Value through longevity: File format migration using open source tools

State Library of North Carolina
Lisa Gregory
Jennifer Ricker

Photo, flickr, DaseinDesign
Digital Preservation Strategies

What is Digital Preservation?

Image, "Orange Marilyn 1962" by Andy Warhol
Digital Preservation Strategies

Options for making files accessible over time

- Emulation
  - Original hardware & software
  - System that mimics original hardware & software

Digital Preservation Strategies

Options for making files accessible over time

- Migration
  - Transformation to “stable” format upon ingest
  - Transformation to “stable” format when current format reaches obsolescence

Image, http://blog.builddirect.com/industryinsights/a-little-better-every-day-really-adds-up/
If We Can Do It, Anyone Can!

- Small library
- Small budget
- Little IT support

Our Approach to Migration Testing

- Investigate what others were doing
- Inventory what we had
- Identify how others were transforming the type of files that we had

Image, http://blog.riskmanagers.us/?attachment_id=2765
Our Approach to Migration Testing

- Determine the transformation path and tool
- Document our expected result
- Perform the transformation
- Evaluate the results
What We Have

- Audio/Video
  - MOV
  - MP3
  - WAV
- Images and Structured Graphics
  - GIF
  - JPG
  - TIF (compressed)
- Web Archives
  - ARC
- Document-Like
  - CSS
  - DOC
  - DOCX
  - HTLM
  - PDF
  - PPT
  - PUB
  - RTF
  - TXT
- Spreadsheets
  - XLS
- Geospatial
  - SHP
  - SHX
  - DBF
Migration Paths/Tools We Chose

- Ffmpeg
  - MP3 to WAV
  - MOV to AVI
- Inkscape
  - AI to SVG
- PLANETS Testbed
  - GIF, JPG, PSD to TIF
  - DOC & DOCX to ODT
  - CSS & HTML to TXT
  - PUB & RTF to PDF/a
- XENA
  - GIF, JPG, PSD to TIF
  - DOC & DOCX to ODT
  - PPT to ODP
  - XLS to ODF
  - CSS & HTML to TXT
  - PUB to PDF/a
- ArcMap/TerraGo
  - SHP, SHX, &DBF to GeoPDF (TerraGo) & Geospatial PDF (ArcMap)
Why these tools?

- Free
- Open source
- Documented
- Supported
- Audit trail/reporting
- Easy to use (preferably with GUI... not command line)
- Versatile (transforms multiple formats, single or batch, etc.)
Our Expectations

- No visual/auditory loss of content
- No loss of metadata
- Minimal degradation in quality (look & feel)
- Minimal degradation in structure (comprehensibility)
- Minimal degradation in interactivity (functionality)
## Findings: ffmpeg

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio/Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.mov</td>
<td>.mpeg-2 + mxf wrapper</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>.mp3</td>
<td>.wav file + bwf header</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Considerable degradation in video and audio quality.</td>
</tr>
</tbody>
</table>
## Findings: Inkscape

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Images and Structured Graphics</td>
<td>.ai</td>
<td>.svg</td>
<td>✓</td>
<td>✓*</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Yes, but with some loss.*
Example of font changes

The Digital Repository is a service of the State Library of North Carolina, a division of the Department of Cultural Resources. First made available in the spring of 2008, this service supports instructional and research needs related to the history and culture of North Carolina by making many of the unique and valuable holdings of the State Library of North Carolina easily accessible via a fully searchable, online database. To this end, the primary focus of the Digital Repository is on current and historical North Carolina state government information.
## Findings: PLANETS Testbed

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document-Like</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.css</td>
<td>.txt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>.doc (all versions)</td>
<td>.odt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>.docx</td>
<td>.odt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>.html</td>
<td>.txt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>.pdf</td>
<td>.pdf/a</td>
<td>✗</td>
<td>n/a</td>
<td>n/a</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>.pub</td>
<td>.pdf/a</td>
<td>✗</td>
<td>n/a</td>
<td>n/a</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>.rtf</td>
<td>.pdf/a</td>
<td>✗</td>
<td>n/a</td>
<td>n/a</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td><strong>Images and Structured Graphics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.psd</td>
<td>.tif (uncompressed)</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>Rendered, but no .psd functionality (layers, etc.) retained.</td>
</tr>
<tr>
<td>.gif</td>
<td>.tif (uncompressed)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>File header's &quot;modified date&quot; was changed to experiment date.</td>
</tr>
<tr>
<td>.jpg</td>
<td>.tif (uncompressed)</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td></td>
</tr>
</tbody>
</table>
## Findings: Xena

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document-Like</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.css</td>
<td>.txt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>.doc (all versions)</td>
<td>.odt</td>
<td>✓</td>
<td>✓*</td>
<td>✓</td>
<td>✓</td>
<td>Tables &amp; tabs did not render exactly in XENA; fine in OpenOffice.</td>
</tr>
<tr>
<td>.docx</td>
<td>.odt</td>
<td>✓</td>
<td>✓*</td>
<td>✓</td>
<td>✓</td>
<td>Bullets did not render exactly in XENA; fine in OpenOffice.</td>
</tr>
<tr>
<td>.html</td>
<td>.txt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>.pdf</td>
<td>.pdf/a</td>
<td>X</td>
<td>n/a</td>
<td>n/a</td>
<td>X</td>
<td>Tool could not accommodate migrating .pdf to .pdf/a.</td>
</tr>
<tr>
<td>.pub</td>
<td>.pdf/a</td>
<td>X</td>
<td>n/a</td>
<td>n/a</td>
<td>X</td>
<td>Tool could not accommodate migrating .pub to .pdf/a.</td>
</tr>
<tr>
<td><strong>Images and Structured Graphics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.psd</td>
<td>.png</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Less crisp than the original.</td>
</tr>
<tr>
<td>.gif</td>
<td>.png</td>
<td>✓</td>
<td>✓*</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>.jpg</td>
<td>.png (uncompressed) or .tif (uncompressed)</td>
<td>X</td>
<td>n/a</td>
<td>n/a</td>
<td>X</td>
<td>Simply wraps in XML.</td>
</tr>
<tr>
<td><strong>Spreadsheets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.xls</td>
<td>.odf</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>X</td>
<td>Author, manager, company metadata lost.</td>
</tr>
</tbody>
</table>

*Yes, but with some loss.*
Example of display issues with the XENA viewer

Discussion Points

• Look at different ways of finding and delivering content
• Utilize existing resources and build from those
• Get feedback from people who don’t know the project to get different perspectives
• Future directions to deliver content

May 3, 2007
# Findings: ArcMap/TerraGo

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Geospatial</td>
<td>.shp, .shx, .dbf</td>
<td>.pdf</td>
<td>✓</td>
<td>✓</td>
<td>✓,*</td>
<td></td>
</tr>
</tbody>
</table>

- **ArcMap:** Embedded metadata is not currently accessible in Adobe.
- **Both tools:** Most metadata is contained in a separate .xml file. The converting tool seems to ignore.

*Yes, but with some loss.

**NOTE:** ArcMap and TerraGo are both proprietary software tools.
Where are we now?
Challenges expected and found

- Complex, related files
- Layers
- Proprietary + less widely used
File Format Observations

- Surprises
  - Audio-video formats have their own complexities
  - Frame rates, compression, and codecs, oh my!
### File Format Observations

- **Surprises**
  - PDF/Argh

<table>
<thead>
<tr>
<th>1a</th>
<th>1b</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 1b restrictions PLUS</td>
<td>• Self-contained</td>
</tr>
<tr>
<td>• Defined document structure (tags)</td>
<td>• No external references</td>
</tr>
<tr>
<td>• Better accessibility</td>
<td>• Lowest level of compliance</td>
</tr>
<tr>
<td></td>
<td>• Digitized materials</td>
</tr>
<tr>
<td></td>
<td>• Metadata is required</td>
</tr>
</tbody>
</table>

---

*Photo, etsy, icehousecrafts*
Want to give it a go?

- Good tools to have:
  - FFmpeg
  - FITS
  - FLAC Frontend
  - Ghostscript
  - Inkscape
  - MPEG Streamclip
  - PLANETS Testbed (RIP?)
  - XENA
Free & open source has downsides

- “Free” in *upfront* costs
- Might be developed by single person, or by hundreds
- Learning curve can be steep
Documentation can be confusing/nonexistent

Permissions/Disclaimers/Warnings!

Downloading instructions will go here. Not yet written.

A snapshot of the HRV software, as of Feb. 3, 1998, is contained in the attached tar file. This software is functional but should be considered at best an "alpha" release. Some parts of the software use functions from the nonlinear dynamics time series software as well as from various Matlab toolboxes such as statistics. Eventually I will make a clean distribution, but for the present this is the best I have time to do.

Can you rock the command line?

- gswin32c -dPDFA -dBATCH -dNOPAUSE -dNOOUTERSAVE -dUseCIEColor -SDEVICE#pdfwrite -sOutputFile=newfilename inputfilename
More helpful knowledge

- Build in time for stops along the road
  - Tool installation
    - You probably won’t have the ideal configuration
  - Troubleshooting
    - You may not get great error feedback
  - General Googling for assistance
    - You might not be able to rely on the software documentation for help
There are still unknowns

On-the-fly or scheduled bulk migration?

QA – what should we use/rely on?

How can we facilitate batch processing?

QA – how much should we do?

ARC to WARC?
Where do we go from here?

- Overcoming challenges to production implementation
  - Usual culprits: staff time, resources, IT restrictions, programming skills
  - The existence of technologies and workflows are not the main problems
Where do we go from here?

- Testing Archivematica (archivematica.org), by Artefactual Systems
- “Archivematica is a comprehensive digital preservation system. Archivematica uses a micro-services design pattern to provide an integrated suite of free and open-source tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model.”
Where do we go from here?

OUR “TOOLS TO HAVE”

- FFmpeg
- FITS
- FLAC Frontend
- Ghostscript
- Inkscape
- MPEG Streamclip
- PLANETS Testbed (RIP?)
- XENA

ARCHIVEMATICA

- digiKam DNG Convertor
- Document Converter
- FFmpeg
- FITS
- Imagemagick
- Inkscape
- OpenOffice.org
- PyODConverter Daemon
Where do we go from here?

- Formal workflow description
  - OAIS compliant
  - Multiple sources
  - Multiple stakeholders
  - On- and off-site storage
  - Tenuous IT capabilities
Where do we go from here?

- At-risk files
- At-riskier files
- Older files
- Older formats (Word 6, etc.)
- Obsolete formats
- Databases
- More work on a/v formats
Thank you

Jennifer Ricker
Jennifer.ricker@ncdcr.gov

Lisa Gregory
Lisa.gregory@ncdcr.gov

Photo, flickr, HarshLight