Active and Retroactive Digital Newspaper Preservation

Andrew McDonnell
Digital Archivist
University of Kentucky Libraries
Special Collections Research Center
mcdonnell@uky.edu
Our Commitment to the Ephemeral

• Temporal value
• Digital is the new high-acid paper stock
• Fragile Bits | Bits are Fragile
• Two University of Kentucky Case Studies
The University of Kentucky’s Newspaper preservation

- Microfilm, starting ~1949
- Digitization, starting ~1998
- Born-digital, starting ~2018
- Well over 1 million pages:
  - kentuckynewspapers.org

For a tremendous history, read Kopana Terry and Eric Weig: https://uknowledge.uky.edu/libraries_present/137/
Case Study: La Voz de Kentucky

Bilingual, Bi-weekly Newspaper: 2003-2019
Optical Media

La Voz in its raw state:
362 discs, combination CD-Rs and DVD-Rs

Sometimes labeled.
Sometimes not.
Our friend the Nimbie

100 Discs at a Time!

• Format agnostic software
• Skipped disc imaging, conducted straight file migration
• Rip first, ask questions later (in this case)
Our enemy: Scratches and Goop

Too shiny to fully convey, but physical condition played a role in the process.
Post rip analysis: initial manual evaluation and tracking.

<table>
<thead>
<tr>
<th>Disc #</th>
<th>Transcription from case and/or disc</th>
<th>Contains Newspaper?</th>
<th>File Types</th>
<th>Size</th>
<th>Files</th>
<th>Folders</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0198</td>
<td>Musica Mac 3</td>
<td></td>
<td>Audio/Music</td>
<td>660.3 MB</td>
<td>145</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>0199</td>
<td>Musica 1</td>
<td></td>
<td>Audio/Music</td>
<td>574.6 MB</td>
<td>126</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>0200</td>
<td>Suplemento 10-10-10</td>
<td></td>
<td>Multiple types</td>
<td>233.4 MB</td>
<td>97</td>
<td>3</td>
<td>La Voz Vol 6 No 8: Apr 20-May 3, 2006, 32 P</td>
</tr>
<tr>
<td>0202</td>
<td>Recursos La Voz</td>
<td></td>
<td>Corrupt Disc</td>
<td>0.0 GB</td>
<td>0</td>
<td>0</td>
<td>Disc unopenable</td>
</tr>
<tr>
<td>0203</td>
<td>Blank</td>
<td></td>
<td>Photograph</td>
<td>101.2 MB</td>
<td>38</td>
<td>0</td>
<td>June-July 2007-08</td>
</tr>
<tr>
<td>0204</td>
<td>Blank</td>
<td></td>
<td>Blank disc</td>
<td>0 Bytes</td>
<td>0</td>
<td>0</td>
<td>empty</td>
</tr>
<tr>
<td>0205</td>
<td>Blank</td>
<td></td>
<td>PDF</td>
<td>30.5 MB</td>
<td>22</td>
<td>0</td>
<td>La Voz Vol 8 No 16: Aug 7-20 2008, 28 Page</td>
</tr>
<tr>
<td>0206</td>
<td>La Voz</td>
<td></td>
<td>Multiple types</td>
<td>12.4 MB</td>
<td>3</td>
<td>0</td>
<td>Logo</td>
</tr>
<tr>
<td>0207</td>
<td>Jet Set 2 Airport Lounge Riffage</td>
<td></td>
<td>Multiple types</td>
<td>586.9 MB</td>
<td>258</td>
<td>4</td>
<td>mostly mp3 files/ graphics/ mac install files/</td>
</tr>
<tr>
<td>0208</td>
<td>Adobe folders</td>
<td></td>
<td>Multiple types</td>
<td>127.5 MB</td>
<td>2192</td>
<td>156</td>
<td>La Voz Vol 5 No 7: Mar 31-Apr 13, 2005, 36</td>
</tr>
<tr>
<td>0209</td>
<td>Elmer, Español final</td>
<td></td>
<td>Audio/Music</td>
<td>88 Bytes</td>
<td>2</td>
<td>0</td>
<td>cda files</td>
</tr>
<tr>
<td>0210</td>
<td>Fonts P.C.</td>
<td></td>
<td>Multiple types</td>
<td>192.8 MB</td>
<td>2139</td>
<td>1</td>
<td>Fonts</td>
</tr>
<tr>
<td>0211</td>
<td>Blank</td>
<td></td>
<td>Multiple types</td>
<td>187.1 MB</td>
<td>96</td>
<td>3</td>
<td>La Voz Vol 6 No 20: Sep 21-Oct 4, 2006</td>
</tr>
<tr>
<td>0212</td>
<td>Adobe Periodico</td>
<td></td>
<td>Multiple types</td>
<td>64.6 MB</td>
<td>5</td>
<td>0</td>
<td>Dr. Carrillo photo, Photoshop images</td>
</tr>
</tbody>
</table>
Appraisal while you work.

TreeSize Analysis

- TreeSize overall file characterization and file manifest
- File extension analysis revealed more than newspaper files in the collection
- Multiple ways of visualizing, exploring a collection
Appraisal while you work.

Treesize Analysis
Manual Intervention revealed
• iTunes library backups
• System files
• Full-length feature films
Rebuilding InDesign

Adobe CS3 for Mac

- Fonts incompatible with most of our Windows machines
- Links, etc. formatted for Mac
- Happily, we are pack rats

2009 iMac
End Result

Panning for Gold

• Initial appraisal: 54,873 files across 1,902 directories
• 102 full PDF issues across eight years, 2003-2011
• Gaps, but a significant addition to the collection, and only known copy of any type in many instances

Coda:
More where that came from. Maybe.
Web Archiving

Newspapers in Kentucky

• UK Libraries Web Archives established in 2018
• Established workflow for Kentucky Kernel
  (Libraries have collected since 1915)
• Workflow + manual in place, but Kentucky Lantern was our first external newspaper website accession
The Kentucky Lantern

Live Site

Archived
Metadata across platforms

Title: Kentucky Lantern website
URL: https://kentuckylantern.com

Description: The Kentucky Lantern is an independent, nonpartisan, free news service based in Frankfort, Kentucky, part of States Newsroom, a nonprofit network of journalists that works to fill gaps in state government reporting caused by the declining numbers of state and local journalists. Captured 14 times between July 21, 2023 and May 06, 2024

Videos: 3 Videos Captured
Subject: Health – Kentucky, Government – Kentucky, Environment – Kentucky, Economy – Kentucky, Journalism – Kentucky
Creator: Kentucky Lantern
Coverage: Kentucky
Format: website
Type: electronic newspaper
Rights: "Our stories may be republished online or in print under Creative Commons licence CC BY-NC-ND 4.0."
Collector: University of Kentucky Libraries Special Collections Research Center, University of Kentucky Libraries

Archive-It: Dublin Core
Metadata across platforms

Kentucky Lantern website

Basic Information
- **Title**: Kentucky Lantern website
- **Identifier**: 2023mes006
- **Accession Date**: 2023-08-17
- **Public?**: True
- **Content Description**: The Kentucky Lantern is an independent, nonpartisan, free news service based in Frankfort, Kentucky, part of States Newsroom, a nonprofit network of journalists that works to fill gaps in state government reporting caused by the declining numbers of state and local journalists. Records harvested through Archiva-l, crawled monthly.

- **Acquisition Type**: Gift
- **Resource Type**: Papers
- **Restrictions Apply?**: False
- **Access Restrictions?**: False
- **Use Restrictions?**: False

Dates
- **Creation**: 2023

Expression
- **2023**

Type
- **Inclusive Dates**

Label
- **Creation**

ArchivesSpace: DACS
Preparing for problems

Archiveready.com

Is your website Archive Ready?

What is ArchiveReady?
An online tool which evaluates if a website will be archived correctly by web archives, such as the Internet Archive.

Why bother?
Because not all websites are archive ready and this may result in invalid web archives and ultimately in information loss.

Who is it for?
**Web professionals** who need to check if their websites are archive ready.
**Web archive engineers** who need to evaluate target websites before harvesting and ingestion.
**Students** who need to learn more about digital archiving.

How does it work?
ArchiveReady analyses your website (i.e. HTML, Images, CSS, JS, Sitemaps) and performs complex evaluations in order to calculate **Website Archivability** from a set of Archivability Facets: Accessibility, Cohesion, Metadata & Standards Compliance.
Preparing for problems

Webrecorder

Web archiving for all!

ReproZip

The Reproducibility Packer!
Takeaways

• Paper --> Microfilm --> Digitization
• Born-Digital--> Stay flexible.
• Automation is great, but only as good as your media sometimes.
• Donor/publication participation is key, but sometimes independent intervention is necessary.
• Advocacy is crucial. Benign neglect won’t save us this time. Mostly.
References


Thank You

Andrew McDonnell
Digital Archivist
University of Kentucky
Special Collections Research Center
mcdonnell@uky.edu