LTO Tapes at WGBH

and the American Archive of Public Broadcasting

The Project

- American Archive of Public Broadcasting (AAPB) – 40,000 hours of a/v content from around the country
- Partnership between WGBH and the Library of Congress to work with stations to deliver material, with Crawford Media Services as vendor
- At WGBH, 300 TB of digital a/v material (11,561 files) identified for inclusion in the American Archive

The Data Transition

- WGBH backup data stored on LTO 4
- Administered through a SAMFS/QFS storage management system
- Metadata in Artesia Digital Asset Management system

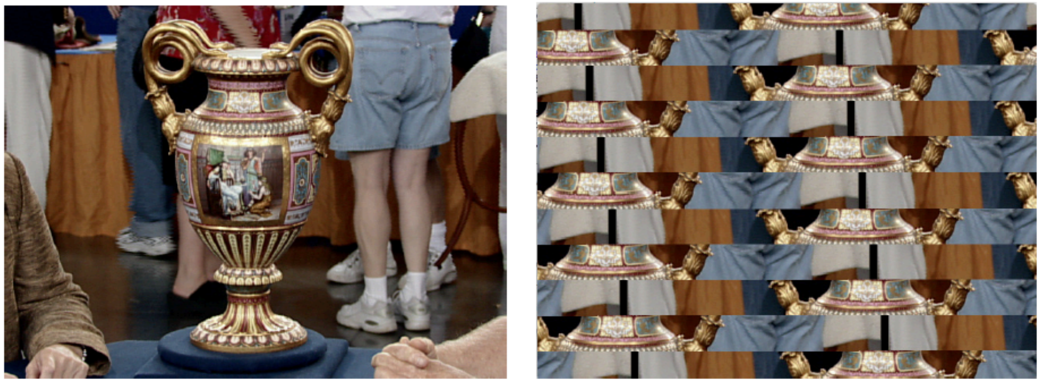
Use of Artesia DAM being discontinued at WGBH – opportunity to pull media and data out of Artesia-SAMFS/QFS system & store it locally on the archive’s dedicated LTO-6 tapes

The Data Transition

- List of files characterized as media reported by Artesia
- SAMFS/QFS queried to generate report on which LTO-4 tapes those files were stored on
- Files organized into ‘batches’ for download to local storage based on reported storage location
- Files transferred remotely from LTO-4 via ssh and downloaded onto external hard drives
- Drives shipped to Crawford Media Services for transcode and upload to Archival Management System

Initial Problems

- 57% of files in first large batch of video (2069 files) sent to Crawford proved to be incomplete or unreadable
- Several types of failure: 0-byte files, files that failed analysis by standard tools such as ffmpeg and mediainfo, files that could be analyzed but failed QC

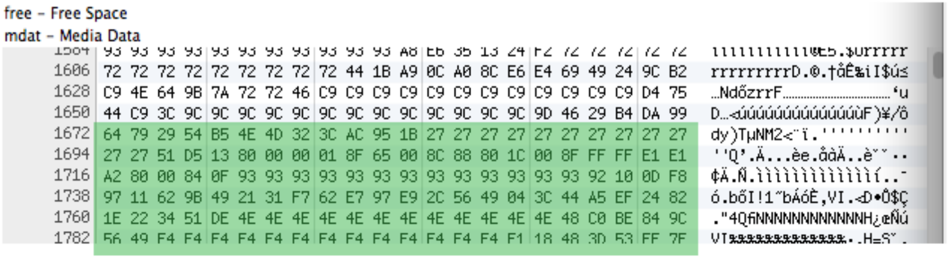


QC Failure: File reads as normal when tested, but content is replaced with repeated glitching image or green screen a portion of the way through the video

Deeper Analysis

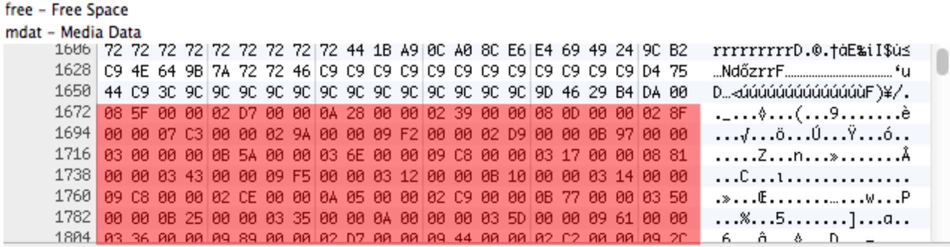
- Most failed files transferred a certain amount of data successfully, then eventually cut out and replaced with nonsense ‘filler’ data

Media data for uncorrupted file



```
free – Free Space
mdat – Media Data
1584 93 93 93 93 93 93 93 93 93 93 93 A8 E6 35 13 24 F2 72 72 72 72 72   ìììììììììì®Ê5.$òrrrrr
1606 72 72 72 72 72 72 72 72 72 44 1B A9 0C A0 8C E6 E4 69 49 24 9C B2   rrrrrrrrrD.©.†åÊ‰iI$ú≤
1628 C9 4E 64 9B 7A 72 72 46 C9 C9 C9 C9 C9 C9 C9 C9 C9 C9 C9 C9 D4 75   …NdőzrrF……………………‘u
1650 44 C9 3C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9D 46 29 B4 DA 99   D…<úúúúúúúúúúúúúùF)¥/ô
1672 64 79 29 54 B5 4E 4D 32 3C AC 95 1B 27 27 27 27 27 27 27 27 27 27   dy)TµNM2<¨ï.''''''''''
1694 27 27 51 D5 13 80 00 00 01 8F 65 00 8C 88 80 1C 00 8F FF FF E1 E1   ''Q’.Ä...èe.åàÄ..èˇˇ··
1716 A2 80 00 84 0F 93 93 93 93 93 93 93 93 93 93 93 93 93 92 10 0D F8   ¢Ä.Ñ.ììììììììììììí..¯
1738 97 11 62 9B 49 21 31 F7 62 E7 97 E9 2C 56 49 04 3C 44 A5 EF 24 82   ó.bõI!1˜bÁóÈ,VI.<D•Ô$Ç
1760 1E 22 34 51 DE 4E 4E 4E 4E 4E 4E 4E 4E 4E 4E 4E 4E 48 C0 BE 84 9C   ."4QﬁNNNNNNNNNNNNH¿æÑú
1782 56 49 F4 F4 F4 F4 F4 F4 F4 F4 F4 F4 F4 F4 F4 F1 18 48 3D 53 FF 7F   VIôôôôôôôôôôôôôò.H=Sˇ.
```

Media data at point of corruption



```
free – Free Space
mdat – Media Data
1606 72 72 72 72 72 72 72 72 72 44 1B A9 0C A0 8C E6 E4 69 49 24 9C B2   rrrrrrrrrD.©.†åÊ‰iI$ú≤
1628 C9 4E 64 9B 7A 72 72 46 C9 C9 C9 C9 C9 C9 C9 C9 C9 C9 C9 C9 D4 75   …NdőzrrF……………………‘u
1650 44 C9 3C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9C 9D 46 29 B4 DA 00   D…<úúúúúúúúúúúúúùF)¥/.
1672 08 5F 00 00 02 D7 00 00 0A 28 00 00 02 39 00 00 08 0D 00 00 02 8F   ._...◊...(...9.......è
1694 00 00 07 C3 00 00 02 9A 00 00 09 F2 00 00 02 D9 00 00 0B 97 00 00   ...√...ö...Ú...Ÿ...ó..
1716 03 00 00 00 0B 5A 00 00 03 6E 00 00 09 C8 00 00 03 17 00 00 08 81   .....Z...n...».......Å
1738 00 00 03 43 00 00 09 F5 00 00 03 12 00 00 0B 10 00 00 03 14 00 00   ...C...ı..............
1760 09 C8 00 00 02 CE 00 00 0A 05 00 00 02 C9 00 00 0B 77 00 00 03 50   .»...Œ.......…...w...P
1782 00 00 0B 25 00 00 03 35 00 00 0A 00 00 00 03 5D 00 00 09 61 00 00   ...%...5.......]...a..
1804 03 36 00 00 09 89 00 00 02 D7 00 00 09 44 00 00 02 C2 00 00 09 2C   .6...â...◊...D...¬...,
```

Process Evolution

- Each ‘batch’ of files was analyzed by ffmpeg immediately after download to determine which files failed
- Many files that initially failed could be successfully downloaded on a second or third try
- Other failed files (QC failure) were not detected until re-analysis after the drives returned to WGBH
- Final tally: 9957 out of 13492 were successfully re-ingested at WGBH -- 74% of files overall

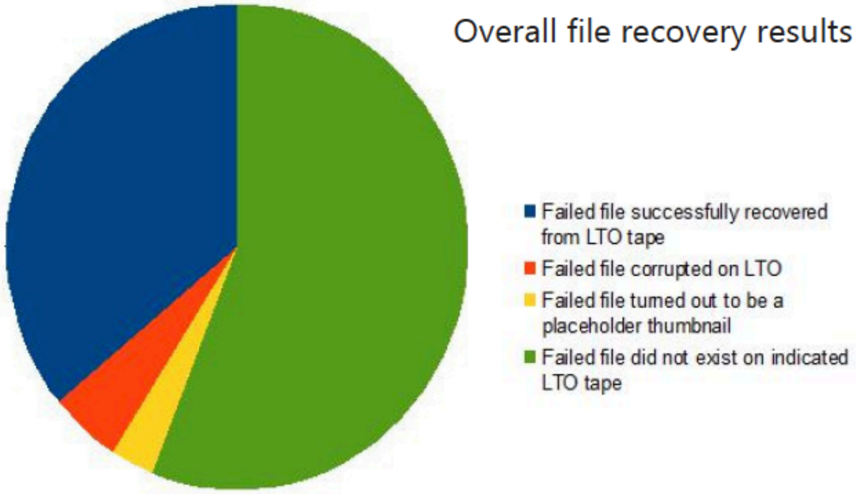
LTO Investigation

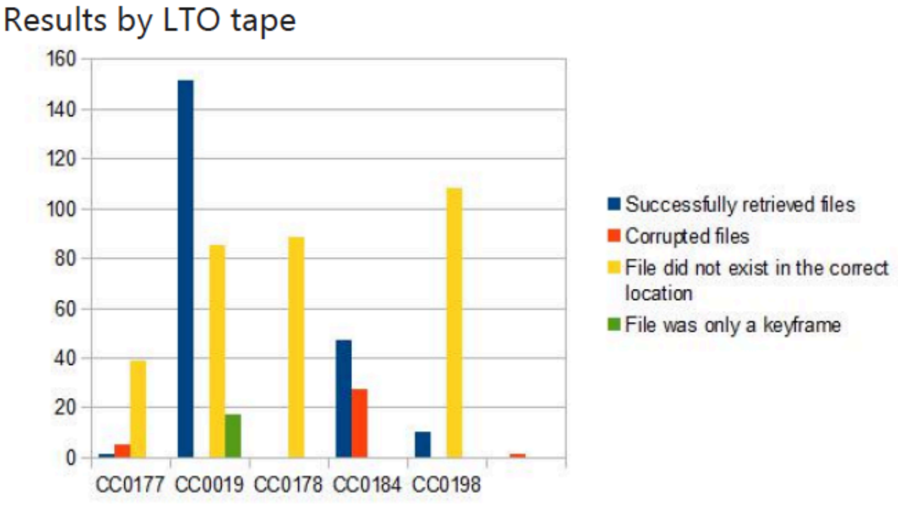
- Ran checksums on all the files that had existing checksums recorded – 1005 failed files passed checksum analysis when checked in storage, 20 checksums could not be generated

- Generated analysis of failed files by LTO-4 storage tape and identified tapes with greatest number of failures, then requested then from IT

- Used MLA LTO-6 decks to do a direct data dump with dd of all the tar files written to the LTO-4 tapes

LTO File Recovery Results





Inconclusive Conclusions

- Most files are still good on tape
- Some of the failures may be due to corruption on the LTO tapes themselves, but only a small percentage
- Processing problems may have been caused by inaccurate reporting from SAMFS/QFS

Next Steps

- Re-ingesting successful files onto local LTO-6 through direct connection
- Tracking location in XML files entered into local Filemaker-based DAM system
- Re-attempting networked transfer of the other 26% and analyzing results

Ideas or discussion welcome!

rebecca_fraimow@wgbh.org