The “M” word: What works and what doesn’t in file format migration

SAA Research Forum

August 10, 2010

The State Library of North Carolina's Digital Information Management Program

- Archive-It
- CONTENTdm
- local storage/OCLC's Digital Archive

Our Approach

- Look at what others were doing
- Look at what we had
- Identify how others were transforming the type of files that we had
- Determine the transformation path and tool
- Perform the transformation
- Evaluate the results

http://www.intersema.ch/company/strategy/

What did we have?

http://www.entertainmentearth.com/prodinfo.asp?number=HPD23500

- A/V
  - MOV
  - MP3
- Image
  - GIF
  - JPG
  - TIFF
  - PSD
  - AI
- Text
  - CSS
  - DOC
- Text, cont.
  - DOCX
  - PDF
  - PPT
  - PUB
  - RTF
  - TXT
  - XLS
- Web
  - HTM
  - ARC
  - WARC

What was our transformation process?

- MP3 to WAV (ffmpeg)
- GIF, JPG, PSD formats to TIFF (PLANETS/ ImageMagick)
- AI to SVG (Inkscape)
- DOC & DOCX formats to ODT (Xena)
- RTF to PDF/A (ConvertDoc)
- MOV to AVI (ffmpeg)
- PPT to ODP (Xena)
- CSS to TXT (Xena)
- PUB to PDF/A (Zamzar)
- XLS to ODF (Xena)

(CC) Larry D. Moore

What did we look for?

http://www.talentgate.ro/career_dev_process/index/item/evaluate

- In the tools
  - Free, open source, documented/supported
  - Easy to use (preferably with GUI…not command line)
  - Versatile (transforms multiple formats, single or batch, etc.)
  - Reporting
- In the transformations
  - Minimal loss of content/minimal degradation of audio/video quality
  - Retention of basic data structure (i.e., text structure of .css)
  - Retention of metadata
  - Retention of interactive functions (i.e., layers in images, formula/macros in spreadsheets, hyperlinks, etc.)
  - Retention of security
  - Retention of embedded content (i.e., audio, images, etc.)

Drum Roll Please…

Walt Disney Studios

Findings: Audio/Video Files

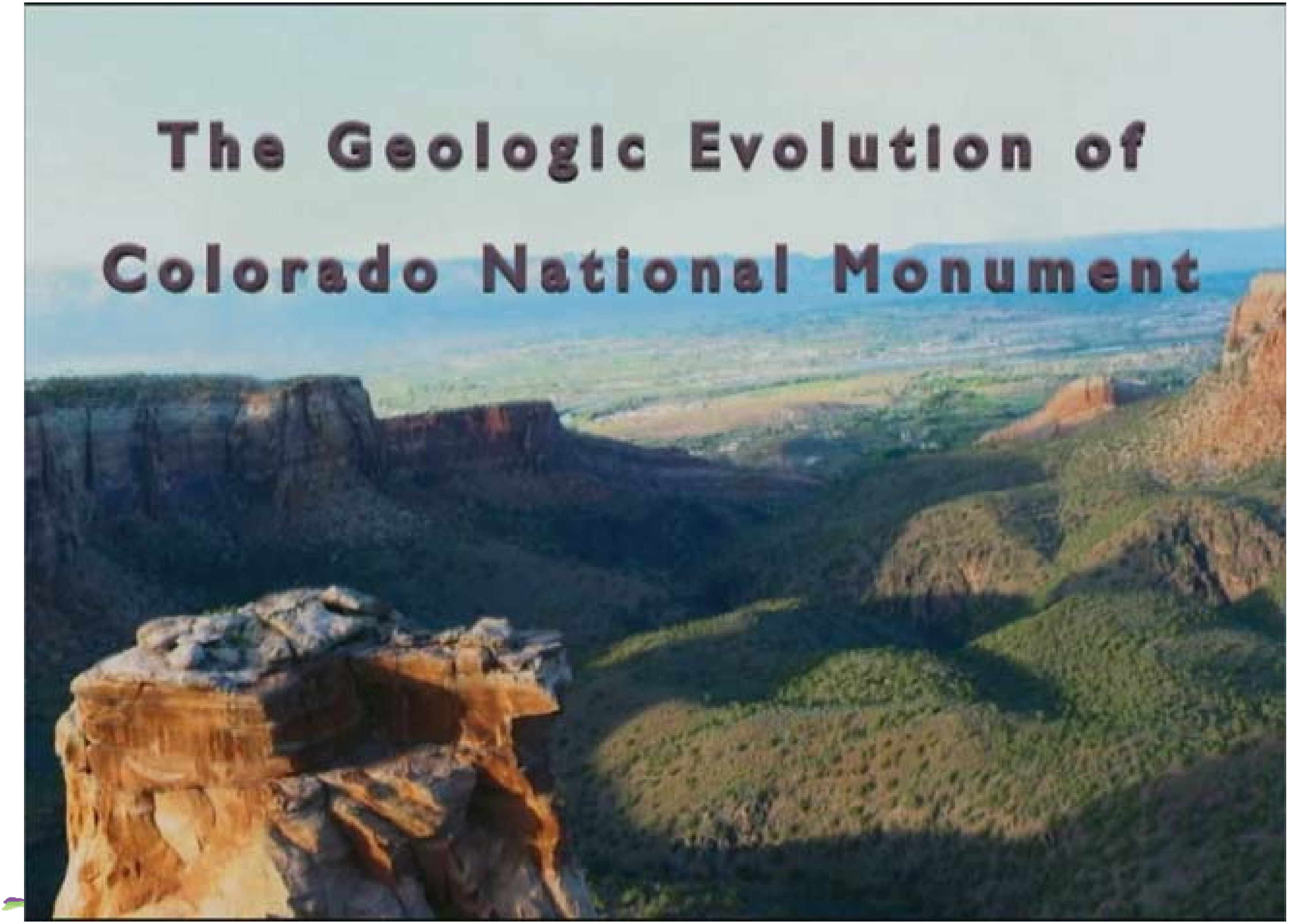
Couldn't find a free/open-source .mov to .mj2 conversion tool so…

- Converting .mov to .avi resulted in significant degradation of the quality of the video
- Converting .mov to .mpeg resulted in a similar degradation (as we expected)

Converting .mp3 to .wav resulted in no noticeable loss of content or quality

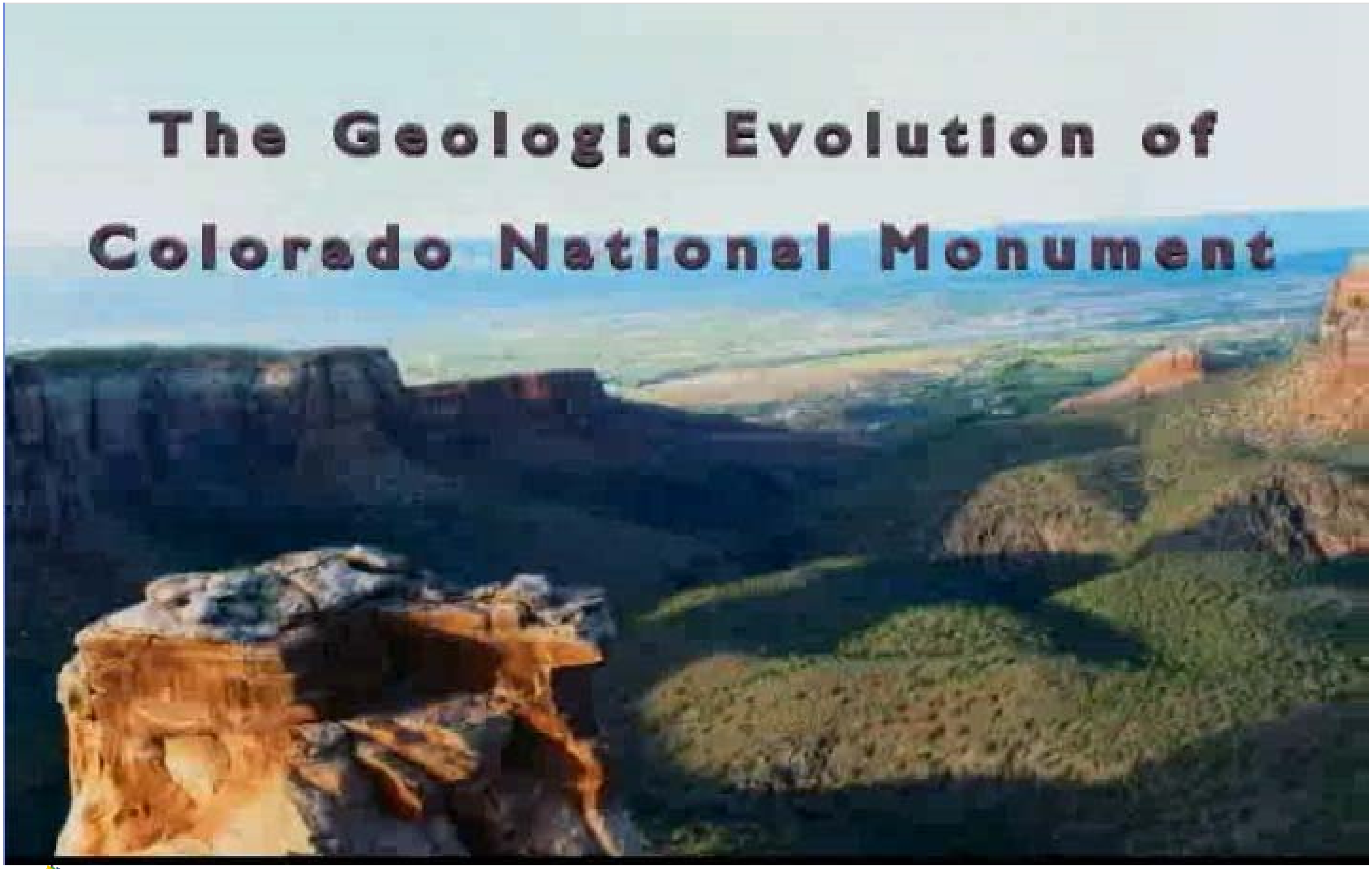
.mov (original)

Example: Video Degradation



.avi (transformation)

Example: Video Degradation



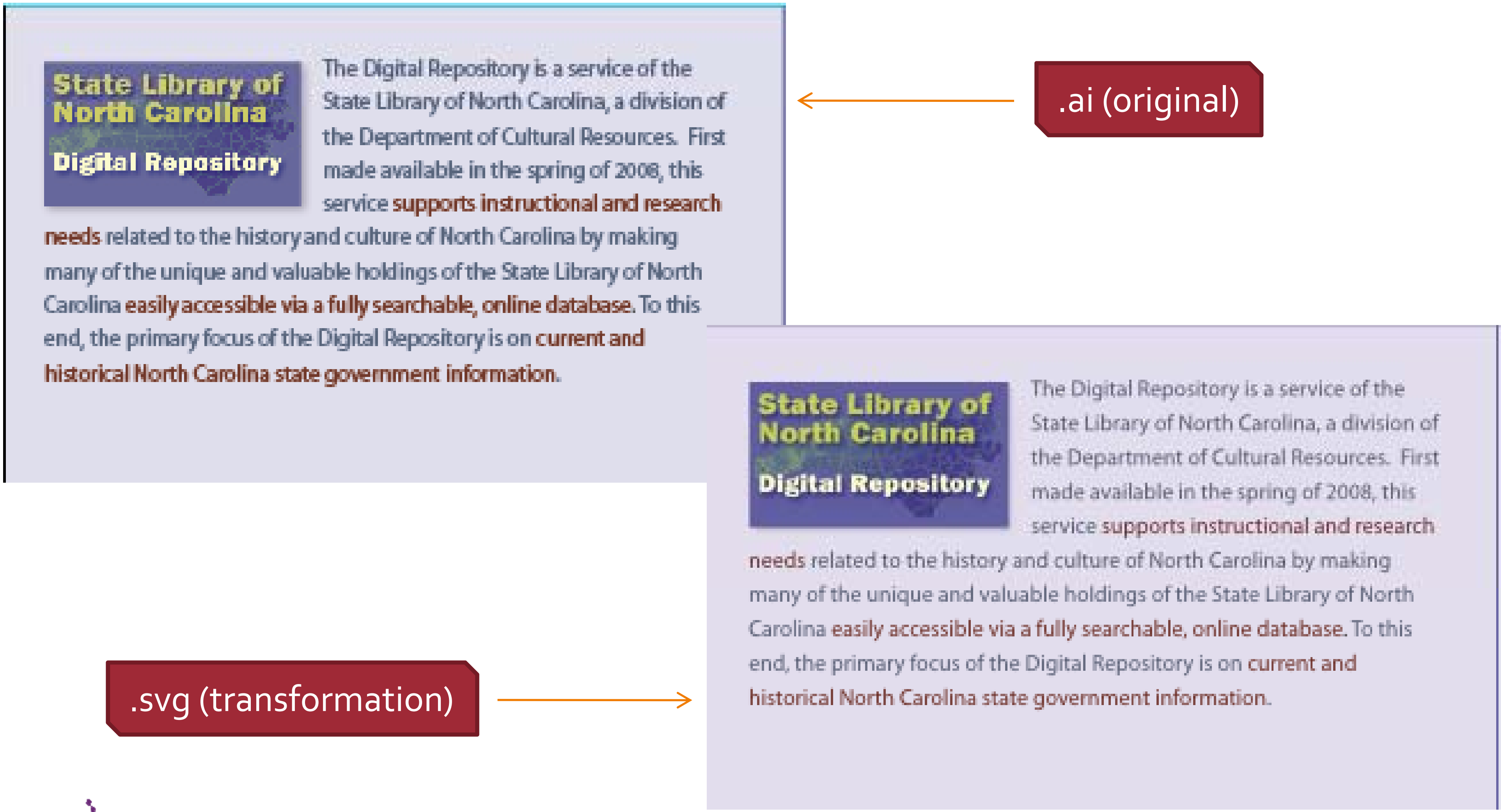
Findings: Image Files

None of the conversions resulted in any noticeable loss of content , however…

- Converting .ai to .svg caused some font formatting and coloring to be a bit off
- Converting .gif to .png resulted in degraded image quality
- Converting .jpg to .tif resulted in some metadata loss

Interestingly, converting .psd to .tif resulted in marginal image quality improvement

Example: Font Change



.ai (original)

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.

.svg (transformation)

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: Text Files

Again, none of the conversions resulted in any noticeable loss of content and

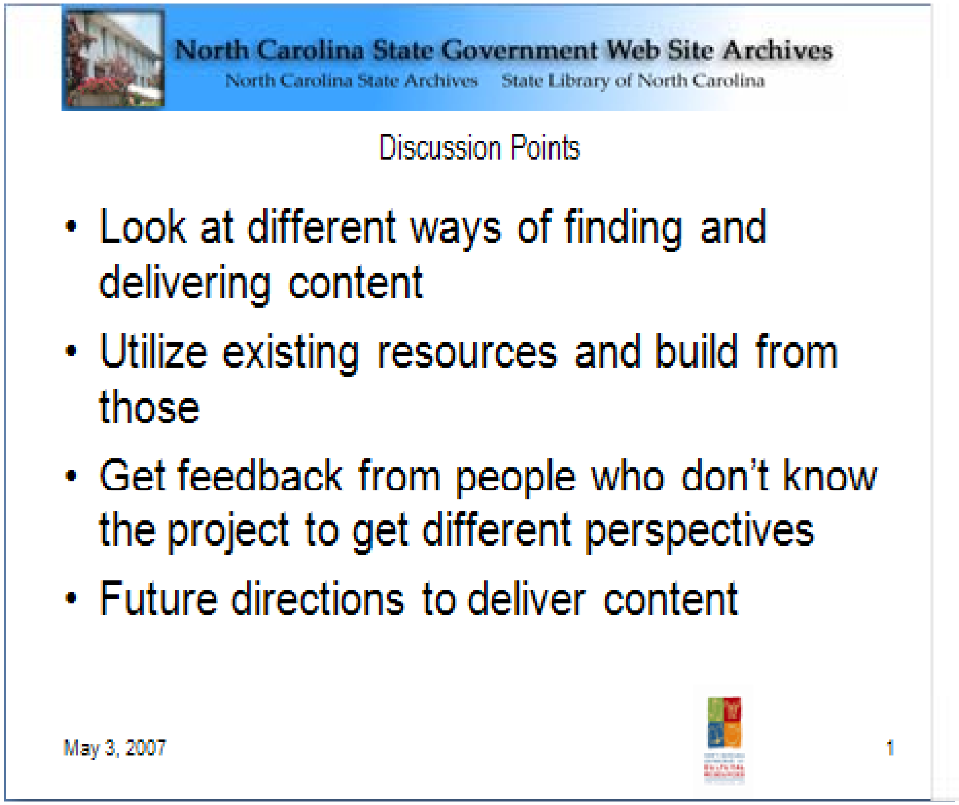
- Converting .css to .txt resulted in no noticeable loss of data structure
- In general, converting to open document formats resulted in links and embedded content remaining intact, as well as interactive functionality like formula and macros

However…

- Converting .doc to .odt resulted in table and tab structures being lost
- Converting .docx to .odt resulted in some formatting (like bullets) was altered
- Converting .ppt to .odp resulted in the loss of page numbers and changes to fonts and footers
- Converting .pub to .pdf/a resulted in the ability to edit the file was lost
- Converting .rtf to .pdf/a and .xls to .odf resulted in the loss of some metadata

Example: Wonky Footer

.ppt (original)



North Carolina State Government Web Site Archives

North Carolina State Archives State Library of North Carolina

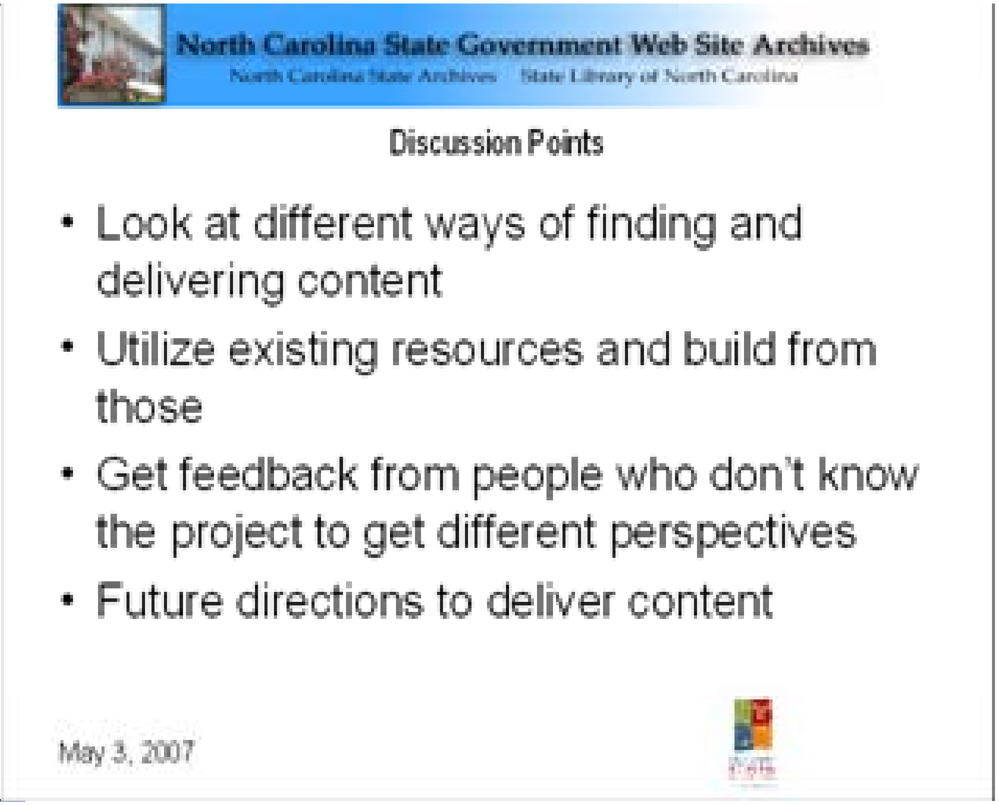
Discussion Points

- 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

1

.odp (transformation)



North Carolina State Government Web Site Archives

North Carolina State Archives State Library of North Carolina

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

What we found

- In general, the migrations went much better than we had expected - the loss we experienced was loss that we could tolerate
- Loss of content doesn't seem to be an issue for these more "modern" file types
- Need to be comfortable using a command line to invoke migration actions with the open source tools available today

http://physics.wku.edu/~bonham/Talks/FacultyWorkshop/index.php?part=Motivation&item=Dump

What can we learn from you?

- Do you know any free, open source tools to convert:
  - .mov to .mj2
  - .tif (compressed) to .tif (uncompressed)
  - .arc to .warc
- How does your process differ from ours? Do you have any suggestions for us?
- If you have transformed older files, how did your results differ from ours? Are there issues we might not expect that we should be prepared to face?

http://garrmdc.blogspot.com/2007/06/knowledge-management.html jennifer.ricker@ncdcr.gov