# Practical Workflows for Large Born-Digital Collections

ODNFEST 2026

**Lara Friedman-Shedlov** | Digital Records Archivist | University of Minnesota
**Andrew McDonnell** | Digital Archivist | University of Kentucky

# Your presenters



**Lara Friedman-Shedlov**
University of Minnesota Libraries
lib.umn.edu

- Digital Records Archivist since 2020
- SAA DAS (Jan 2014)
- In Archives & Special Collections dept since 2002



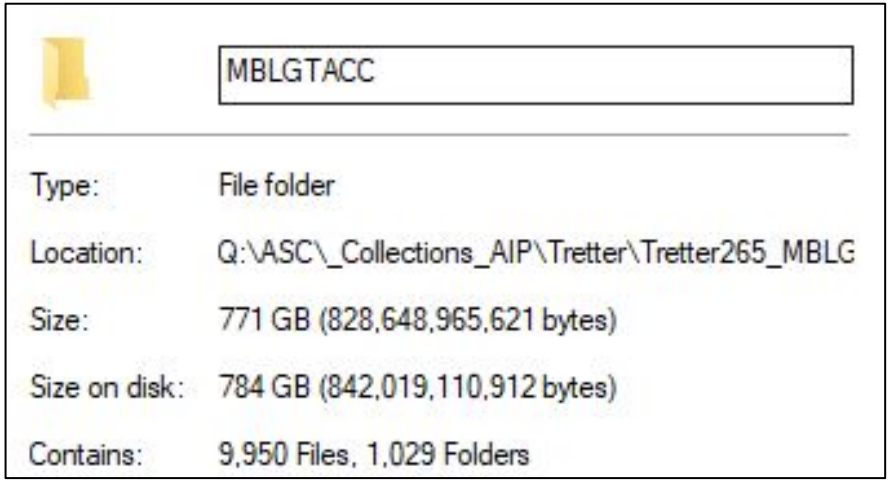
**Andrew McDonnell**
University of Kentucky Libraries
Special Collections Research Center
libraries.uky.edu

- Digital Archivist since 2023

# Born digital files are just like any other archival files, but...

- No control over file formats
  - Proprietary and obsolete software and formats
  - New types of “documents”
- Varying modes of transfer:
  - Obsolete media
  - Whole computers
  - Cloud-based files
- Malware and viruses
- Sheer volume →

- Easy to generate many versions
- Easy to duplicate
- Harder to quickly grasp content
- Sensitivity & privacy concerns



# Cloud-based born digital files are just like other born digital files, but…

- File number and size can be difficult to determine
- Metadata is often lost
- Very large amounts of files break built-in export utilities
- Proprietary file formats that can't live outside the cloud
- File ownership / permissions issues
- File system architecture/ directory structure/ shortcuts
- No real control of disposition of original/source files













# An umbrella solution for a cloudy day?




**Free**, open-source, command-line program to manage files on cloud storage

- Interfaces with 70+ cloud storage systems, including DropBox, iCloud Drive, Box, and OneDrive, and **Google Drive** (rclone.org/overview/)
- Compatible with Linux, Windows and Mac systems

# Rclone – why/why not?




**Pros**

- Copies files not owned by you (as long as shared with you)
- Preserves date modified
- Verifies with checksums
- Converts Google file formats on export (configurable)
- Can calculate total file size and there’s no data limit
- Handles intermittent connections

**Cons**

- Command line/ learning curve
- Can’t export Google forms (but can anything?)

# Setting up Rclone

```
C:\Users\ldfs\Desktop\rclone>rclone config
Current remotes:

Name                 Type
====                 ====
ASC_GDrive           drive

e) Edit existing remote
n) New remote
d) Delete remote
r) Rename remote
c) Copy remote
s) Set configuration password
q) Quit config
e/n/d/r/c/s/q>
```

```
C:\Users\ldfs\Desktop\rclone>rclone size ASC_GDrive:"Test-files"
2026/04/15 11:40:47 NOTICE: Google drive root 'Test-files': Size may be underestimated due to 3
objects with unknown size
Total objects: 112
Total size: 162.784 MiB (170691899 Byte)
Total objects with unknown size: 3 (3)
```

# Running Rclone

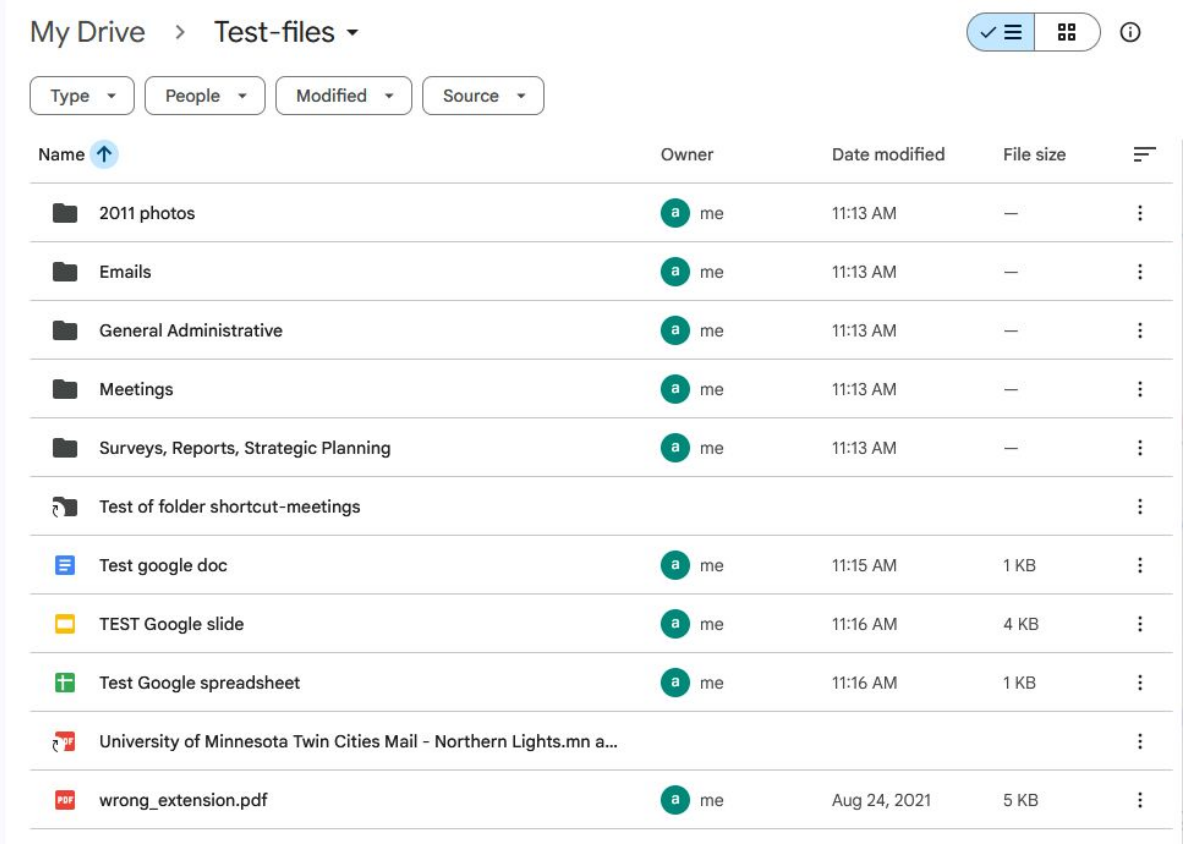
```
C:\Users\ldfs\Desktop\rclone>rclone -v -P copy ASC_GDrive:"Test-files" "D:\_Accessions_SIP\Rclone-transfers\newtest"
```

`rclone -vP copy [remote]:”[source-folder]” “[destination-path]”`

**-v** = verbose
**-P** = progress (real time transfer stats)

And now we wait....

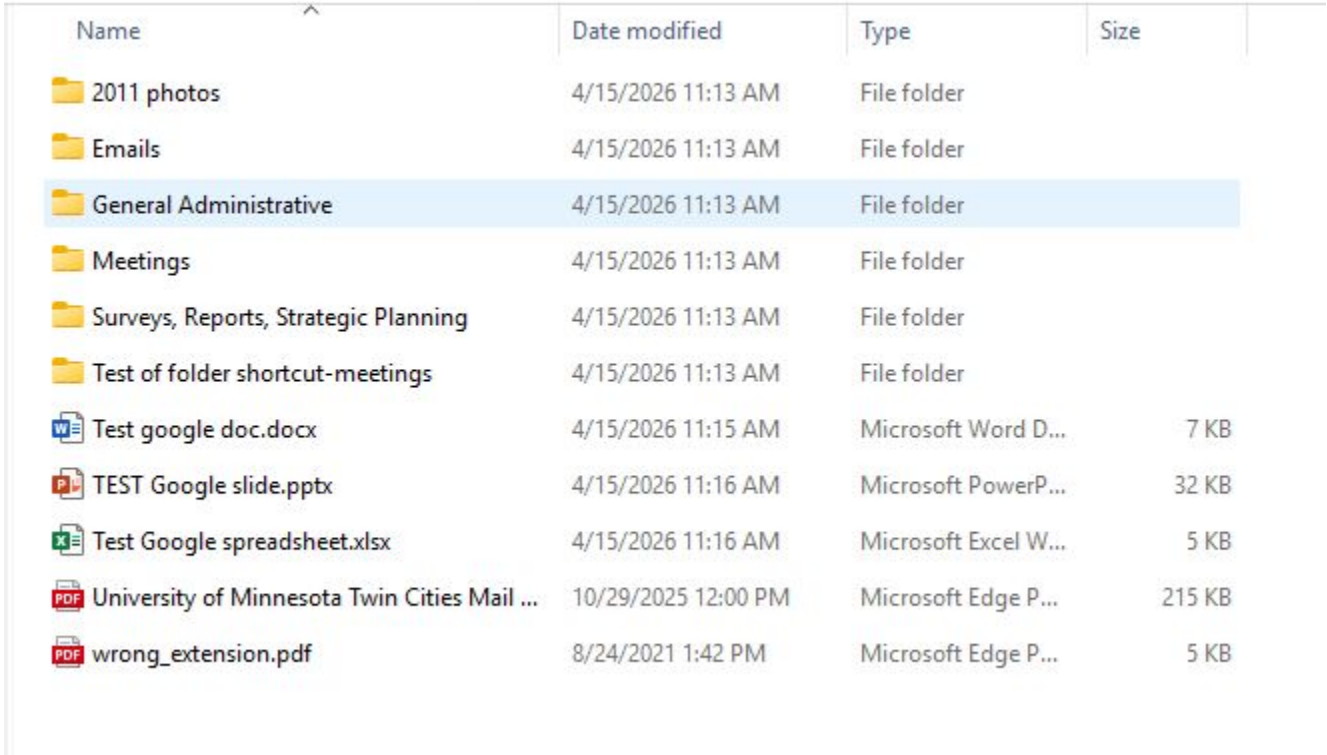
# Google Drive files to be transferred




**Challenges/concerns**

- Folder structure maintained?
- Google file formats converted?
- Shortcutted file?
- Shortcutted folder?

# Rclone copy from Google Drive



| Name | Date modified | Type | Size |
|---|---|---|---|
| 2011 photos | 4/15/2026 11:13 AM | File folder | |
| Emails | 4/15/2026 11:13 AM | File folder | |
| General Administrative | 4/15/2026 11:13 AM | File folder | |
| Meetings | 4/15/2026 11:13 AM | File folder | |
| Surveys, Reports, Strategic Planning | 4/15/2026 11:13 AM | File folder | |
| Test of folder shortcut-meetings | 4/15/2026 11:13 AM | File folder | |
| Test google doc.docx | 4/15/2026 11:15 AM | Microsoft Word D... | 7 KB |
| TEST Google slide.pptx | 4/15/2026 11:16 AM | Microsoft PowerP... | 32 KB |
| Test Google spreadsheet.xlsx | 4/15/2026 11:16 AM | Microsoft Excel W... | 5 KB |
| University of Minnesota Twin Cities Mail ... | 10/29/2025 12:00 PM | Microsoft Edge P... | 215 KB |
| wrong_extension.pdf | 8/24/2021 1:42 PM | Microsoft Edge P... | 5 KB |

# Running Rclone – advanced

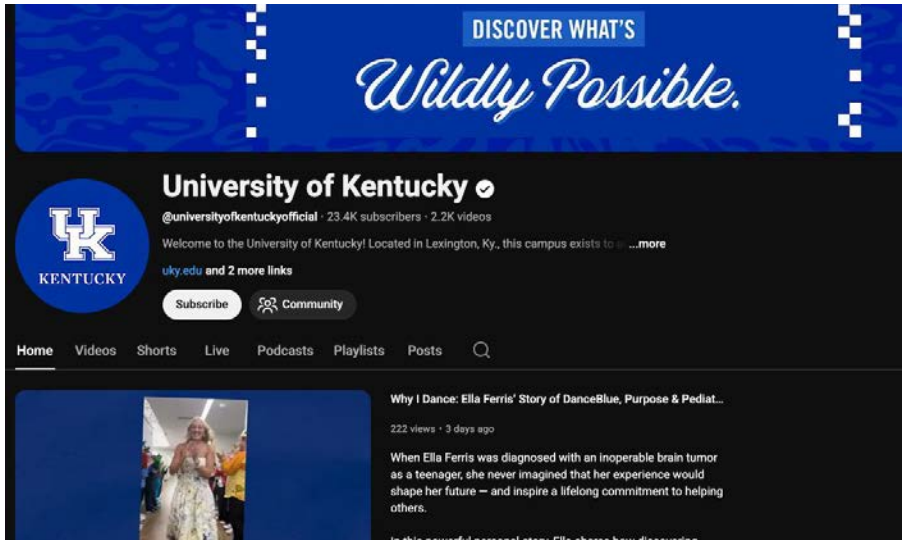
```
rclone copy -cMP --log-file [logfilename].txt --log-level DEBUG
[remote]:“[source-folder]” “[destination-path]”
```

**-c** = generate checksums
**-M** = preserve metadata (if possible – works on some remotes)
**-P** = progress (real time transfer stats)

-- log-file generates a logfile
-- log-level determines the level of detail. INFO= verbose, DEBUG = very verbose

# YouTube, Google Drive, and TakeOut

University of Kentucky Archives
Preserving the institutional YouTube account




www.youtube.com/universityofky

The Challenge:

- 16 years of videos
- Many videos no longer on any other drive, server, or known location
- Valuable metadata also held only within the YouTube channel

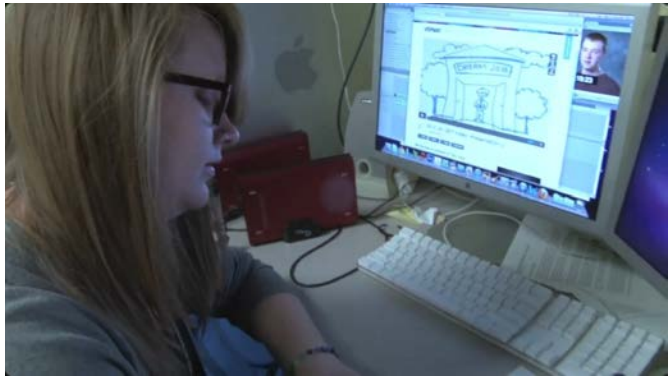
# YouTube impermanence




All is impermanence.

16 years of videos means, especially for a brand and recruitment focused account, there is significant outdated information or dated content that PR would rather not have front-facing.

Videos are being deleted.



Machines change.




Will we even eat ice cream in the future?

# An existing relationship pays dividends: UK Digital Communications Archiving Working Group



- Established three years ago to improve University Archives' web archiving program
- Expanded into photography and, for this project, the video team
- Without this: probably no means of preserving everything we'd like to

# Downloading videos one at a time is pretty easy: YT-DLP

You can download YouTube videos one at a time ***without*** credentials:
Public Domain Tool: YT-DLP: https://github.com/yt-dlp




The Good!

- Updated frequently, works on more than YouTube
- Sample command: yt-dlp -P “download/location” “Video-URL”
- Metadata downloadable, various formats

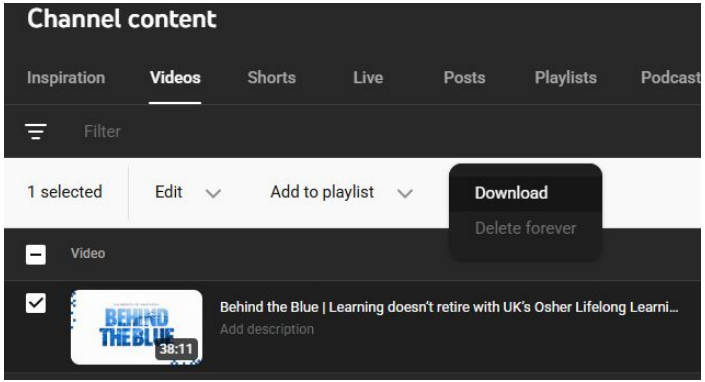
The Less Good!

- Caveat emptor:  pirate vibes, proceed with caution
- Redditors report: overuse can get your IP address blocked by YouTube

# Downloading videos one at a time is easy: within YouTube

You can download YouTube videos one at a time ***with*** credentials

- With back-end access you can easily download YouTube videos one at a time



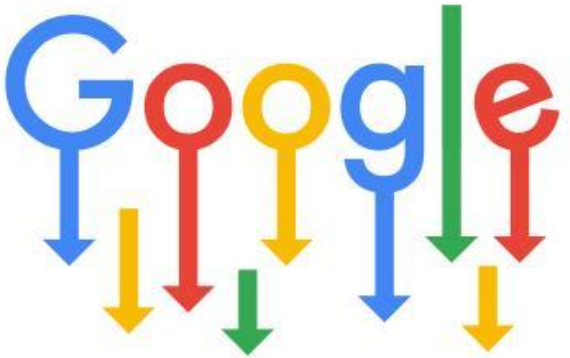

The Good:

- Easy, straightforward
- Channel owners can share access without sharing credentials

The Bad:

- No automatic accompanying metadata
- Reformats the video, so if you require original upload quality, not an option.

# Downloading all videos all at once: potentially challenging




**Our Solution: Google TakeOut**

- Only possible with owner's Google account login credentials
- Shared YouTube editor permissions: insufficient. Requires owner credentials or collaboration with the owner to directly do the work.

# Google Takeout challenges: time



Hiccups along the way to downloading

- 4,619 videos, 5.143 TB

**Challenge 1: Time**

Google Takeout downloadable in packages maximum 50 GB

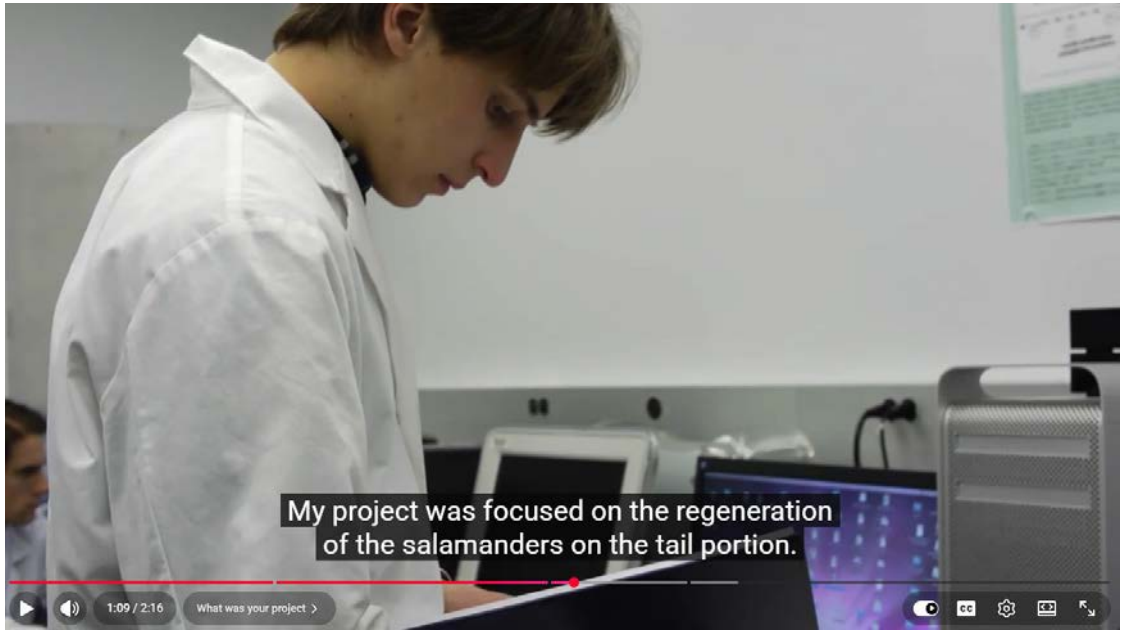
- In our case 93 packages, and ~20 individual files larger than 50 GB
- My browser could only download 3 at a time without crashing
- Each file: ~25-30 minutes to download
- Problem: Takeout downloads expire after 7 days

Lesson learned: If you run out of time and the downloads expire, you can re-generate a Google Takeout request. However files are packaged/zipped in a different order, so you must do it all over again to ensure complete collection.

# Google Takeout challenges: more time

**Challenge 2: More Time**
Even extracting that many files took a full week.

This process exports *everything*. Public, private, unlisted videos inclusive.



# Google Takeout challenges: files

**Challenge 3: Corrupt files and problematic formats**

- Around 150 of the 4,619 unzipped files were corrupt and needed to be re-downloaded individually
- Some of the older livestreamed events exported as .flv videos. Rather than convert, re-downloaded as mp4 files individually



# Google Takeout challenges: metadata

**Challenge 4: Google TakeOut file naming/metadata correlation**

- Unique IDs for videos do not correlate to the downloaded files in a meaningful way.
- Exact file names for videos: not in the csv metadata.
  - Abridged versions of the video titles as they appeared on YouTube.
  - Capped at 47 characters, any duplicates appended with (1) (2) … etc.
  - Forbidden file-name characters auto-replaced, unnoted in metadata

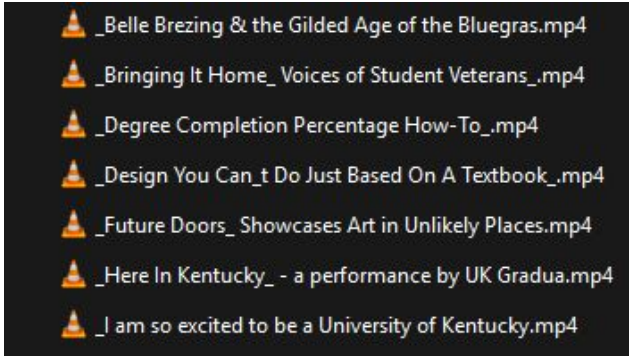
| Video Title (Original) | Video ID | Approx Du | Duration (h | Video Category | Video Description (Original) | C |
|---|---|---|---|---|---|---|
| University of Kentucky 2011 Evening Commencement Pre-show | M4MzBCC5E54 | 597000 | 0:09:57 | Education | This is a copy of the University of Kentucky 2011 | U |
| University of Kentucky 2011 Evening Commencement Processional | gqrdqLR4Xoc | 1067000 | 0:17:47 | Education | This is a copy of the broadcast of the processic | U |
| University of Kentucky 2011 Evening Commencement Speech by Pres. Todd | g0eYuuaDUdw | 798000 | 0:13:18 | Education | This is a copy of part of the broadcast of the Un | U |
| University of Kentucky 2011 Graduate & Professional Student Commencement Diploma Prese | UzkK0sopxus | 5239000 | 1:27:19 | Education | This is a copy of the broadcast of the University | U |
| University of Kentucky 2011 Graduate & Professional Student Commencement Final Remarks | gFJAeQGjne8 | 1617000 | 0:26:57 | Education | This is a copy of the broadcast of the University | U |
| University of Kentucky 2011 Graduate & Professional Student Commencement Pre-show | a1neTt7L2Rk | 703000 | 0:11:43 | Education | This is a copy of the University of Kentucky 2011 | U |
| University of Kentucky 2011 Graduate & Professional Student Commencement Processional | S7_Tkfp3ncA | 1366000 | 0:22:46 | Education | This is a copy of the broadcast of the University | U |
| University of Kentucky 2011 Graduate & Professional Student Commencement Speech | ytbNB2AZOhc | 774000 | 0:12:54 | Education | This is a copy of the broadcast of the University | U |
| University of Kentucky Alma Mater | VvCIUhRHXSU | 42000 | 0:00:42 | Education | The University of Kentucky Alma Mater has play | U |

# Google Takeout challenges: metadata

**Challenge 4: Google TakeOut file naming/metadata correlation**

- Unique IDs for videos do not correlate to the downloaded files in a meaningful way.
- Exact file names for videos: not in the csv metadata.
  - Abridged versions of the video titles as they appeared on YouTube.
  - Capped at 47 characters, any duplicates appended with (1) (2) … etc.
  - Forbidden file-name characters auto-replaced, unnoted in metadata

| Video Title (Original) | Video ID |
|---|---|
| "A College Degree Will Always Open More Doors" | HiBU7PFs2zM |
| "Belle Brezing & the Gilded Age of the Bluegrass" | _XSwtbU4IFU |
| "Bringing It Home: Voices of Student Veterans" | uKryOEs9E7A |
| "Degree Completion Percentage How-To" | Ab7ZjhO7eN8 |
| "Future Doors" Showcases Art in Unlikely Places | tHUaVRYXfx4 |



# Google Takeout takeaways

**Limited use cases**

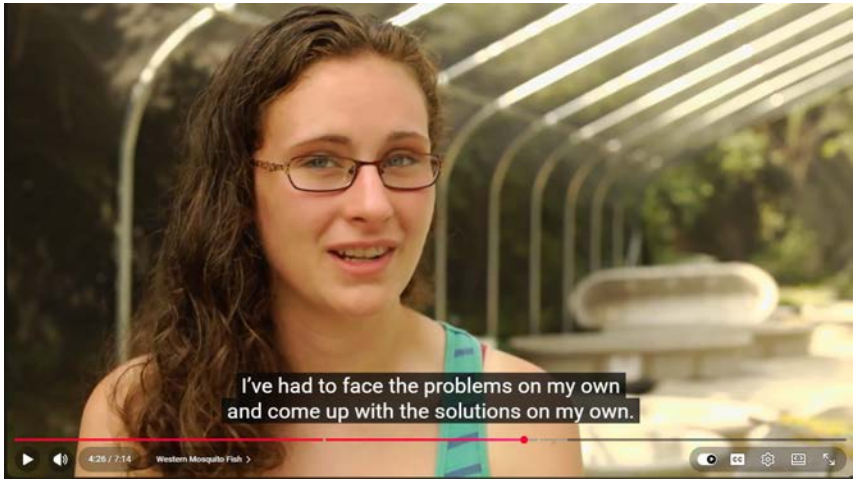
- Requires rare, unfettered access and/or a donor willing to work in concert.

**Hiccups are baked in**

- Knowing is half the battle.
- But then there is still half a battle.

**But for the right project, it's worth it**

- For our work, this would have been otherwise infeasible.



# Optical disks



- Time consuming to transfer files from this carrier
- Can be structured many ways
  - Data
    - Multi-session?
  - Audio
  - DVD
- More info
  - Color books (www.cdfs.com/cdfs-color-books.html)
  - A. Duryee, “An introduction to optical media preservation”, Code4lib Journal, Issue 24 2024-04-16
  - Developing a Robust Migration Workflow for Preserving and Curating Hand-held Media (British Library)

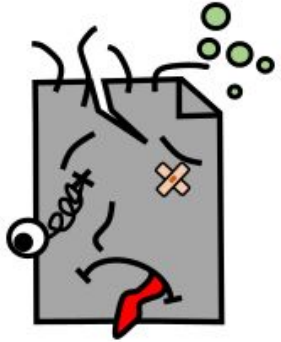
# Enter Nimbie

Acronova Nimbie (NB21-BR) is a disk autoloader / “disk robot”



https://store.acronova.com/nimbie-usb-plus-bd-cd-dvd-autoloader-nb21.html

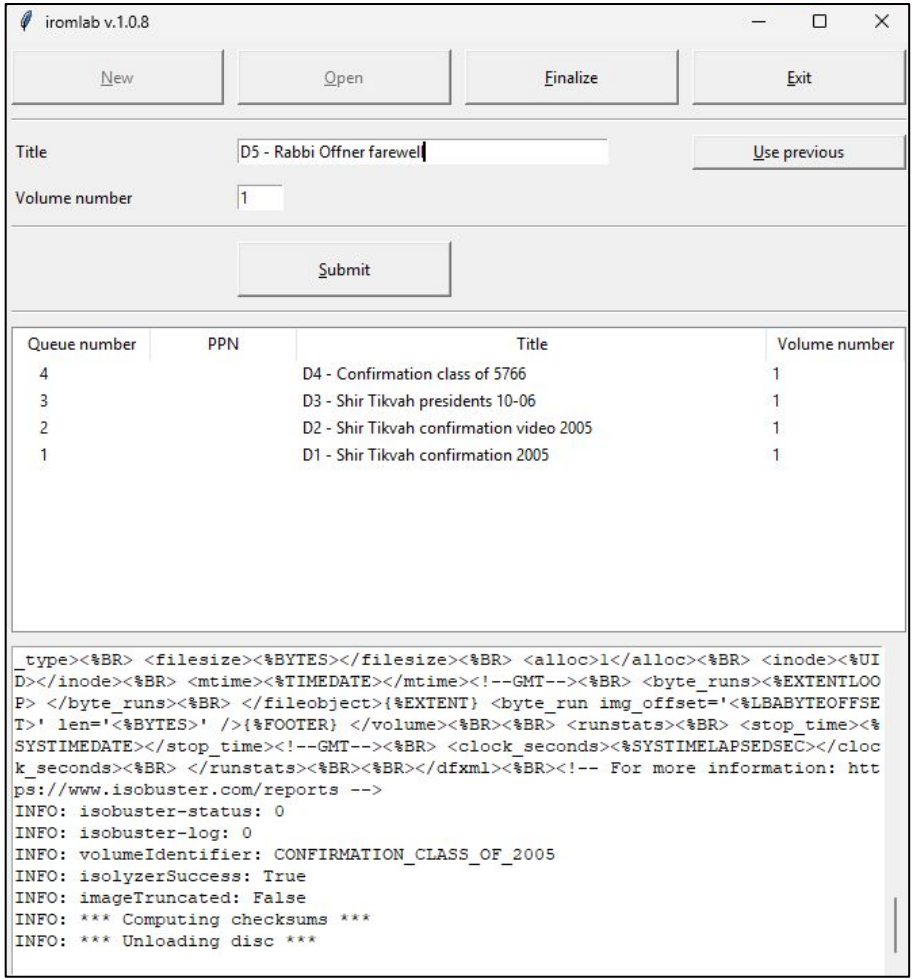
# Iromlab



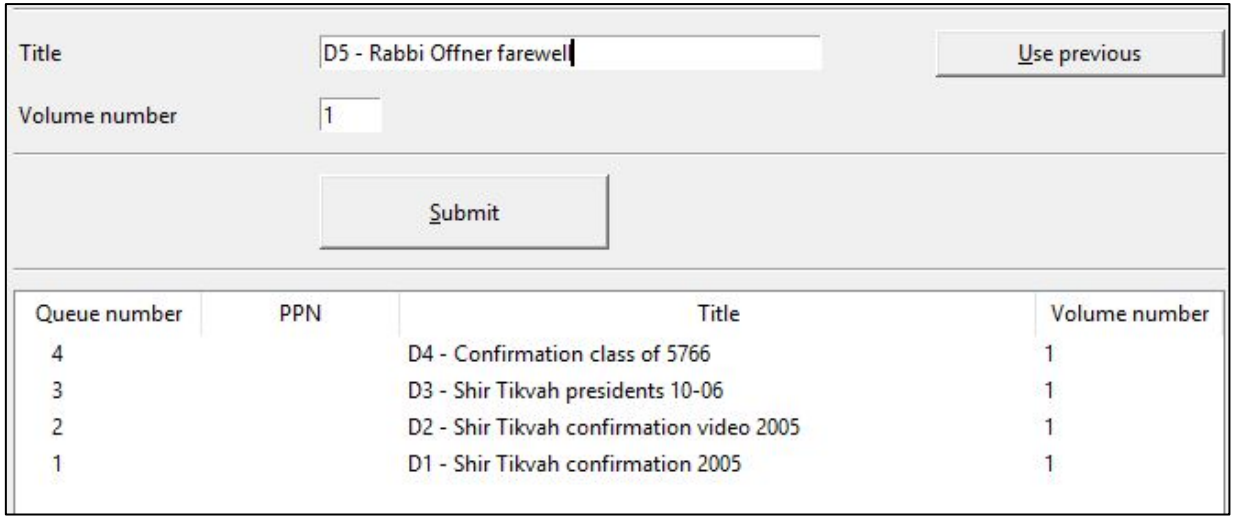
- “Image and Rip Optical Media Like A Boss”
- Developed by Johan van der Knijff for National Library of the Netherlands
- “Assessing High-volume Transfers from Optical Media at NYPL,” 2021 (https://journal.code4lib.org/articles/15908)
- Free, open source, Python-based
  - BUT employs IsoBuster + dBpoweramp
  - Windows (may be adaptable)
  - GUI interface
  - Reports/logs
  - Supports multiple optical disk carrier types (CD-ROM, DVD video, Audio CD, mixed modes, etc.)

# Iromlab interface

- Enter title of each disk
- Drop disk into well
- Submit
- When all disks in batch entered, finalize
- Disks begin processing immediately



# Running Iromlab

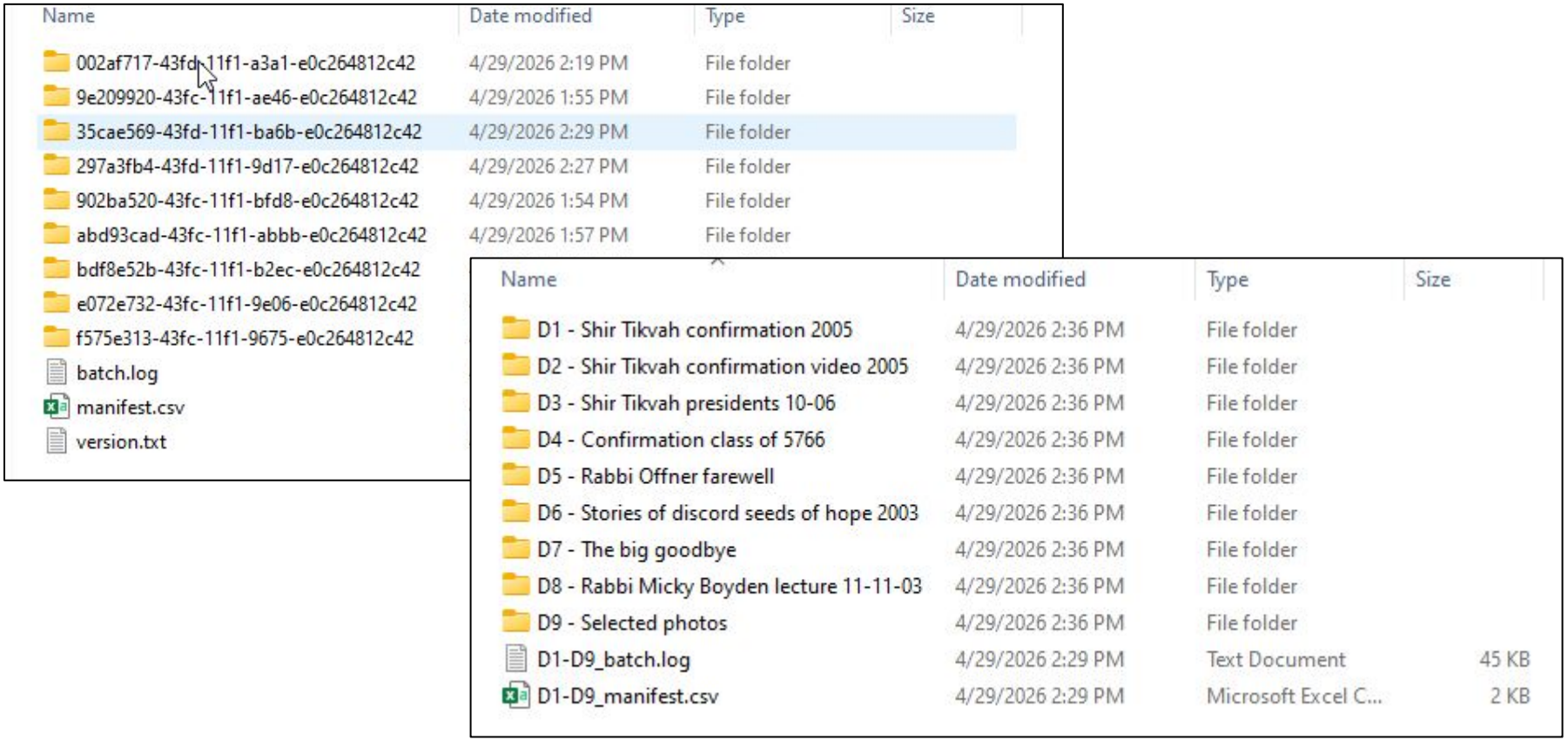



```
_type><%BR> <filesize><%BYTES></filesize><%BR> <alloc>1</alloc><%BR> <inode><%UI
D></inode><%BR> <mtime><%TIMEDATE></mtime><!--GMT--><%BR> <byte_runs><%EXTENTLOO
P> </byte_runs><%BR> </fileobject>{%EXTENT} <byte_run img_offset='<%LBABYTEOFFSE
T>' len='<%BYTES>' />{%FOOTER} </volume><%BR><%BR> <runstats><%BR> <stop_time><%
SYSTIMEDATE></stop_time><!--GMT--><%BR> <clock_seconds><%SYSTIMELAPSEDSEC></cloc
k_seconds><%BR> </runstats><%BR><%BR></dfxml><%BR><!-- For more information: htt
ps://www.isobuster.com/reports -->
INFO: isobuster-status: 0
INFO: isobuster-log: 0
INFO: volumeIdentifier: CONFIRMATION_CLASS_OF_2005
INFO: isolyzerSuccess: True
INFO: imageTruncated: False
INFO: *** Computing checksums ***
INFO: *** Unloading disc ***
```

# Iromlab file manifest

| jobID | PPN | volu | title | volumeID | success | containsA | containsD | cdExtra | mixedMo | cdInteract | multiSession |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 902ba520-43fc-11f1-bfd8-e0 | | 1 | D1 - Shir Tikvah confirmation 2005 | CONFIRM | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| 9e209920-43fc-11f1-ae46-e0 | | 1 | D2 - Shir Tikvah confirmation video 2 | CONFIRM | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| abd93cad-43fc-11f1-abbb-e | | 1 | D3 - Shir Tikvah presidents 10-06 | SHIR_TIKV | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| bdf8e52b-43fc-11f1-b2ec-e0 | | 1 | D4 - Confirmation class of 5766 | PLAY_MO | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| e072e732-43fc-11f1-9e06-e0 | | 1 | D5 - Rabbi Offner farewell | THE_BIG_ | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| f575e313-43fc-11f1-9675-e0 | | 1 | D6 - Stories of discord seeds of hope | VOLUME_ | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| 002af717-43fd-11f1-a3a1-e0 | | 1 | D7 - The big goodbye | THE_BIG_ | TRUE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| 297a3fb4-43fd-11f1-9d17-e0 | | 1 | D8 - Rabbi Micky Boyden lecture 11-11-03 | | TRUE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE |
| 35cae569-43fd-11f1-ba6b-e0 | | 1 | D9 - Selected photos | ST-Freddi | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |

# File clean-up using Python script



# Robot vs human

| Transfer method | Average transfer time per disk |
|---|---|
| Manually (disk by disk) | 6 min |
| Autoloader - machine time | 3.6 min |
| Autoloader - active/person time | 1.2 min (DVD)<br>3.2 min (mixed) |

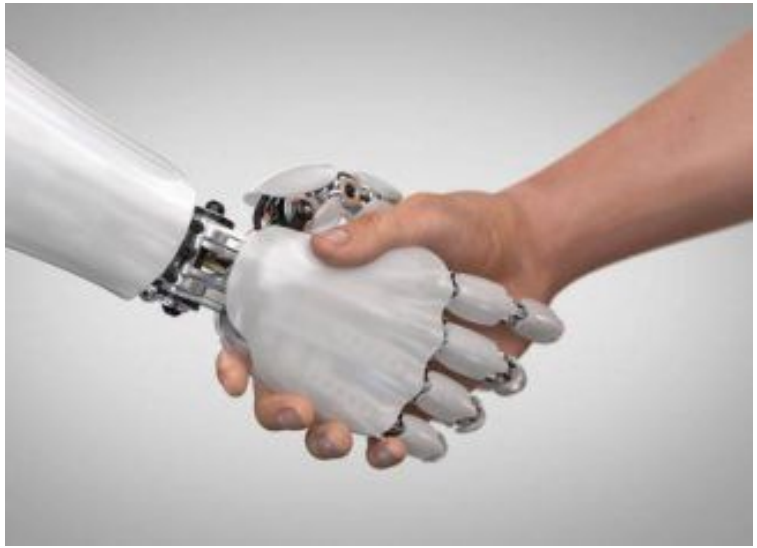


Image: CC-BY 3.0
https://www.inteldig.com/seccion/ciencia/robotica/

# More strategies for large optical disk collections

- Sub-maximal batches are better, at least at first
- Scripts!
  - Renaming and deleting unwanted log files
  - Extracting files from .iso
- Pre-inventory template (number, photograph, list)
- Cut and paste disks titles from inventory into Iromlab
  - Faster
  - Less error-prone than manually typing
- Feed spreadsheet into Iromlab
- **Automated disk metadata capture (with OCR)**

# Automation to harvest disk metadata

**The Bottleneck:**

Transcribing disk metadata from labels and cases

- Limited student budget
- How to go faster and limit drudgery?



# Automation to glean disk metadata

**A solution: Microsoft Power Automate**

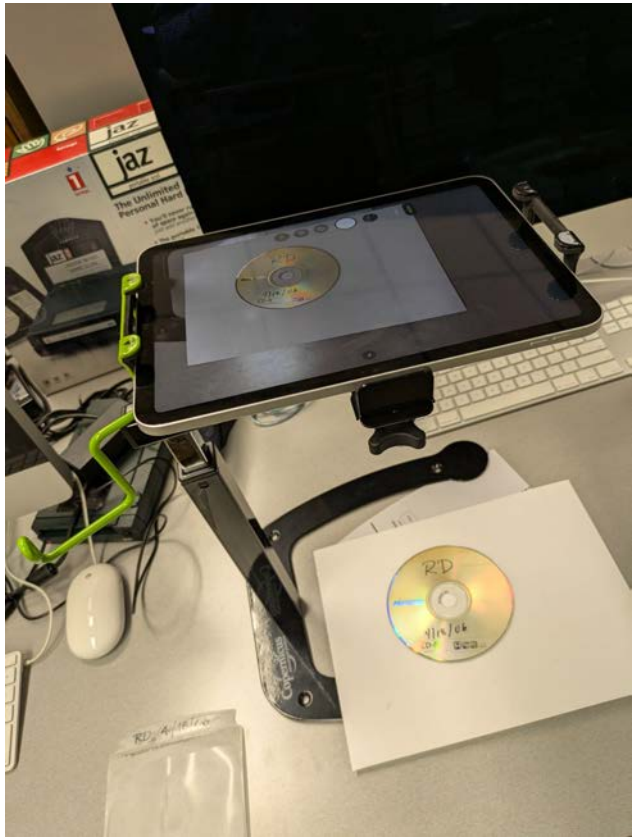



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Unique Identifier | PM format | Transcription from case and/or PM | Virus Scan Results | # of extracted files | Total size | Migration date | Server Location | Tech |
| 3 | 32 | CD | MICHAEL JOHNATHON<br>WOODSONES<br>OLD-TIME RADIO HOUR<br>1. The Show / 2. Celebrity Promo / 3. Ric<br>Recorded by<br>PLANET III<br>POET.ALAN<br>Lexington, KY<br>606 225-0333<br>RECORDS USA<br>606-255-5700<br>SHOW<br>32<br>AFFILIATE COPY<br>NOT FOR SALE<br>E 1999 POETMAN RECORDS USA<br>All rights reserved. Made in the USA | Scan - 2026-03-27 09_03_30.pdf | | | 3/27/2026 | | McDonnell |

# Disk label photo setup

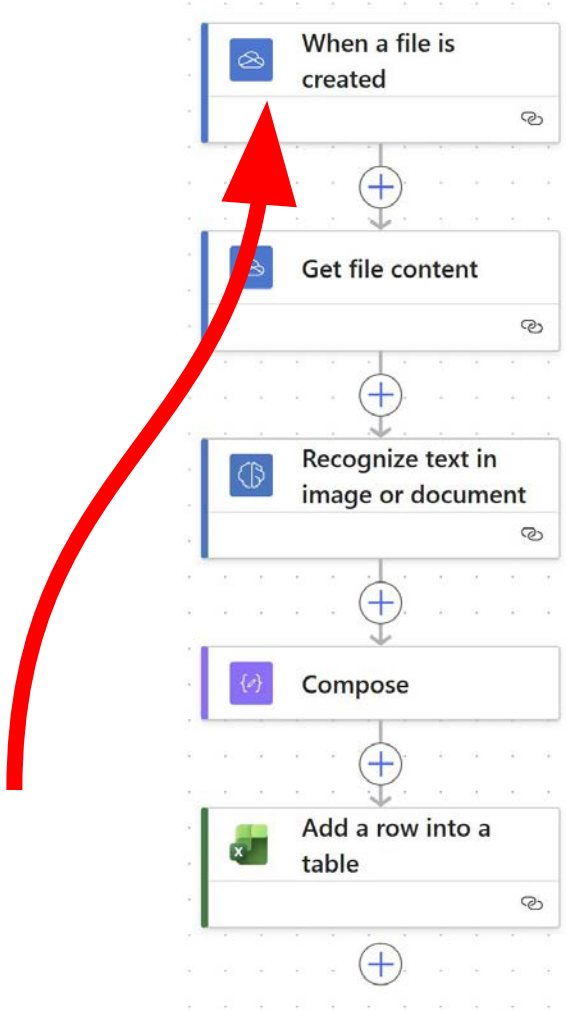
**A simple setup:**

- Initial experiments using OneDrive app on my cell phone were successful.
- Transitioned to $60 overhead stand with an iPad



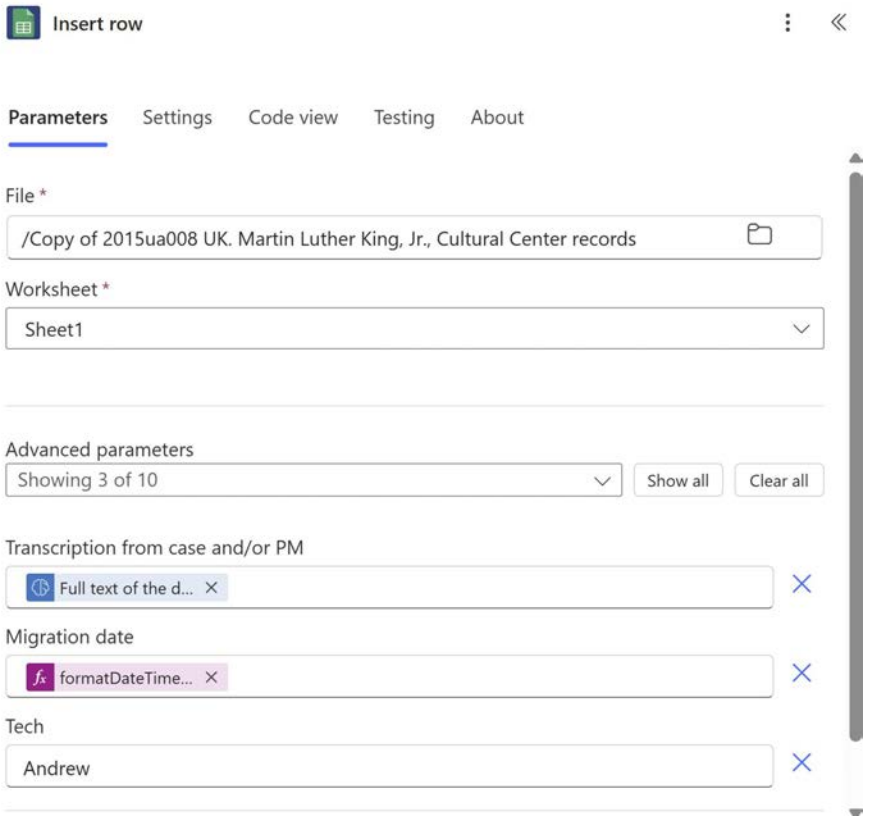
# Power Automate “flow”

1. Create a project folder for disk photos
2. Set photo app on phone/tablet to save to that folder.
3. Within OneDrive app, you can add photos to the folder within the app.
4. When a photo is added to the project folder, the Flow begins:



# Power Automate “flow”

1. The image is automatically OCR’d (an AI option reads handwritten labels)
2. Text is added to relevant fields in the migration log including transcription, image file name, timestamp, and technician



# Automation to glean disk metadata




**A Solution: Microsoft Power Automate**

Flows can be duplicated and edited for future projects

Flows can be shared within and across repositories

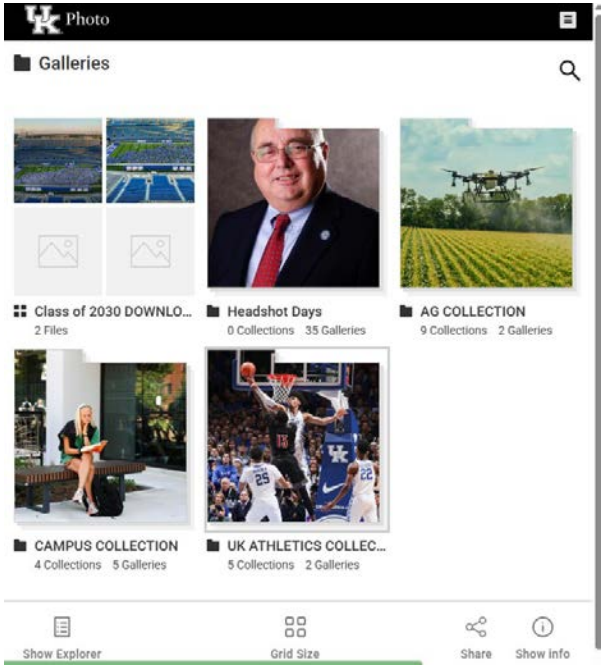
Flows can connect to non-Microsoft tools, including Google Drive, Zapier, Trello, and many others

Human quality control checking: still necessary

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Unique Identifier | PM format | Transcription from case and/or PM | Virus Scan Results | # of extracted files | Total size | Migration date | Server Location | Tech |
| 3 | 32_CD | | MICHAEL JOHNATHON<br>WOODSONES<br>OLD-TIME RADIO HOUR<br>1. The Show / 2. Celebrity Promo / 3. Ric<br>Recorded by<br>PLANET III<br>POET.ALAN<br>Lexington, KY<br>606 225-0333<br>RECORDS USA<br>606-255-5700<br>SHOW<br>32<br>AFFILIATE COPY<br>NOT FOR SALE<br>E 1999 POETMAN RECORDS USA<br>All rights reserved. Made in the USA | Scan - 2026-03-27 09_03_30.pdf | | | 3/27/2026 | | McDonnell |

# Testing the Waters:
# Other AI for Wrangling Big Born-Digital

1. Bespoke scripting for obtaining challenging collections: Photoshelter for instance —---------->
2. Whisper AI (via Blaze) for audio/video transcription and appraisal
3. EAD and Spreadsheet analysis and generation



# Thank You

ODNFEST 2026

**Lara Friedman-Shedlov** | ldfs@umn.edu
**Andrew McDonnell** | mcdonnell@uky.edu