Wrangling Workflows

ossarcflow

investigating, modeling and testing workflows for libraries and archives to curate born-digital content

Rebecca Russell
Woodson Research Center
Rice University
#ossarcflow

OSSArcFlow project slides courtesy of Educopia

Introduction

OSSArcFlow: A two-year IMLS grant exploring …

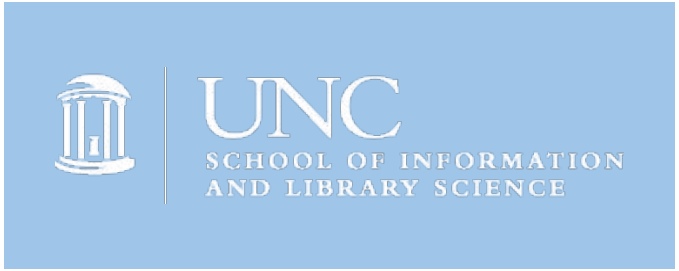
Open Source Software Archival Workflows with BitCurator, ArchivesSpace, and Archivematica

Research Questions

- How can institutions combine tools to support workflows that meet local institutional needs?

- How can institutions implement “handoffs” between systems that perform different functions on the same data?

Project Team



OSS Systems

BitCurator

Forensic disk imaging
File system analysis and reporting
Identification of PII

ArchivesSpace
a community served by LYRASIS

Supports core collection management
Authority control
Event tracking & reporting

archivematica

Lower barrier to digital preservation
Standards compliant - OAIS+
Microservice approach

Partner Institutions

academic libraries

MIT Libraries

EMORY LIBRARIES & INFORMATION TECHNOLOGY

MOUNT HOLYOKE COLLEGE 1837 PS.CXLIV:XII

Atlanta University Center Robert W. Woodruff Library

DUKE UNIVERSITY LIBRARIES

Stanford University LIBRARIES

NYU

RICE

public libraries

New York Public Library

DC public library check it out!

and more!

Kansas Historical Society

UNC THE ODUM INSTITUTE

Digital Curation Dossier

- Ahead of the partner meeting on December 4-5, 2017, project partners created digital dossiers outlining the form, function, and future of digital curation at their home institutions. https://educopia.org/research/ossarcflow

- Snapshot of each partner's digital preservation landscape

WOODSON RESEARCH CENTER
RICE UNIVERSITY

REBECCA RUSSELL

OVERVIEW

Woodson Research Center (WRC) is the Special Collections and University Archives for Rice University. Rice University is a private research university with an undergraduate focus located in Houston, Texas.

There are 5 professional Archivists in the department. 4 staff members work directly with digital preservation efforts, and have SAA-DAS certification. As a small department, we do not have dedicated curatorial foci, but do 'a bit of everything' including digital preservation. We typically spend about 20% of our time working on digital curation. A staff member could easily to do this full-time.

Project Timeline

Year One ...

... Year Two

Written Workflow Representations

Procedural Narratives ...

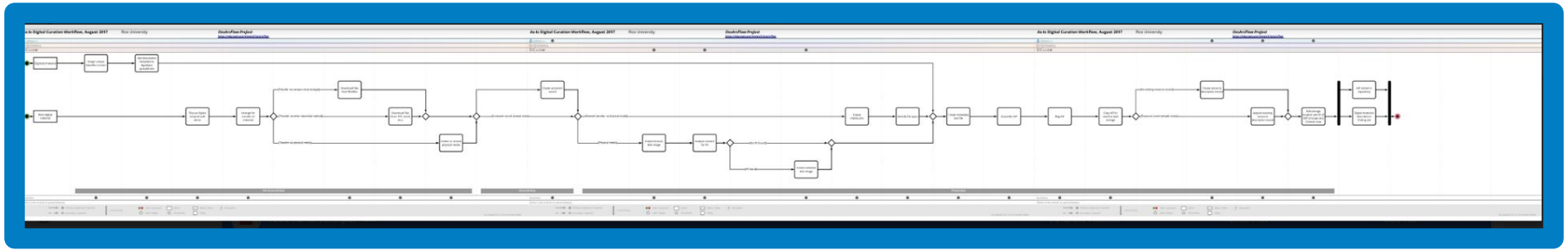
- **TRANSFER/ACCESSIONING**
  - Digital content is acquired:
    - [Digital Archivist or Archives Staff] IF on physical storage media THEN
      - [Archivist] takes picture of the media
      - IF filesystem is known THEN:
        - Triaged from physical media using either Guymager or Forensic Toolkit (FTK) Imager
      - IF filesystem is unknown THEN:
        - Triaged from physical media using Kryoflux
    - IF
  - Data is mo

Tabular step-by-step details ....

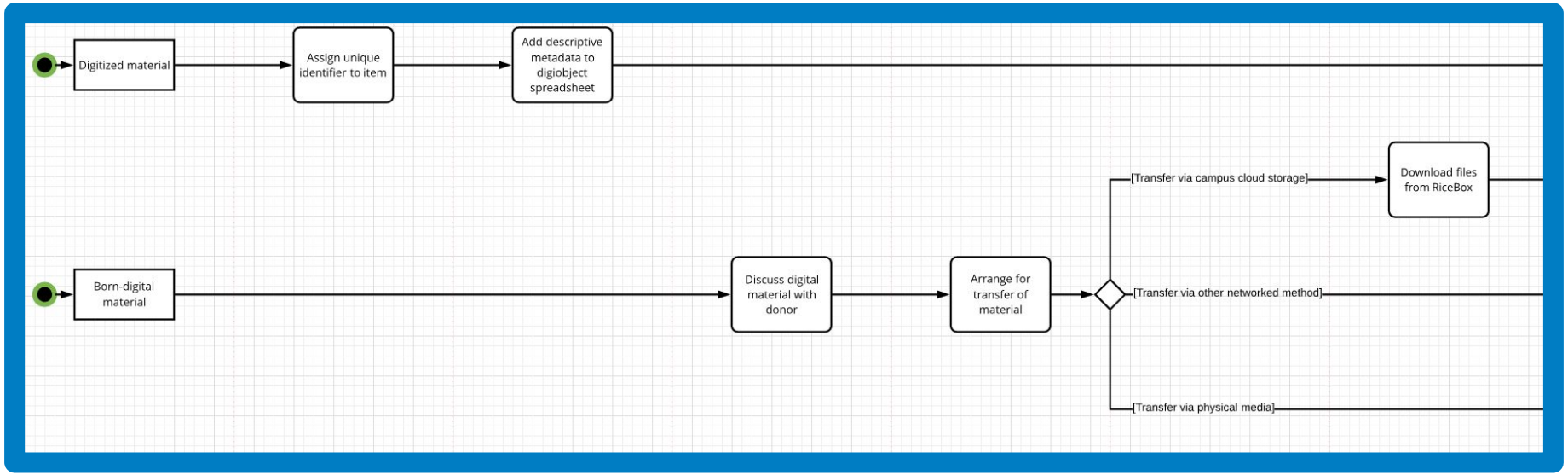
phase

A	B	C	D	E
phase	**step**	**description**	**hardware**	**software**
pre-accessioning	high-level content analysis, donor negociation	[Michael and/or Curator] Appraisal meeting/donor negotiation using a survey instrument that gathers information about what the collection is and what it contains.		survey instrument (Google Sheet? Excel spreadsheet?)
transfer/accessioning	photographic documentation	IF materials are on physical storage media, THEN [Digital Archivist or Archives Staff] takes a picture of the media		
transfer/accessioning	forensic disk imaging	IF materials are on physical storage media AND the filesyetem in known THEN forensic disk image is created		Guymager in the BitCurator Environment or Forensic Toolkit Imager
transfer/accessioning	forensic disk imaging	IF materials are on physical storage media AND the filesyetem unknown THEN forensic disk image is created		Kryoflux

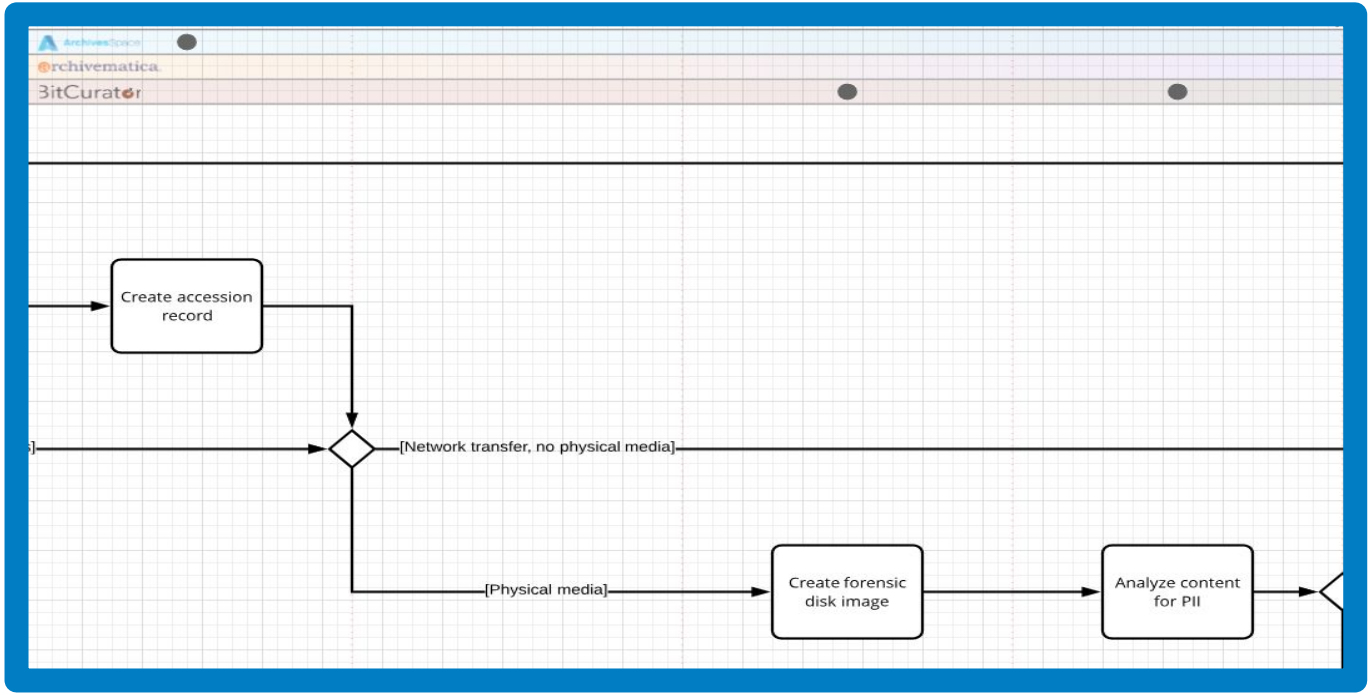
Visual Workflows



Visual Workflows



Visual Workflows



Visual Workflows

Some Preliminary Findings …

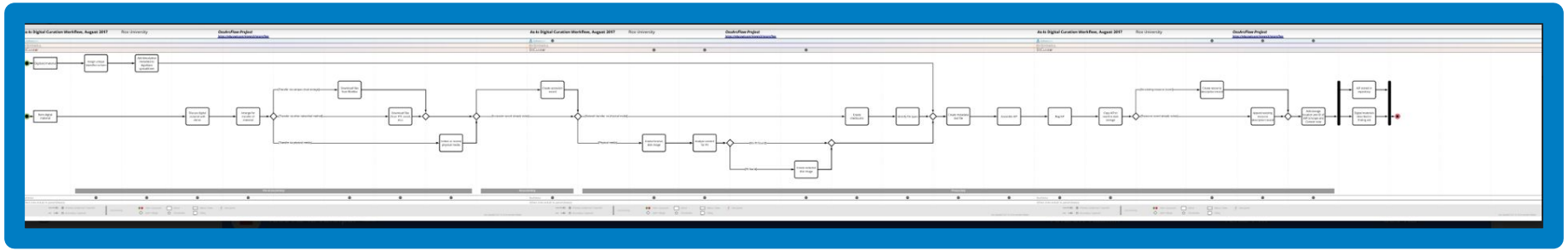
Every institution is unique, but our problems aren't!

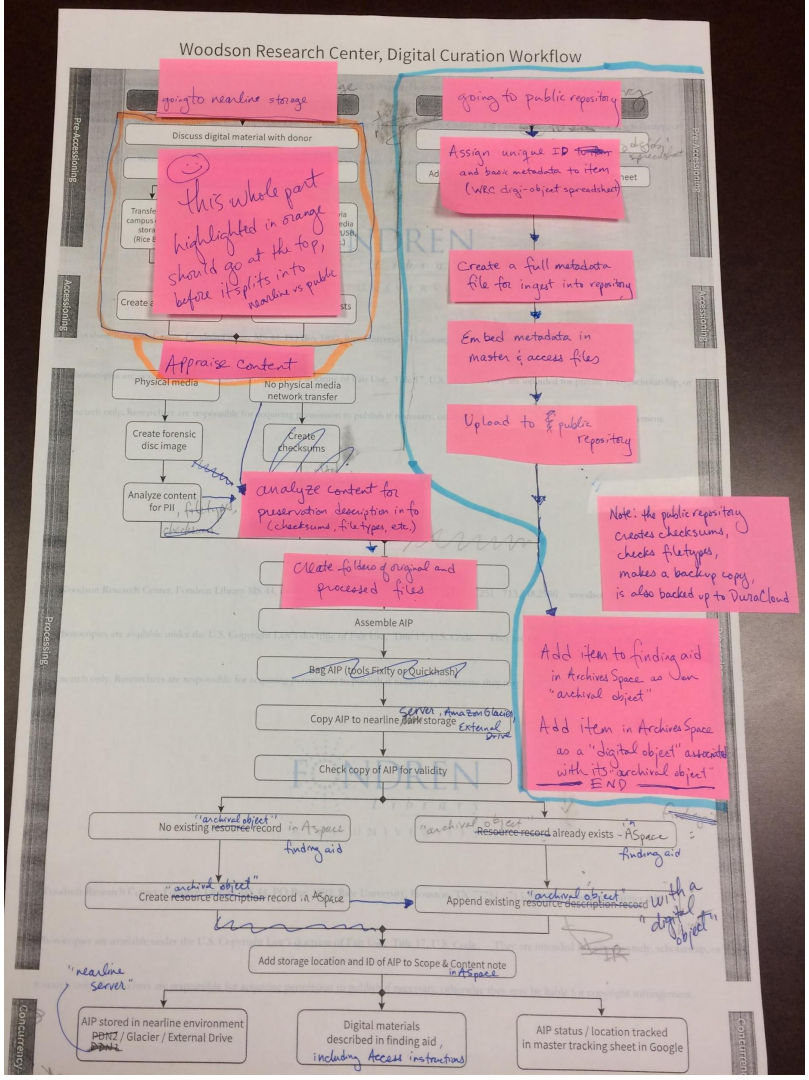
- Transition from “silos” of digital activity to a holistic approach
  - *There's no "Swiss army knife" of digital curation tools*
  - *Tools are made as stand-alones to be used in a complex network of tools*

- Too many manual processes, too much "data massaging"
  - *Output from System A doesn't work as input to System B*
  - *Rice scenario: Create disk image and reports in BitCurator > Physically transfer reports on USB drive to create AIP, describe in ArchivesSpace*

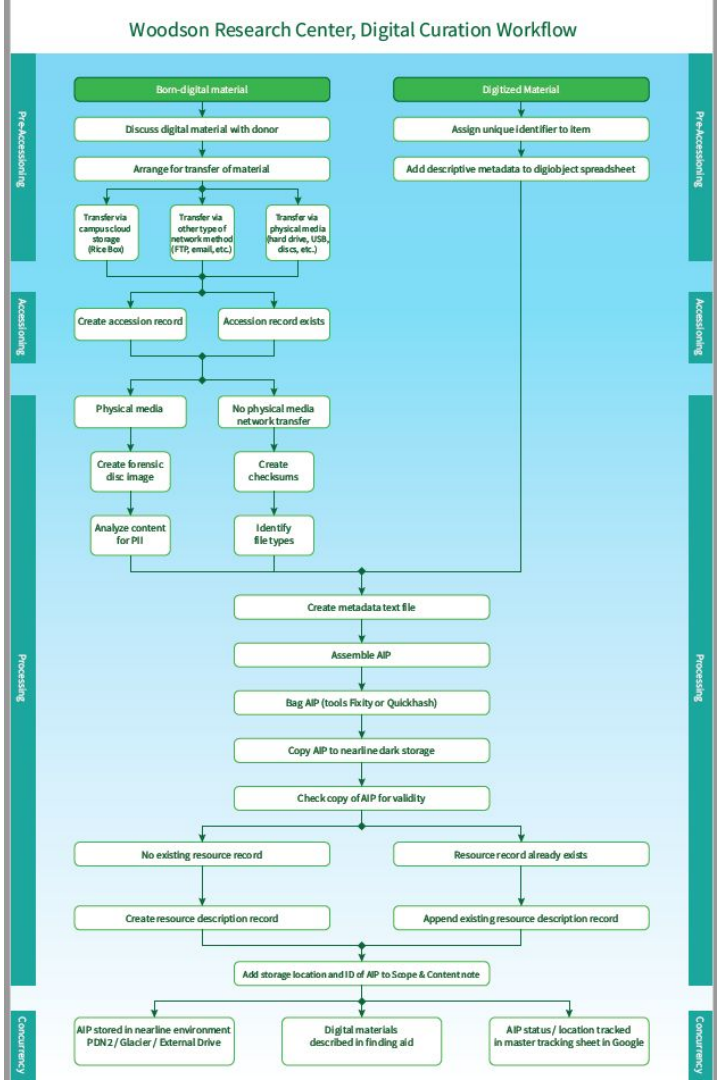
- Nobody has the 30,000-foot view
  - *It's hard to manage digital collection management*
  - *Tracking content across systems and tools*

Visual Workflows



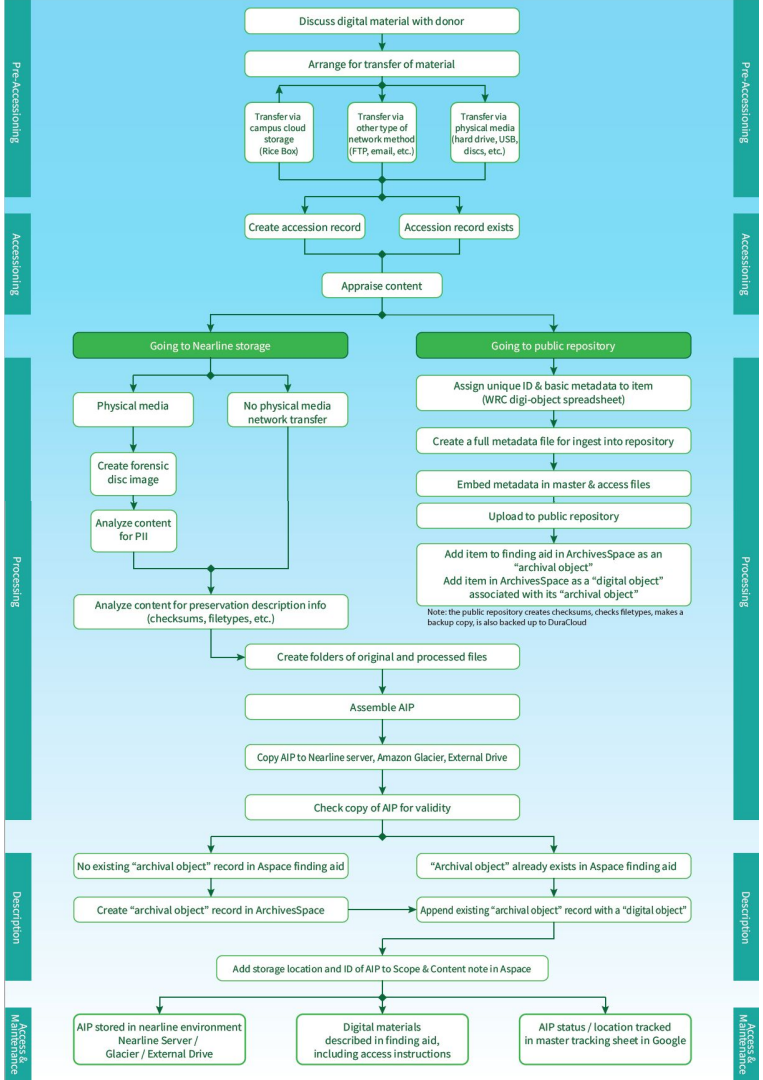


Visual Workflows – iterative process



Visual Workflows – iterative process

- Help current staff understand our workflow
- Orient new staff
- Share with other repositories



Some More Preliminary Findings …

Some cultural challenges:

- Workflows evolve quickly and digital curation has a steep learning curve
  - *Acknowledging the dynamic nature of our tools and workflows*
- Laws, policies, organizational cultures, and available resources all influence curation decisions
  - *From lawyers to Help Desk techs, everybody has a say*
- Iteration is a challenge … this is supposed to be a digital curation *cycle*, right?
  - *Curation is dynamic, but tools can encourage lock-in*

P.S. It's not just you … digital curation challenges everybody!

Project Team Activities

- **Design** training modules that will promote the use of the OSS workflow documentation and scripts

- **Create and disseminate** “Implementation Guide” to help institutions implement digital curation workflows in their own environments

Rice takeaways…from the first year

- **Prioritize** our digi pres goals, activities, update our internal workflows
- **Articulate** our digi pres needs to administration, dedicated staff time for Digital preservation
- **Support** network to bounce ideas off of, reach out to experts
- **Investigate** tools

More Rice takeaways…from the first year

- **Pockets** of deep knowledge
- **Perpetual** workflow
- **Partner** Comparisons

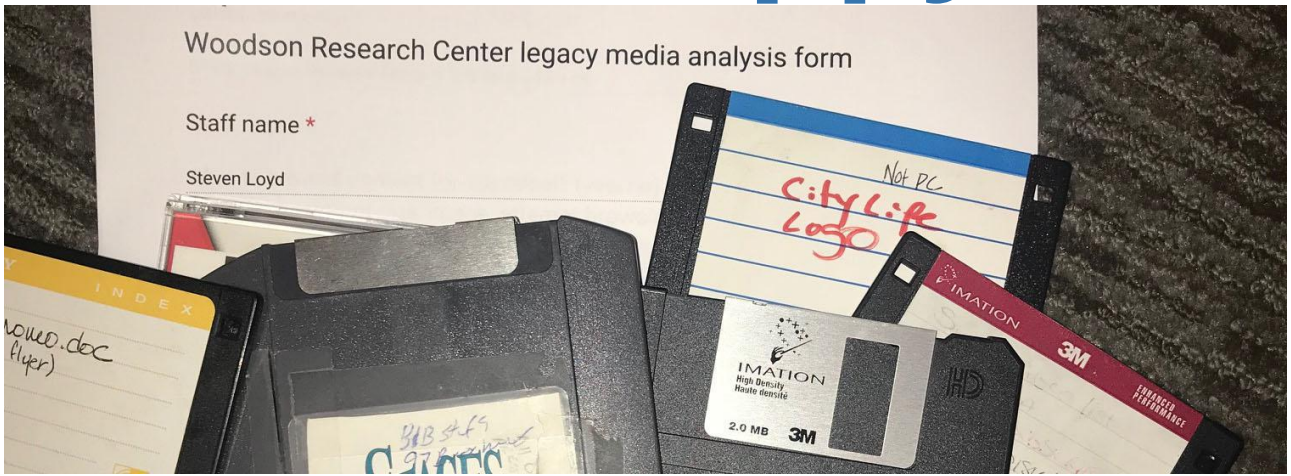
Project Webpage:

http://educopia.org/research/ossarcflow

OSSArcFlow Listserv:

https://groups.google.com/a/educopia.org/d/forum/ossarcflow_public

Fun with Floppy Discs



a case study of preserving content from legacy media formats

Amanda Focke
Woodson Research Center
Rice University

Overview & goals of our legacy media project

- Concern about our legacy media being vulnerable to physical failure – time to act!
  - Goal 1: identify, locate and track our legacy media in a spreadsheet
  - Goal 2: prioritize the capture of data off these discs
  - Goal 3: appraise, capture and normalize the data
  - Goal 4: describe and store the data

Goal 1: identify, locate & track legacy media in a spreadsheet

- Created a shared spreadsheet

Legacy eRecords inventory

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

ox 9; Box 10, Box 15

A	B	C	D	E	F	G
UA/MS	Collection Name		Location of digital objects	Media Type	Quantity	Completed-Date and Initial
2 p1, p2, p3, p4	Neal Lane papers		Box 89; Box 91, folder 14	DVD, CD	24 CDs (Box 91), 11 DVDs(Box 89)	RR12/8/2016
501	South Main Alliance records		Box 88, folder 2	Zip disks	2	RR 12/5/2016
501	South Main Alliance records		Box 88	CDs, zip disks	2 zip disks, 14 CDs	RR 12/5/2016
501	South Main Alliance records		Box 88	3.5" floppy disks	9	RR 12/5/2016
79	Walter Widrig papers on the excavation of Via Gabina Villas		Box 8, folder 21, 22; box 22, folder 11; box 5, folder 34; box 8, folder 23; box 22, folders 11-14	cd, disks [?], IBM cards	Box 8: 14 CDs; Box 23 5 IBM cards	NG 05/09/2014
79	Walter Widrig papers on the excavation of Via Gabina Villas		Box 8	3.5" floppy disks	31	NG 05/09/2014
356	Charlotte and Maximilian Collection		box 8	zip disks		NG 05/09/2014
612	Robert Avalon music collection		Box 8	Zip disks, 3.5" floppy disks	3 floppy, 3 zip	NG 05/09/2014
206	Ken Kennedy academic records		Box 63, including folders 39-43	3.5" floppy disks	57	NG 05/09/2014

Goal 2: prioritize the capture of data

- **Possible research value of the files**
- Age of the files
- How many of that media format do we have / should we buy the hardware for it?
- **Our current ability to capture the data**

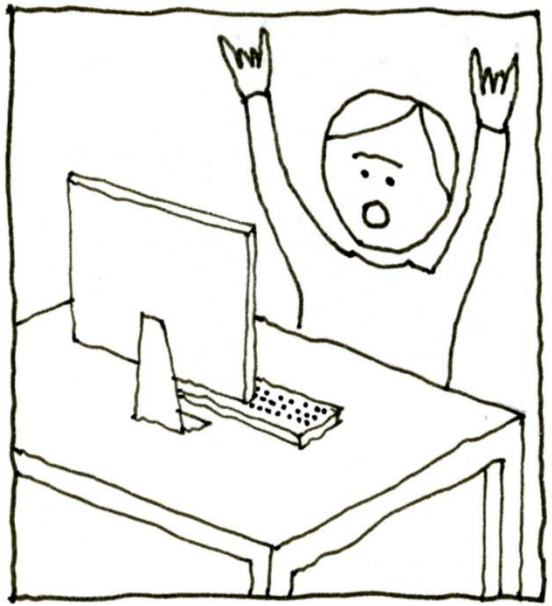
Goal 2: prioritize the capture of data

Our initial estimates of processing time needed

Format	Capture time	Appraisal time	AIP prep	AIP creation	# of items	ESTIMATED Total amount of time needed for this format
8” floppies	5 min	10 min	30 min	30 min	10	~11 hours
5.25” floppies	5 min	10 min	30 min	30 min	100	~108 hours
3.5” floppies	5 min	10 min	30 min	30 min	550	~595 hours
Zip disks	10 min	15 min	45 min	30 min	140	~233 hours
CDs	10 min	25 min	45 min	30 min	1080	1980 hours
DVDs	15 min	25 min	45 min	30 min	178	~341 hours

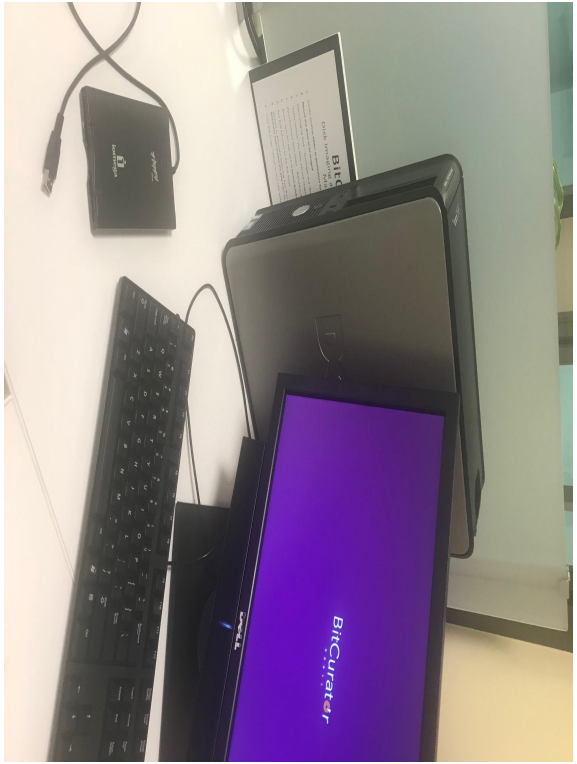
Need for a Temporary Digital Archivist!

- Staff collaborated to write a proposal for a 6 month person @ 30 hours per week
  - Overall description of project and goals
  - Statement of need
  - Examples of collections with legacy formats
  - Table of formats / time needed
  - List of workflow / steps to be performed per item
- Presented to administration – and it was funded!



Goal 3: Appraise, Capture, and Normalize the Data

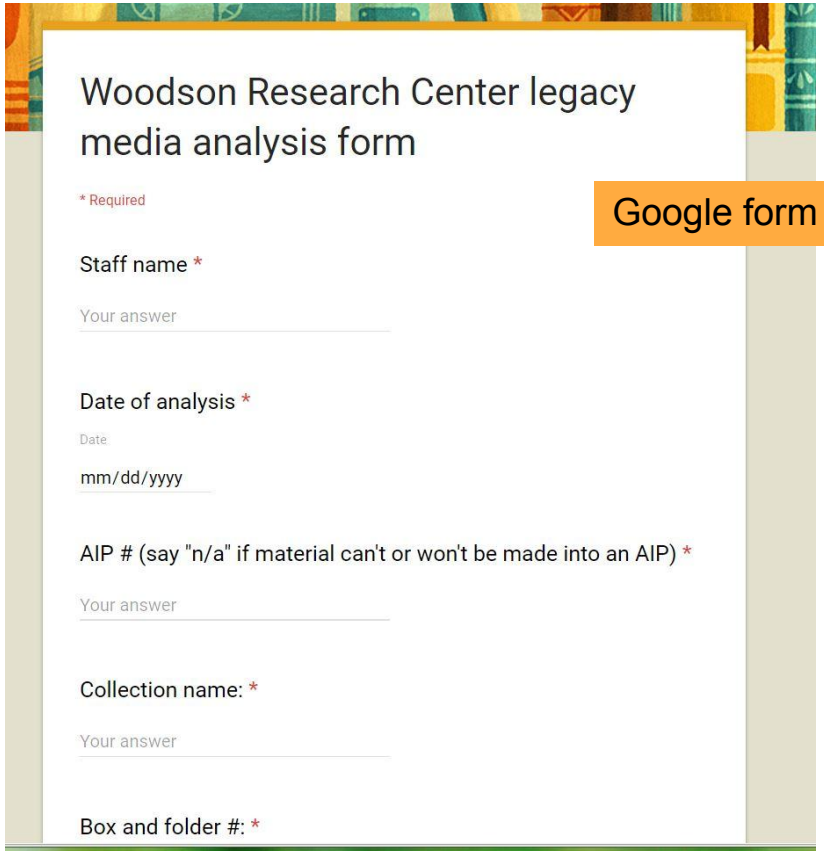
- Browse the content / appraise its value
  - Duplicate info? Has research value?
- What is BitCurator and why do we use it?
- What if you don’t have BitCurator?



https://bitcurator.net/

Goal 3: Appraise, Capture, and Normalize the Data

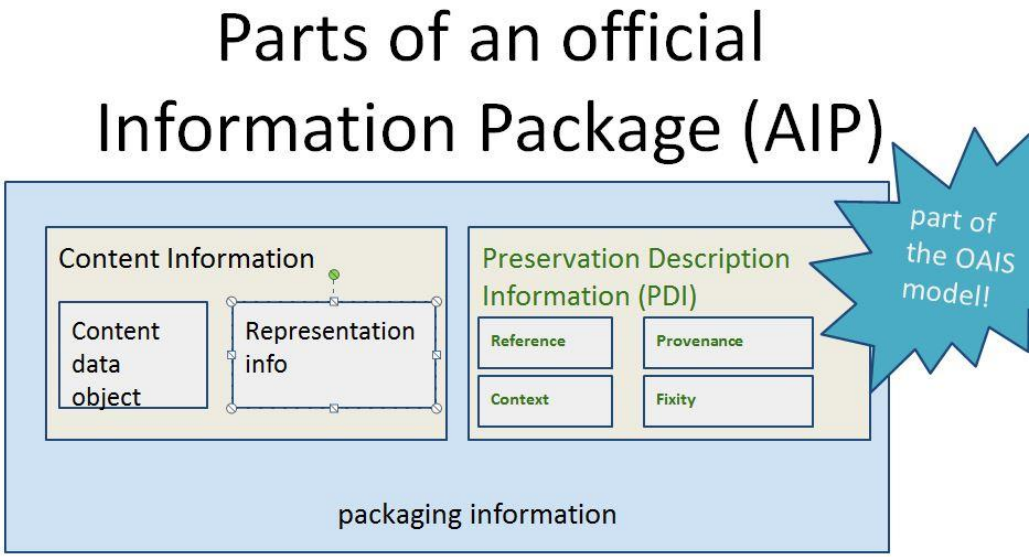
- Media analysis form to track ability to browse content, appraisal decisions, and more
- For each disc / drive – records outcomes (including disc failures) then gets printed and stored with the original media in the box.
- Responses can be seen in spreadsheet form, but don't connect to another system such as ArchivesSpace



Google form

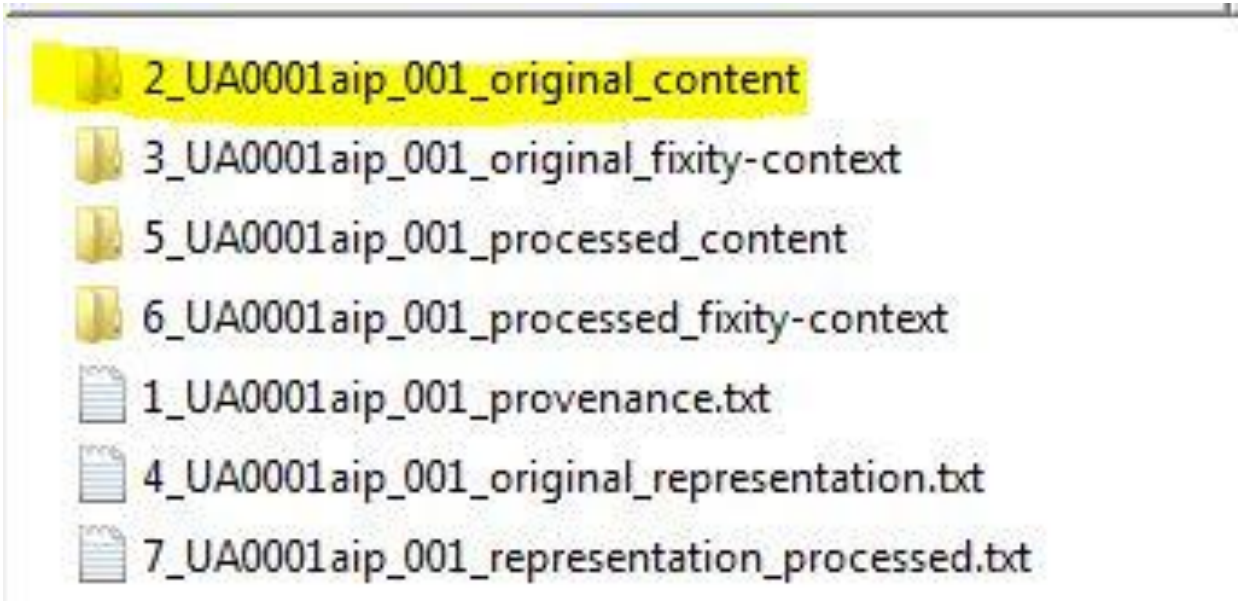
Goal 3: Appraise, Capture, and Normalize the Data

- Generally this kind of data would be stored nearline, not publicly online, so we built it into Archival Information Packages (AIPs)
- the content would be broken out into “original” and “processed” folders



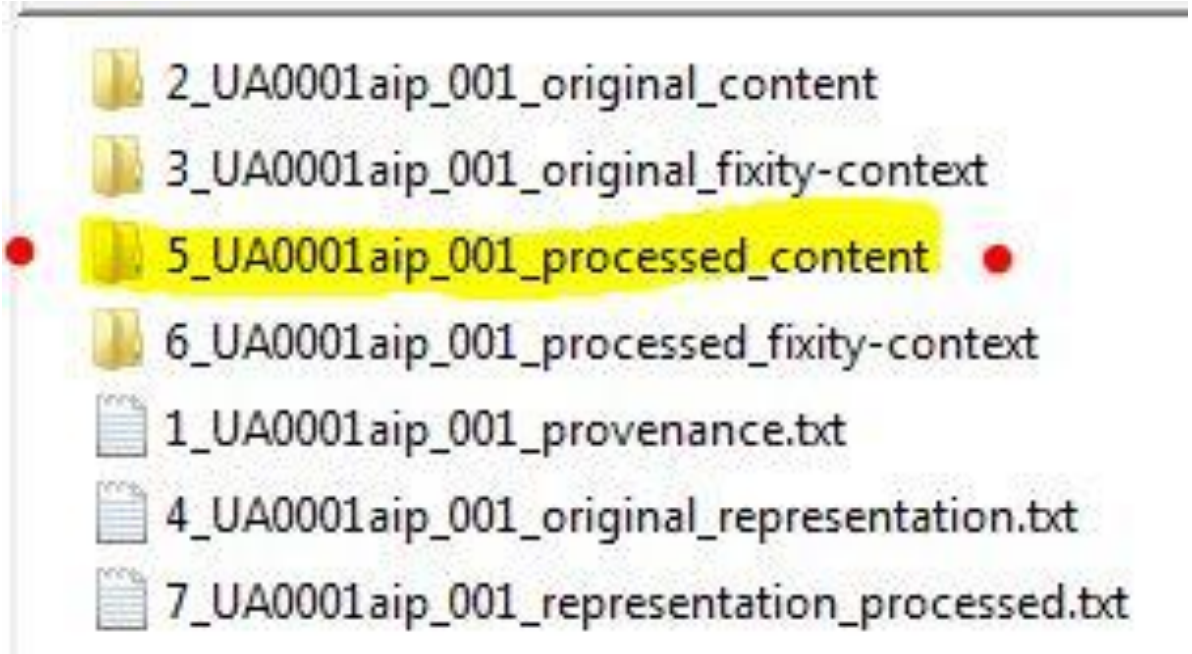
Goal 3: Appraise, Capture, and Normalize the Data

- **Capturing data** – going into the “original content” part of the AIP



Goal 3: Appraise, Capture, and Normalize the Data

- Normalized files go in the “processed” folder to facilitate access



But wait –there was more to do

- Our original plan stopped here, with creating the AIP
- Realization that we needed to build in time for more workflow steps
  - description in finding aid
  - plus storing in our 3 places and tracking those storage steps

Goal 4: Describe and Store the Data

- Described in finding aid, including clear wording on how to access the material

RICE
Fondren Library

Woodson Research Center | woodson@rice.edu | 713-348-2586

Collections | Digital Material | Subjects | Names | Record Groups

Walter Widrig villa excavation pottery report

Digital Work Identifier: MS0079aip_021

Woodson Research Center, Rice University, Houston, Texas | Walter Widrig villa excavation pottery report

Dates

2004

Summary

Reports on the pottery of the 1982 excavations of the Gabina villas. 4 PDF files converted from text files, viewable with a PDF reader.

Conditions Governing Access

This material is open for research. Please contact woodson@rice.edu or 713-348-2586 for access and use identifier # MS0079aip_021 in your request.

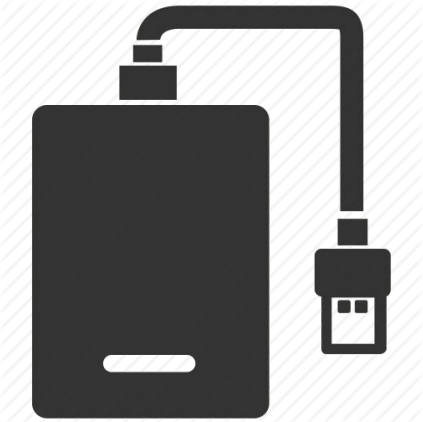
Extent

348 Kilobytes

Goal 4: Describe and Store the Data

- Stored in 3 environments, and tracking those environments

On-campus server
with 1 backup copy



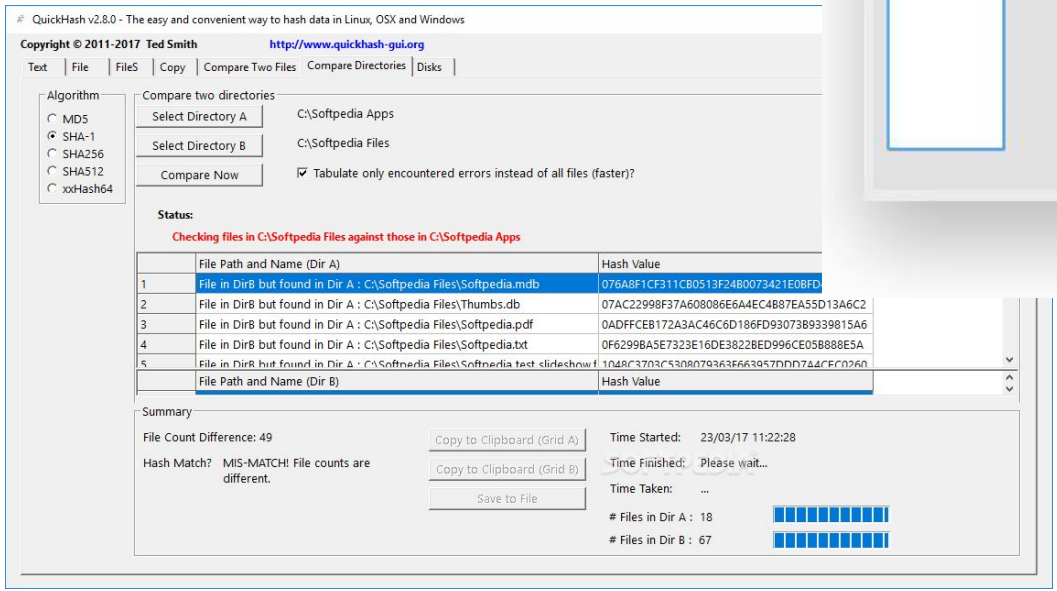
External hard drive in
the archives offices

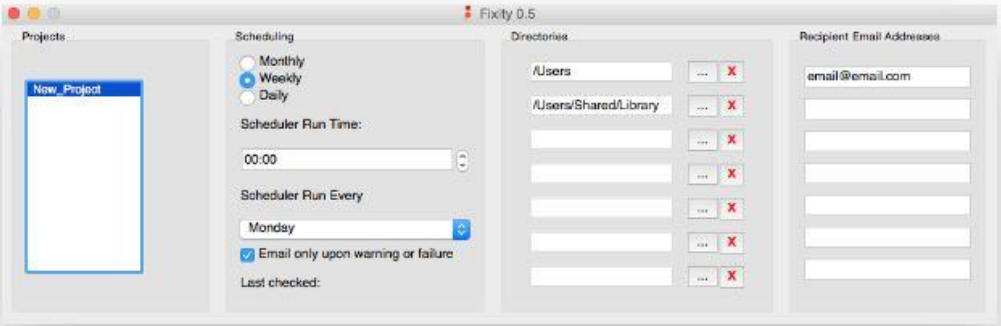
Cloud back up

Goal 4: Describe and Store the Data

- Stored in 3 environments, and tracking those environments

Quickhash

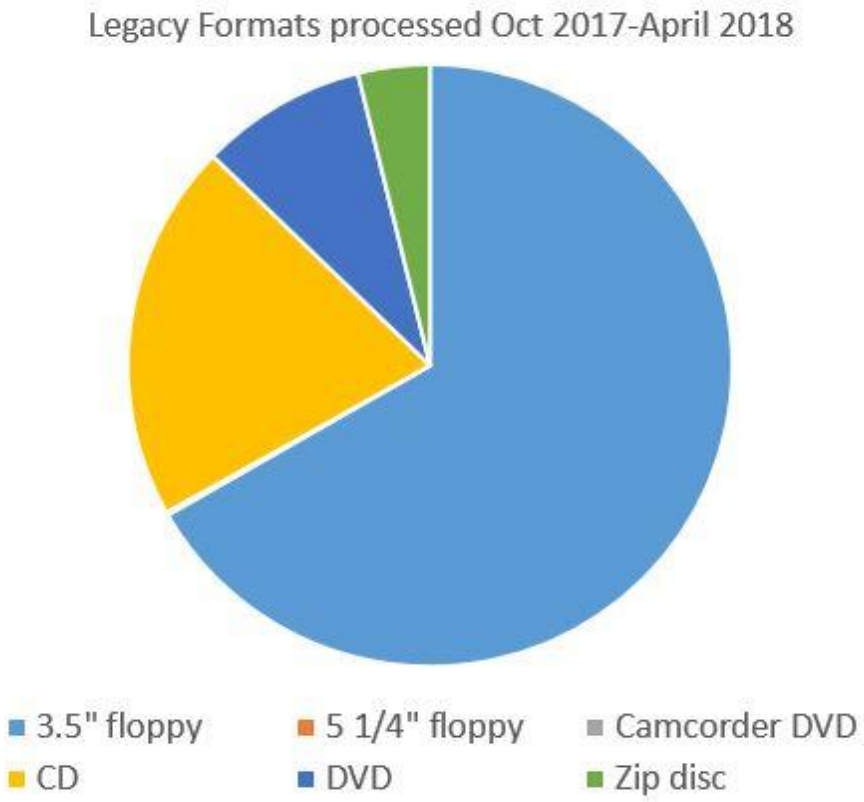




Fixity by AV Preserve

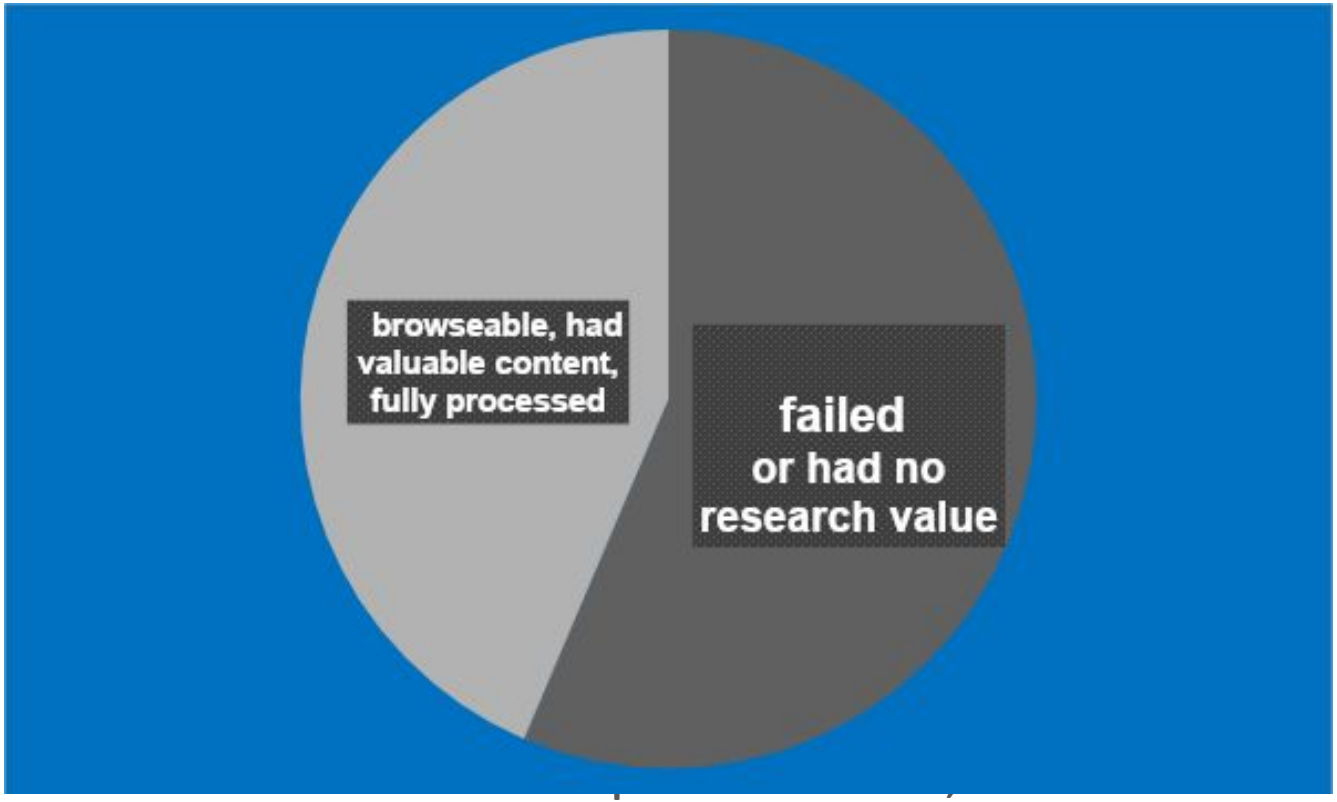
Gathering checksums and tracking them over time to look for differences / file corruption

How much did we **process** in our 6 months?



- *More than we had planned!*
  - **868 3.5” floppy disks**
  - 50 zip discs
  - 265 CDs
  - 116 DVDs
- Plus, refining our documentation & processing some non-legacy media backlog

“Processed”, which has various outcomes…



What did we learn?

Legacy media isn’t as scary as we thought it was.

Internal hard drive with IDE (40 pin) connections

LESSON: Just start somewhere, and you will build confidence.
Communicate to your donors, users, and administrators about your program.

Actual processing time was half as long as we thought – even after adding steps we hadn't originally planned on

Small disc such as a 3.5" floppy taking 30 minutes -- *not 75 minutes*

CD with 5-100MB files, 60 minutes -- *not 1 hr 50 minutes*

LESSON: When we made the original time estimates,
it was based on experience from trying to do this along with our other responsibilities –
showing it is much more efficient to have a dedicated person.

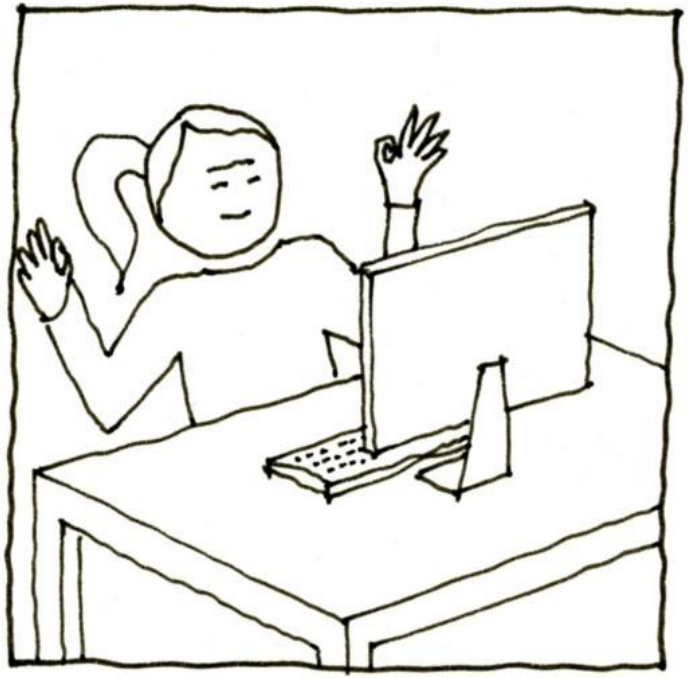
Our workflow documentation wasn’t as good as we thought it was.

- Overview flowchart
- Processing manual
- Media analysis form
- AIP tracking of description and storage
- Checksum logs for the AIPs in 3 storage areas

LESSON: Have other people review your workflows and give you feedback!

To stress?

- Digital preservation is the most challenging work we have encountered by far in archives
- Comparing our program to other repositories' can be intimidating
- This work is very time intensive, can be "invisible"
- Our IT department is for desktop support, not archival digital preservation support
- We always need more training as the materials change, the tools change, the systems change, what researchers want changes…
- There is no Finish Line



or not to stress?

- Trust that what we are doing makes a positive difference
- No one’s program is perfect or easy
- There’s no system out there that does “everything”
- There is a growing community out there to join

Thanks!

- **Rebecca Russell, CA, DAS**
- russellr@rice.edu
- 713-348-5133

- **Amanda Focke, CA, DAS**
- afocke@rice.edu
- 713-348-2124

RICE