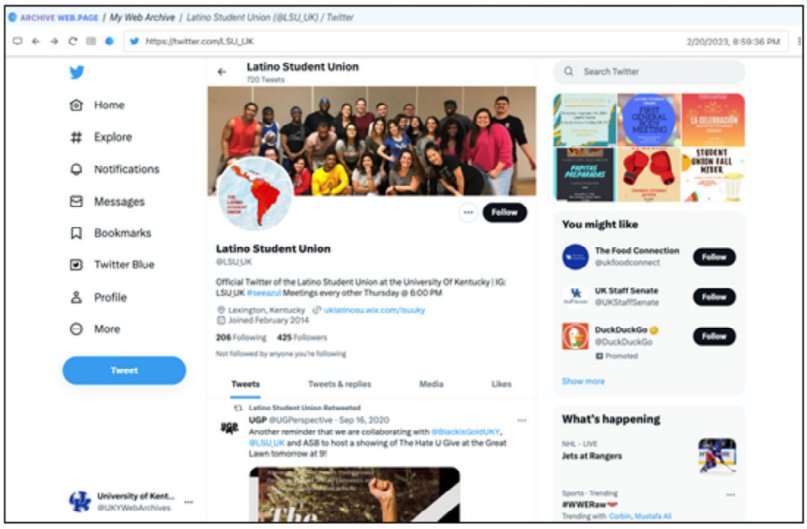
Informational flyer created by our Student Organizations Communication Assistant, Claudia Benito

We also have another goal, and that is to get the student groups themselves invested in their own archives. This is important because it gives control of preserved content to the organizations themselves allowing them to pick and choose what they want preserved. It builds a positive and productive relationship between these groups and Special Collections, and it also increases the chance of sustainability with more hands on-deck. If we can find the easiest methods and tools to use for archiving, match that with a successful delivery of the value of using archives to support an organization's legacy, the more likely the students will be take on their own archiving practices.

Building Forward-Thinking Practice Around Anticipating Failures

Gaps can be left in collections when crawling technologies fail.

Anticipating these failures and preparing backup tools or policy can alleviate this burden.



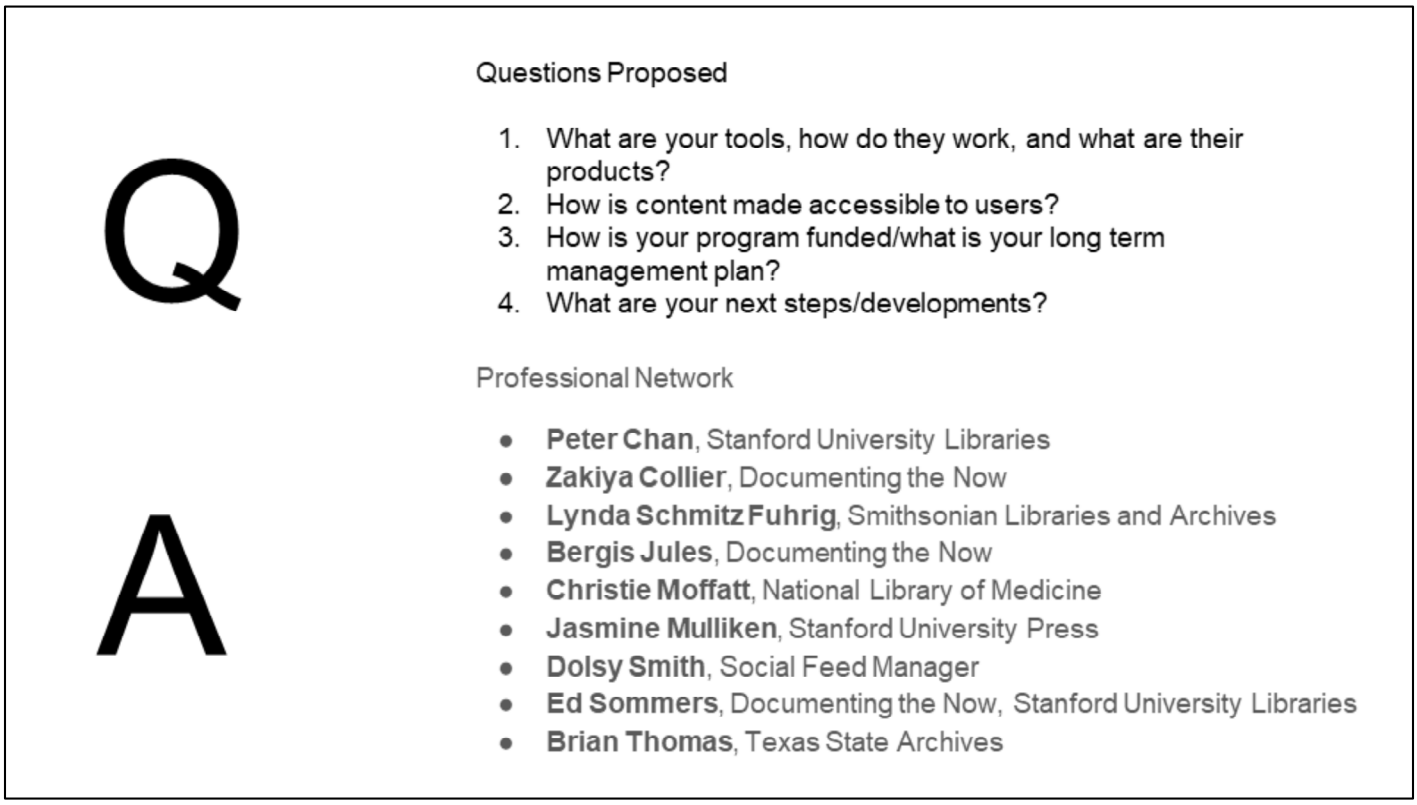
The same Twitter page, successfully captured with Webrecorder

But we need to find those easy solutions first. As I mentioned before, web archiving tools frequently fail with social media, and as tool developers work to keep up with these changes, the down time can result in gaps in our collections. So we want to build a more forward-thinking practice by expecting failures. By anticipating gaps and thinking about how our policy and practice can adapt, we can better prepare ourselves for failures. We are doing this currently by researching more web archiving tools and building a sort of backup arsenal to turn to turn to when Archive-It fails. So Webrecorder, specifically their archiveweb.page plugin, has so far been quite successful where Archive-it fails. Earlier, I showed a slide with a failed Twitter crawl with Archive-It, and here is that same page I managed to get with Webrecorder a couple weeks later. I'd like to stress here that Webrecorder was a tool we had looked in 2019, but had dismissed at the time because we weren't super impressed with it.

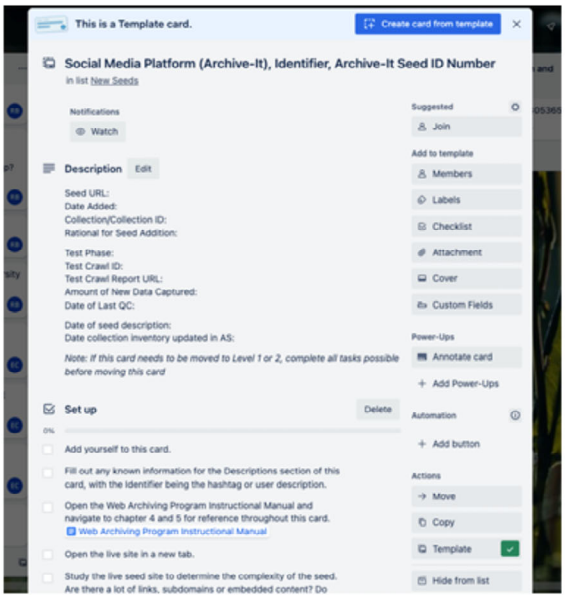
Redundancy in Research

- New technologies always coming into the field
- Existing technologies typically successful may fail in the future
- Old technologies once dismissed may have upgraded to better tools

And this brings me to one of the great take aways of the Wildcat Histories project, which was to learn about the value of redundant research. Web archiving is extremely volatile work. New technologies are always coming into the field, but almost just as importantly, old technology we once dismissed may evolve into much better tools. You can't just look at a tool and totally trash it if it doesn't seem like it will work for you because down the line, it may turn out to be just what you need. Especially when your existing tools start to fail.



But we do have to be careful about what tools we adopt. My position is grant funded and I work very minimal hours. I don't have the time to invest in tools that won't work for us. So rather than take hours to investigate and test new tools, we have used the Wildcat Histories project as a chance to reach out to a professional network of archivists and developers to ask them about their experiences. What tools have they created or used that worked well for them? What kind of outputs do their tools create and how are those outputs made compatible with their existing collections? How is the content in those collections made accessible to users? If we speak with developers, we also want to know how their tool is being funded for maintenance, support and development. The sort of elephant in the room about open source tools is that they can be great solutions, but they can also very risky and not cost effective even if they're "free". If the main developer retires, will that tool go away? Will someone take over? What if the tool was grant funded and the money runs out? If I've invested my time in a tool, I've invested money. If the tool fails and I have to do an overhaul, good chance I've wasted some money. That being said, I'd like to put in here that if an open source tool is truly valuable, the benefits may outweigh the risks and it is important to support these tools so they can be further developed and maintained. It really is about finding a balance and what works best for your institutional needs.



Trello card with steps for crawling social media sites with archive-It

Maintaining Simplicity

- Adopting complex technologies or practices makes long-term management difficult
- If a staff member leaves and takes knowledge with them, how is the work maintained?
- Needs to be simple for students to understand and use readily

For us, one of our institutional needs is maintaining simplicity. If we adopt too many tools, it's going to be too burdensome to maintain workable practices and documentation with our limited resources. And because we do have limited resources, another thing we have to keep in mind is the skill-level of technology we can properly maintain. Will we always have a staff member that can use complex tools? If that staff member leaves and takes all their knowledge with them, how do we maintain the work? I mentioned my position is grant funded or has been funded through Ruth's efforts at reallocating funds, but the end is always a threat on the horizon. So it is extremely important that I do everything with the end in mind. The simpler and more supported the tools are that I adopt into our practices, the more sustainable when my time is up. I have also crafted our technology workflows around a REALLY thorough set of procedures and VERY descriptive (and pictured!) instructions in order to enhance longevity of our web archives. Basically my goal is that almost any new staff member should be able to read my documentation and complete the very base work satisfactorily, even students. And this has been a major part of Wildcat Histories. We want students to take an active role in archiving their own content. Web archiving can be difficult, particularly with social media, but if we can determine the right tools and if the methodology is simple enough, even students can understand it and use it readily. Special Collections will always be here to problem-solve and provide updates to changes, but we want the students to feel confident about using web archiving technology and empower them to take part in preserving their own legacies.

Themes

- Documentation is important
- Record procedure or policy changes, *with* rationales
- View workflows as learning-in-working, or practiced-based and adaptable, rather than a set of rigid tasks
- Flexibility is key
- Relationship building can open doors; sometimes these relationships emerge serendipitously–look out for them!
- Digital preservation requires funding and personnel
- Short-term and ongoing re-appraisal and re-prioritization and re-allocation of existing funds

Future directions and questions

- Reappraisal of born digital materials already migrated and preserved, but not processed
- Access tools for web archives, including data harvesting tools
- Resource allocation and workflows for description and capturing are different for web-based content vs. other born-digital content
- Is appraisal criteria for permanent and culturally significant online university records different from other born-digital content?
- Should we apply digital preservation criteria retroactively? This requires additional development for our tools