Harvard University Library

Digital Library Federation Fall Forum
Charlottesville, November 7-9, 2005

Maintaining Archive Integrity During Inter-Repository Transfer:

Lessons Learned from the NDIIPP Archive Ingest and Handling Test (AIHT)

Stephen L. Abrams
Harvard University Library

stephen_abrams@harvard.edu

Phase I ingest

- Digital Repository Service (DRS) intended for highly "curated" digital objects, those that are
  - Created through known workflows and according to known technical specifications
  - In a small set of approved formats
  - Accompanied by reliable metadata
- These constraints are enforced throughout the repository
  - SIP schema
  - API
  - Database
- Accompanying metadata was non-existent or unreliable

“By Jhove …”

- DSIP (SIP packaging tool) based on JHOVE
- JSTOR/Harvard Object Validation Environment
  - Extensible framework for format-specific object identification, validation, and characterization
  - Existing modules for ASCII, GIF, JPEG, PDF, TIFF, UTF-8, and XML
  - New modules for AIFF, HTML, JPEG 2000, and WAVE
  - Support for 93% of the AIHT corpus
  - The remaining 7% use more than 90 formats
    - PSD, Word, Flash, AVI, MP3, BMP, QT, PNG, MPEG, PS, …

Phase I inventory

Format	*MIME type*	*By manifest*	*By extension*	*By JHOVE*
AIFF	audio/x-aiff	162	151	162
ASCII/UTF-8	text/plain	20,207	10,822	30,887
GIF	image/gif	1,337	1,320	1,339
HTML	text/html	16,677	16,579	3,649
JPEG	image/jpeg	12,752	12,763	12,576
PDF	application/pdf	1,663	1,664	1,659
TIFF	image/tiff	1,538	1,533	1,537
WAVE	audio/x-wave	2,015	2,016	2,015
XML	text/xml	0	1	8
Unknown	application/octet-stream	1,141	10,613	3,618
		57,492	57,492	57,450

Phase I inventory

Format	*MIME type*	*By manifest*	*By extension*	*By JHOVE*
AIFF	audio/x-aiff	162	151	162
ASCII/UTF-8	text/plain	20,207	10,822	30,887
GIF	image/gif	1,337	1,320	1,339
HTML	text/html	16,677	16,579	3,649
JPEG	image/jpeg	12,752	12,763	12,576
PDF	application/pdf	1,663	1,664	1,659
TIFF	image/tiff	1,538	1,533	1,537
WAVE	audio/x-wave	2,015	2,016	2,015
XML	text/xml	0	1	8
Unknown	application/octet-stream	1,141	10,613	3,618
		57,492	57,492	57,450

Phase I inventory

Format	*MIME type*	*By manifest*	*By extension*	*By JHOVE*
AIFF	audio/x-aiff	162	151	162
ASCII/UTF-8	text/plain	20,207	10,822	30,887
GIF	image/gif	1,337	1,320	1,339
HTML	text/html	16,677	16,579	3,649
JPEG	image/jpeg	12,752	12,763	12,576
PDF	application/pdf	1,663	1,664	1,659
TIFF	image/tiff	1,538	1,533	1,537
WAVE	audio/x-wave	2,015	2,016	2,015
XML	text/xml	0	1	8
Unknown	application/octet-stream	1,141	10,613	3,618
		57,492	57,492	57,450

Phase II ingest

- Ingest of the Stanford DIP
  - Technical metadata provided
  - Direct DIP-to-SIP transform using Perl script
- Stanford technical metadata was generated, in part, using the JHOVE beta 2 release
- DSIP used the internal JHOVE beta 3 pre-release

Phase II inventory

Format	*MIME type*	*Phase I GMU/LC DIP*	*Phase II Stanford DIP*
AIFF	audio/x-aiff	162	162
ASCII/UTF-8	text/plain	30,887	30,910
GIF	image/gif	1,339	0
HTML	text/html	3,649	1,222
JPEG	image/jpeg	12,576	10,766
PDF	application/pdf	1,659	1,662
TIFF	image/tiff	1,537	1,537
WAVE	audio/x-wave	2,015	0
XML	text/xml	8	5
Unknown	application/octet-stream	3,618	11,186
		57,450	57,450

Phase II inventory

Format	*MIME type*	*Phase I GMU/LC DIP*	*Phase II Stanford DIP*
AIFF	audio/x-aiff	162	162
ASCII/UTF-8	text/plain	30,887	30,910
GIF	image/gif	1,339	0
HTML	text/html	3,649	1,222
JPEG	image/jpeg	12,576	10,766
PDF	application/pdf	1,659	1,662
TIFF	image/tiff	1,537	1,537
WAVE	audio/x-wave	2,015	0
XML	text/xml	8	5
Unknown	application/octet-stream	3,618	11,186
		57,450	57,450

Phase III migration

- Migration of GIF, JPEG, and TIFF source images to JPEG 2000 format
- “Preservation” goals
  - Pictorial integrity
  - Reuse
- For automated processing, the source images were grouped into 25 unique categories based on
  - Format
  - Color space
  - Compression
  - Bits/sample
  - Image size

Phase III migration specifications

- Aware 3.6.0 codec
  - JP2 profile
  - RLCP (resolution-layer-component-position) progression order
  - Tile size 1024×1024
  - Reversible 5-3 wavelet transform
  - Decomposition levels based on source image maximum pixel size
  - Two quality layers: 50% (35dB pSNR) and 100% (full quality)
  - Reversible channel quantization
  - Highest quality coding predictor offset
  - Grayscale/sRGB colorspaces for single/multi-channel images

```
% j2kdriver –set-input-image-type type file.ext \
  –t JP2 –p RLCP --tile-size 1024 1024 –w R53 levels –y 2 \
  --set-output-j2k-layer-psnr 0 35 –-set-output-j2k-layer-psnr 1 0 \
  -q ALL REVERSIBLE --predictor-offset 0 –o file.jp2
```

Phase III migration results

Format	*Color space*	*Compress*	*Bits/sample*	*Source*	*JPEG 2000*
GIF	Palette	LZW	8	1,339	706
JPEG	YCbCr	DCT	8	67	66
			8,8,8	12,501	8,117
			8,8,8,8	8	0
TIFF	Bitonal	None	8	6	6
	RGB	None	8,8,8	1,510	1,510
		LZW	8,8,8	16	11
		PackBits	8,8,8,8	5	5
				15,452	10,421

Phase III migration results

Format	*Color space*	*Compress*	*Bits/sample*	*Source*	*JPEG 2000*
GIF	Palette	LZW	8	1,339	706
JPEG	YCbCr	DCT	8	67	66
			8,8,8	12,501	8,117
			8,8,8,8	8	0
TIFF	Bitonal	None	8	6	6
	RGB	None	8,8,8	1,510	1,510
		LZW	8,8,8	16	11
		PackBits	8,8,8,8	5	5
				15,452	10,421

Phase III migration results

Format	*Color space*	*Compress*	*Bits/sample*	*Source*	*JPEG 2000*
GIF	Palette	LZW	8	1,339	706
JPEG	YCbCr	DCT	8	67	66
			8,8,8	12,501	8,117
			8,8,8,8	8	0
TIFF	Bitonal	None	8	6	6
	RGB	None	8,8,8	1,510	1,510
		LZW	8,8,8	16	11
		PackBits	8,8,8,8	5	5
				15,452	10,421

Phase III migration results

- Aware worked very quickly to address all the errors uncovered during the test
- All corrections are incorporated into the 3.7.1 codec

Phase III migration QC

- Automated
  - JHOVE verification of codec specifications
  - Photoshop source/target pixel comparison
    - RGB-to-sRGB transform was numerically lossless
    - YCbCr-to-sRGB transform had small round-off error (σ=0.02)
- Manual
  - Side by side viewing of source/target images
    - ISO 3664 conditions
    - Trained observers from Harvard College Library-Digital Imaging Group

Conclusions

- Need for standardization on common SIP and DIP formats
- Externally supplied technical metadata is suspect
- The lack of technical metadata should not prohibit ingestion
- A high degree of repository automation is possible
  - SIP generation using JHOVE
  - Image migration and QC