Automated Migration for Image Preservation

Stephen L. Abrams; Harvard University; Cambridge, Massachusetts, USA

Abstract

Format migration is recognized as one of the most important potential strategies for ensuring the long-term usability of stored digital assets. The Harvard University Library is investigating the issues surrounding the design and implementation of an automated process for large-scale format migration, with initial emphasis placed on transforming GIF, JPEG, and TIFF images to JPEG 2000 form. This paper will review the findings of Harvard's initial exploration of automated migration as a participant in the Library of Congress's Archive Ingest and Handling Test and subsequent activities towards providing a migration capability for its Digital Repository Service. The significant issues in this investigation include: source population eligibility, resultant image profile selection, automated workflows, post-migration quality assurance testing, provenance metadata, final disposition of source images, and applicable preservation policies.

Introduction

Format migration is recognized as one of the most important potential strategies for ensuring the long-term usability of stored digital assets [1]. The trigger event for the migration of an asset can be either the incipient obsolescence of the object's current format (indicated by a growing paucity of systems, tools, or services that can process or render the format appropriately) or a change in user requirements that cannot be accommodated by the current format. The Harvard University Library (HUL) has operated a preservation repository, the Digital Repository Service (DRS, <http://hul.harvard.edu/ois/systems/drs/>), for over five years, with more than 4 million objects (15 TB) under managed storage. Approximately 3 million of these are digital images, primarily GIF, JPEG, and TIFF surrogates of either material culture artifacts (paintings, prints, photographs, sculpture) or text (print and manuscript). While none of the formats in current use are believed to present a preservation risk, many DRS institutional clients have expressed interest in providing enhanced user functions such as interactive zoom, pan, and rotate that would be facilitated through the use of the JPEG 2000 format [2].

As a participant in the recent Archive and Ingest Handling Test (AIHT) organized by the Library of Congress as part of its National Digital Information Infrastructure Preservation Program (NDIIPP), HUL had an opportunity to investigate several of the issues surrounding an automated format migration [3]. During this test, Harvard successfully migrated over 10,000 GIF, JPEG, and TIFF source images to JPEG 2000 form. The results of the test suggest the feasibility of largely automated migration workflows and post-migration quality assurance (QA) testing.

Based on this experience HUL is now engaged in activities towards providing an automated image migration capability to the DRS. The potential source population for this conversion exhibits substantial heterogeneity in format, quality, color space, and compression. Moreover, many of these images do not exist independently, but rather, within families of related images associated through derivation or dependency relationships. The complexity of these relationships complicates the selection of the appropriate source object for a potential migration. Furthermore, a migration strategy requires consideration of preservation policies regarding the retention of source data: since some migrations will be undertaken in response to incipient obsolescence, should the source data be discarded as redundant, or should they be retained "just in case" they might prove useful in the future?

This paper will review the findings of Harvard's work in the AIHT and its subsequent migration activities with emphasis on source population eligibility criteria, resultant image profile selection, codec configuration, automated workflow design, post-migration quality assurance testing, documentation of provenance, final disposition of source images, and consideration of appropriate preservation policies.

Archive Ingest and Handling Test

Although the primary intent of the AIHT was to test the premise that significant bodies of digital content can be transferred easily and without loss between institutions utilizing radically different preservation architectures and technologies, the test also included a component during which participants investigated format migrations. The HUL investigation was organized around the transformation of all GIF, JPEG, and TIFF images found in the test corpus to JPEG 2000 form. This process was simplified by the assumption that all images in the corpus represented independent digital objects. In other words, HUL did not attempt to determine if derivation or other dependency relationships existed between images. As none of the source images were detected to include embedded color profiles and none of the TIFF images utilized colorimetry tags (Whitepoint or PrimaryChromacitities) [4], the advanced color management capabilities of the JPEG 2000 JPX profile were not necessary [5]. The resultant image profile was therefore based on the JP2 profile using lossless compression (reversible 5-3 wavelet), the sRGB and greyscale color spaces, decomposition levels based on image size, and two quality layers, corresponding to archival master and delivery roles.

The GIF format was not supported as a source format by the commercial codec used for the migration. Consequently, all GIF images were first converted to equivalent TIFF images using an open source tool. The final JPEG 2000 conversion for this GIF population failed in just under half of the cases (see Table 1). Post-mortem investigation revealed the source of the failure to be the use of transparent background color in the original GIF images, which was carried forward to their TIFF counterparts. The version of the codec used for the test did not support TIFF transparency. However, a subsequent release of the codec added this feature. Unfortunately, the AIHT schedule did not provide sufficient time to reprocess any of the migration failure cases.

Table 1: GIF Migration Results

Color space	Com-pression	Bits per sample	GIF files	JPEG 2000 files
Palette	LZW	8	1,339	706

Significant numbers of JPEG-to-JPEG 2000 conversions also failed (approximately 35%, see Table 2). Four component JPEG images were not supported by the codec and the remaining failures were traced to a codec bug. All of these problems were addressed by the codec vendor.

Table 2: JPEG Migration Results

Color space	Com-pression	Bits per sample	JPEG files	JPEG 2000 files
YCbCr	DCT	8	67	66
		8 8 8	12,501	8,117
		8 8 8 8	8	0

The success rate for direct TIFF-to-JPEG 2000 conversions was over 99% (see Table 3). The five failures were due to a codec bug, which was again quickly corrected by the vendor.

Table 3: TIFF igration Results

Color space	Com-pression	Bits per sample	TIFF files	JPEG 2000 files
Bi-tonal	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

Automated and manual QA testing was performed subsequent to the migration. The open source JHOVE validation tool was used to verify that all of the resulting JPEG 2000 images met the established specifications. JHOVE is an extensible Java framework for format-specific object identification, validation, and characterization (see <http://hul.harvard.edu/jhove/>) that supports all three of the source formats as well as JPEG 2000 .

Manual side-by-side viewing of before and after images was performed on a small, through representative, sample under ISO 3664 calibrated viewing conditions [6]. As an additional step, a commercial image processing application was used to perform pixel-by-pixel comparisons of source and target image data. All TIFF-to-JPEG 2000 conversions showed exact numerical equivalence. While the JPEG-to-JPEG 2000 conversions did show a small, statistically insignificant variance, this was judged to be clearly beneath the visually-lossless threshold by trained observers of the Harvard College Library Digital Imaging Group (HCL-DIG). The most likely cause of these discrepancies appears to be numerical round-off during the YCbCr-to-RGB color transform needed for JPEG source images (but not necessary in the TIFF RGB-to-sRGB case). Nevertheless, based on these results the transformation process can be considered mathematically lossless at best, while at worst, it was perceptually lossless.

Automated Migration

Eligibility

The majority of images in the DRS do not exist independently: for the most part, a single *intellectual* work is represented in the DRS by a set of related digital images. Within this set, most of the images can be generally classified into one of three broad functional roles:

- *Archival master* (AM) – an image optimized for the greatest range of potential outputs
- *Production master* (PM) – the result of further processing of an AM (or PM), e.g. crop, de-skew, sharpen, color correction
- *Delivery* – a (generally) lower-quality derivative use copy

In addition to the derivation relationships that link AM/PM to PM, and AM/PM to delivery images, additional dependency relationships may associate these images with calibration targets and color profiles.

For a given digital object comprised of multiple images, it is preferable on both operational and conceptual grounds to first migrate forward the PM to a new PM′ and then re-derive any required delivery images from PM′. In essence, this preserves the procedural relationship *between* the images, as well as the images themselves. Furthermore, the PM encapsulates significant manual assessment and processing not found in the AM. Thus, if the AM were migrated to a new AM′, the derivation of a new PM′ (necessitating manual intervention) would be prohibitively expensive. Migration eligibility is thus predicated on identifying the PM image. Unfortunately, the heterogeneity with respect to inter-image relationships in the DRS is significant (see Table 4; derivation is indicated by a greater-than sign (>); a forward slash (/) indicates sibling formats at the same derivation "level"). This heterogeneity complicates the selection of the appropriate source image from within a given family of images.

Table 4: Image Derivation Families

1	GIF							
2	JPEG							
3	JPEG	>	JPEG					
4	JPEG	>	JPEG	>	JPEG			
5	JPEG	>	TIFF	>	JPEG			
6	PhotoCD	>	PhotoCD	>	JPEG			
7	PhotoCD	>	TIFF	>	JPEG			
8	PhotoCD	>	TIFF	>	JPEG	>	JPEG	
9	TIFF							
10	TIFF	>	JPEG					
11	TIFF	>	JPEG	>	JPEG			
12	TIFF	>	TIFF					
13	TIFF	>	TIFF/JPEG					
14	TIFF	>	TIFF	>	JPEG			
15	TIFF	>	TIFF/JPEG	>	JPEG			
16	TIFF	>	TIFF	>	JPEG	>	JPEG	
17	TIFF	>	TIFF	>	TIFF/JPEG			

The initial approach was to simplify the selection by definition: within a given derivation hierarchy, the root image would always be considered the AM; the leaves, delivery copies; and intermediates, PM images. Whenever multiple PM images co-exist, TIFF would be preferred over JPEG and where PM images exist at two derivation levels (Categories 8 and 16), the upstream (high-density) TIFF would be preferred over the downstream (mid-density) JPEG.

However, consultation with collection managers indicated that this approach was too simplistic. In some instances of Category 14, for example, the intermediate TIFF is not fully processed (some color correction but no cropping or sharpening) and the leaf JPEG thus fills both the PM and delivery roles and is the proper source image for a migration. Unfortunately, functional tagging in the DRS has not always been applied uniformly over the past five years, so certain classes of objects will have to be evaluated on a collection-by-collection basis to ensure that the appropriate source images are identified.

Category 6 is problematic as an appropriate PhotoCD-to-JPEG 2000 codec has not been identified. It is hopeful that recent work in PhotoCD preservation will result in usable tools in the future [7]. For Categories 7 and 8 the intermediate TIFF does function as a PM and is appropriate as the migration source.

The majority of the independent TIFF images in category 9 are CCITT T.6 (Group 4)-compressed bi-tonal surrogates of printed or manuscript text pages. The curatorial demand for the enhanced user features enabled by the use of JPEG 2000, however, is focused on art images, not text. Thus, bi-tonal TIFF page images are considered out of scope for the initial phase of the project. HUL is evaluating the JPM profile for possible future use in representing bi-tonal page images using JBIG2 compression [8].

Source Image Classification

Consistent with the process developed during the AIHT project, the automated migration workflow requires the classification of source images based on format, color space, compression, number of components, and component bit depth in order to parameterize the codec properly. Color spaces need to be classified as either uncalibrated or calibrated (e.g. RGB vs. sRGB). The existing DRS metadata model does not track the existence of an external color profile as a metadata property, but only as a relationship; while the existence of an embedded profile is not tracked at all. It is therefore not always possible to determine an image's color space through a direct repository query. (These deficiencies will be rectified in a future enhancement round.) However, external profiles are retrievable through a procedural traversal of the relationship network, and the presence of internal profiles can be determined through invocation of JHOVE. Image classification is therefore performed as a batch pre-processing step.

Resultant Image Profiles

Implicit in the design of the DRS image migration service are the following four goals:

- Preserve the visual integrity of the source images
- Maximize the utility of the resultant images
- Optimize for rendering performance on the widest range of commercial and open source systems
- Maximize the homogeneity and sustainability of DRS content

Homogeneity is an important administrative concern of repository operation. Any further growth in the complexity of image derivation families beyond the 17 already extant is undesirable, if not unsupportable, over time. Sustainability is important for producing durable digital objects that are amenable to preservation efforts [9]. The use of an open, non-proprietary standard format is an important component of ongoing durability.

These high-level goals can be met by defining JPEG 2000 profiles that:

- Use lossless compression
- Enable the dynamic generation of derivatives of arbitrary size and sub-region
- Give preference to speed of decoding, rather than encoding
- Minimize the use of extension features
- Are self-contained and self-documenting

More specifically, the baseline specifications shared by these profiles are:

- Resolution-layer-component-position (RLCP) progression order
- 1024x1024 tile size
- Reversible (lossless) 5-3 wavelet transformation
- 1 quality layer
- *n* decomposition levels (dependent on image size)
- Reversible channel quantification
- Highest quality coding predictor offset
- Reversible component (de-correlation) transformation (RCT)
- Embedded Tile Length Marker (TLM) segments
- Embedded Packet Length, Tile-part Header Marker (PLT) segments
- No built-in error resilience
- Inclusion of optional Capture Resolution ('resc') box, if the data are available
- Inclusion of optional Intellectual Property ('jp2i') and XML ('xml ') boxes

Given an image's maximal pixel dimension p, the number of decomposition levels n is calculated as:

$$n = 1 \quad \text{, for } p \le 150 \tag{1}$$

$$n = \lceil \ln(p / 150) / \ln(2) \rceil \text{, for } p > 150 \tag{2}$$

Color component de-correlation permits somewhat higher compression ratios. The specification of TLM and PLT segments facilitates fast decoding. Internal error resilience is not necessary, as the assurance of bit fixity is a repository-level service already provided by the DRS. The Intellectual Property box is populated with a rights statement expressed in the JPX IPR schema. The XML box is populated with a technical characterization of the file and a persistent identifier pointing to appropriate descriptive metadata in external public discovery systems. A number of candidate schemas are being evaluated for expressing these properties, including the NISO Z39.87/MIX schema [10], the XMP EXIF schemas [11], and the schemas defined by JPX.

These baseline specifications are extended into four specific profiles:

- JP2 greyscale
- JP2 sRGB
- JP2 with embedded restricted color profile
- JPX with embedded unrestricted color profile

The first profile is used for single-component JPEG and TIFF source images; the second, for RGB or sRGB TIFF images; the third, for calibrated color TIFF images whose color space can be expressed in terms of the ICC Three-Component Matrix Based Input profile [12]; and the fourth, for color TIFF images whose color space cannot be expressed in those restricted terms. Note, however, that the JPX profile is included only to allow embedding of unrestricted color profiles; no other JPX extensions are used.

Quality Assurance

The automated JHOVE-based QA program will follow that established during the AIHT project. The source-to-resultant pixel comparison step that was manually invoked in AIHT should be susceptible to full automation. The commercial tool used for this purpose is scriptable, but is available only on Windows and Macintosh platforms. The DRS infrastructure, however, operates in a Unix/Linux environment. Due to the number and size of files involved in the migration, it is undesirable to require inter-platform file transfers. The construction of an equivalent QA application using available open source toolkits for JPEG, TIFF, and JPEG 2000 appears feasible. However, an in-depth investigation into the technical practicality of this step has not yet been initiated. In the event that this solution proves too difficult or expensive, HUL would revert to manual processing of a representative sample of the files.

Disposition

The existence of new JPEG 2000 PM images, from which arbitrary derivatives can be dynamically created at the point of request, obviates the need to retain pre-formed deliverables in the DRS, which will be deleted. In cases where source PM images exist at more than one derivation level, or in multiple formats at the same level, those PM images not used as the source of the migration will also be considered redundant and will be deleted. Note that with regard to the image derivation families enumerated in Table 4, this deletion policy will have the added benefit of reducing object complexity, since following the migration several of the categories will conflate together.

With regard to the versioning of the PM images, the current retention policy is that the DRS will always maintain the initial, the previous, and the current version of a given file. The original version is retained as documentation of the original deposit and as the possible subject of future digital archeology efforts. The previous version is retained in case the migration is determined to be flawed. Intermediate version are not retained as they are presumably obsolete and thus of limited value. Subject to constraints of available storage capacity, and curatorial willingness to pay for that storage, these intermediate may be discarded. This process that can be illustrated diagrammatically as:

$$1, n\text{-}1, n \;\rightarrow\; 1, [\, n\text{-}1, \,] \, n, n{+}1 \;\rightarrow\; 1, n, n{+}1$$

Thus, following a migration, the newly created PM version and the formerly current (now previous) version will be retained, while the formerly previous version will be considered redundant and potentially discardable. Note that while image *datastreams* will be physically deleted, complete *metadata* about those datastreams are always retained.

Provenance

The existing DRS metadata model provides minimal support for recording object event history. The PREMIS event model will be implemented at a future stage [13]. In terms of that model, the following properties should minimally be recorded:

- eventIdentifier
- eventType = "migration"
- eventDateTime
- eventDetail
- eventOutcomeInformation

The eventDetail property will include the file's source classification, codec version identification, and invocation parameters. The eventOutcomeInformation property will record the results of post-migration QA testing. In the interim before full PREMIS support is available in the DRS, this provenance information will be recorded in a structured manner in available free text fields.

Conclusions

An automated image migration service following the general direction outlined in this paper will be added to the DRS during the summer of 2006. While this activity arose through curatorially-defined changes to end-user requirements, rather than obsolescence, the process that is being developed is intended to be appropriate for migrations instigated by either situation. Based on the limited, though representative, experience of the AIHT project, HUL believes that the automated large-scale retrospective conversion of approximately 1 million existing digital images to JPEG 2000 form, meeting measurable quality standards, is achievable.

References

[1] Testbed Digitale Bewaring, *Migration: Context and Current Status* (2001).

[2] ISO/IEC 15444-1, *Information technology – JPEG 2000 image coding system – Part 1: Core coding system* (2000).

[3] Stephen Abrams, Stephen Chapman, Dale Flecker, Sue Kriegsman, Julian Marinus, Gary McGath, and Robin Wendler, "Harvard's Perspective on the Archive Ingest and Handling Test," *D-Lib Magazine*, 11 (2005).

[4] Adobe Systems Incorporated, *TIFF Revision 6.0* (1992).

[5] ISO/IEC 15444-2, *Information technology – JPEG 2000 image coding system: Extensions* (2004).

[6] ISO 3664, *Viewing conditions - Graphic technology and photography* (2000).

[7] Peter D. Burns, Thomas E. Madden, Edward J. Giorgianni, and Don Williams, "Migration of Photo CD Image Files," *Proc. Archiving 2005*, pp. 253-58. (2005).

[8] ISO/IEC 5444-6, *Information technology – JPEG 2000 coding system – Part 6: Compound image file format* (2003).

[9] Caroline Arms and Carl Fleischhauer, "Digital Formats: Factors for Sustainability, Functionality, and Quality," *Proc. Archiving 2005*, pp. 222-27 (2005).

[10] NISO Z39.87/AIIM 20, *Data Dictionary – Technical Metadata for Digital Still Images* (2002).

[11] Adobe Systems Incorporated, *XMP Specification* (2005).

[12] ICC.1:1998-09, *ICC Profile Format Specification* (1998).

[13] OCLC/RLG, *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group* (2005).

Author Biography

Stephen Abrams is the Digital Library Program Manager at the Harvard University Library, where he provides technical leadership for strategic planning and coordination of the Library's digital systems and projects. He was the project manager for the JHOVE format validation tool and the ISO project leader for the PDF/A standard, and is leading efforts to establish a Global Digital Format Registry (GDFR). Mr. Abrams is a member of ACM, ALA, ASIS&T, IEEE Computer Society, and LITA.