Keeping the Bits in Place: A Case Study of Raster Image Migration

Jacob Nadal
Craig Preservation Lab, Indiana University Libraries
Bloomington, Indiana, USA

Abstract

This case study shows that raster image data is inherently preservable, as a logical implication of content contained in these files. This presentation will address these issues through a report on a project to migrate images from Kodak's Photo CD technology to the now standard TIFF and the emerging standard JPEG 2000 format. Although Photo CD is not a truly obsolete file format, it has clearly become and also-ran in the file format race, and it is desirable to migrate these files to standard formats for use in image repository and delivery systems. Recommendations are made for insuring accurate migrations of image data.

The Portable Pixmap (PPM) format was used to facilitate analysis of the iamge files. PPM was chosen because it allows the RGB values from the source file to be directly re-encoded into ASCII, rendering the binary content of a raster image file into an ASCII text file that indicates the RGB values in series for each pixel in the image. The resulting files were compared using text and numeric processing tools. Further comparative information about the binary files was gathered from image analysis software and comparisons of derivative iamge data.

A secondary consideration addressed is the maintenance of image metadata across format migrations. Because of the different levels of support for metadata amongst the file formats considered, automated metadata capture and subsequent image processing may be greatly affected by migration. The means of retaining metadata during migration are considered, reflecting on the value of internally versus externally contained metadata.

Introduction

It is generally accepted that within the current horizons of digital library development, there will be a need to migrate data from one format to another, either to avoid outright obsolescence or to provide a consistent technological foundation for future development.[1] Although increasing levels of standards awareness and standards compliance has helped make the potential problem more manageable, numerous questions are making the rounds about the likely outcomes and feasibility of data migration projects.

Raster images and textual data account for a great share of the digital resources that libraries depend upon and of the digital objects they produce. In their simplest form, uncompressed raster images list the color values for each pixel in an image. This simple data should be transferable without loss across any number of formats, provided they all share a common understanding of how to represent color.

This study addresses raster image data in the form of Kodak Photo CD files, which were migrated to several different formats directly from a source file, as well as across generations of file formats. Each new version was compared to the source and across generations to determine where errors had developed in the migration process.

History of the Project

The Photo CD files used in this study were created in 1995, under the auspices of a Title VI grant that supported the reformatting of numerous items from Indiana University Libraries' African Collections. Although microfilm was the standard medium for reformatting at the time the grant was received, a set of African market posters from the collection was photographed and transferred to Kodak Photo CD, then emerging as a promising format for library imaging needs.

During the course of moving to the Indiana University Libraries new web system, we realized that the on-line gallery of African market posters had grown outdated, with images sized for an era of lower display resolution s and slower connections. Fortunately, the Kodak Photo CD containing the original images was still viable as both a digital format and a physical object. It was still desirable to not only create better images for on-line viewing, but to migrate the source images to Tagged Image File Format (TIFF), currently the standard file format for the Indiana University Libraries' imaging projects.

A gallery of the files discussed in this study are available on-line at the Craig Preservation Laboratory's website: http://www.libraries.iub.edu/craiglab/

Experimental Setup and Methodology

This study involved the use of four graphics files formats. The original files were stored as Kodak Photo CD (PCD). To integrate these into other Indiana University imaging efforts, it was necessary to migrate them to Tagged Image File Format (TIFF) images. In the interest of examining the potential for continued use of these files, they were also migrated to JPEG 2000, an emerging standard.[2]

The Photo CD format has a provision for scaling to different dimensions, from a base image of 768 x 512 pixels, up to a maximum of 2048 x 3072. Initial comparisons showed that as long as all migration tests used the same scaling factor, there was no further inconsistency observed based on pixel dimensions. Consequently, resolution was kept to base (768 x 512 pixels) for this trial, to limit processing time and facilitate comparisons.

Conversions were carried out in both an uncontrolled color space and using the KODAK Photo CD Color Negative profile, as recommended in Kodak Photo CD Technical Paper 043.[3] All results refer to the Color Negative profiled version.

Images were migrated directly from the original PCD to each of the destination formats (PPM, TIFF, and JPEG2000). A second migration was made from the second generation TIFF to a third generation JPEG 2000. All files were then converted to Portable Pixmap (PPM) format for comparison.

The conversions from PCD to TIFF and JPEG 2000 were carried out in three different applications: LemkeSoft Graphic Converter 4.8,[4] ImageMagick 6.1.9,[5] and Adobe Photoshop 7.1.[6]

Conversions to PPM were made using the netpbm tool to create ASCII encoded PPM files. The ASCII file allowed further comparison within the migration path of each application, and across applications, by providing a human readable file as well as numeric data that could be manipulated and compared in a variety of software.

The netpbm tool was used to read out RGB values from a binary file and re-encoded them to ASCII PPM files. Image comparisons between source and PPM files were conducted in each application to verify that color values were not changed in the PPM conversion.

To compare the resulting files across conversion tools and formats, the Unix bdiff utility and Graphic Converter's image comparison functions were also used to compare binary files. Further comparison involved creation of derivative images based on subtractive differences in RGB values from two source files. color counts, and histograms.

Results

Three migration tests were performed during this study. In the first, each application was used to generate a TIFF and JPEG 2000 file from the original PCD file. In the second, each application was used individually to make a series of migrations from PCD to TIFF to JPEG 2000. In the final test, the applications were used in combination to carry color information from PCD to TIFF to JPEG 2000, based on the information gained about their effectiveness in the preceding trials.

Direct Migration from Source

In the first trial, Graphic Converter, Photoshop and ImageMagick were used to generate a set of images in TIFF and JPEG 2000 from the original PCD image. The netpbm tool was used to convert these files to PPM to enable comparison.

Photoshop and ImageMagick produced identical RGB values between their own TIFF and JPEG 2000 files. While each application was faithful in its rendering of image data across formats, they did not provide an identical interpretation of the source file. These differences occurred in both uncontrolled and profiled color spaces.

Graphic Converter produced identical RGB values from the source PCD to TIFF. Graphic Converter does not implement the complete JPEG 2000 specification, however, providing no option for truly lossless compression. As a consequence, the JPEG 2000 files showed notable differences in their RGB content from the TIFF and PCD files. The differences were presumably decompression artifacts, and especially notable in areas of high detail. To the credit of the JPEG 2000 compression algorithms, however, differences were not significantly greater than any of the other discrepancies observed across the software.

Comparison of non-matching pixels across the applications revealed that differences were significant, from 2 to 22 degrees of difference in color value per pixel, per channel in the images inspected. Mismatches did not run consistently through the entire file, either, indicating that some color values were more faithfully converted than others.

Sample RGB differences across different applications.

	ImageMagick	PhotoShop	Graphic Converter	Difference (Max - Min)
R	83	88	77	11
G	68	70	64	6
B	57	57	56	1
R	100	103	90	13
G	84	85	76	9
B	76	76	72	4
R	130	134	115	19
G	110	110	98	12
B	98	98	89	9
R	145	149	127	22
G	120	120	106	14
B	107	107	97	10

A comparison of the binary files was performed by creating a new image file based on the differences in RGB values between each pixel in two images. Darker colors correspond to RGB values closer to 0, indicating lower degrees of difference between images, while lighter colors correspond to higher RBG values, indicating larger degrees of difference. The differences in RGB values for the images shown have doubled for viewability. The number of separate colors in the resulting comparison image is equal to number of non-matching colors between the images compared.

Figure 1 shows a grayscale reduction of the comparison file generated from the TIFF images generated by Graphic Converter and Photoshop. The color comparison image contained 9.051 separate colors, and the grayscale reduction shown contains 32 colors.

Figure 2 shows a grayscale reduction of the comparison image generated from the TIFF files produced by ImageMagick and Graphic Converter. The grayscale image shown contains 28 separate colors, although the original comparison image also contained 9,051 colors, equal to the result of Graphic Converter-Photoshop comparison. An inspection of histograms for these images showed the overall color profile to contain higher numbers of dark pixels than the Graphic Converter to Photoshop comparison, indicating a closer color match.

Figure 3 shows the comparison image generated from TIFF files created by ImageMagick and Photoshop. This image contained 3,761 colors, resulting in 19 colors in the grayscale image shown.

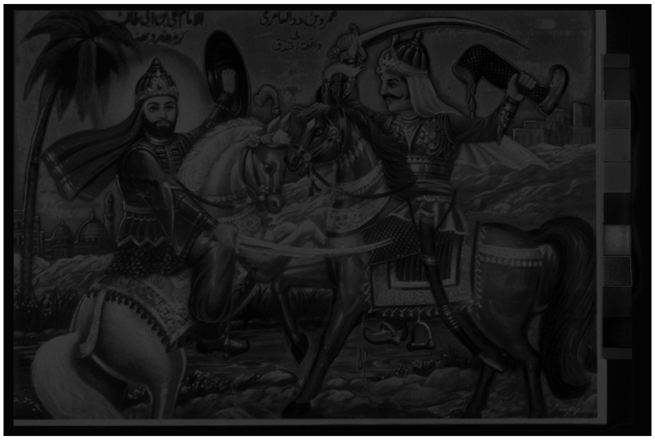

Figure 1. Comparison image from Graphic Converter and Photoshop (grayscale reduction to 32 colors)

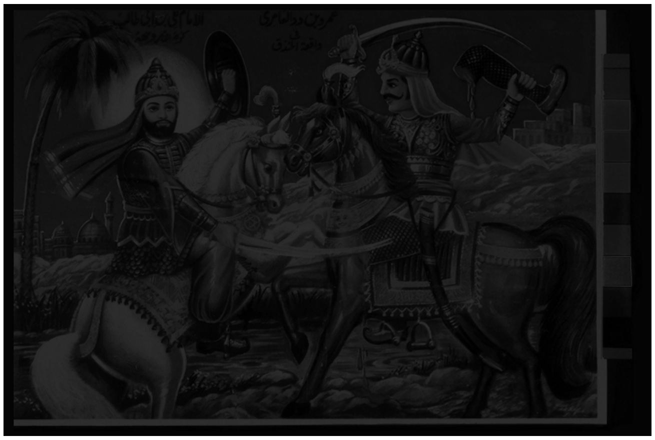

Figure 2. Comparison image from Graphic Converter and ImageMagick (grayscale reduction to 28 colors)

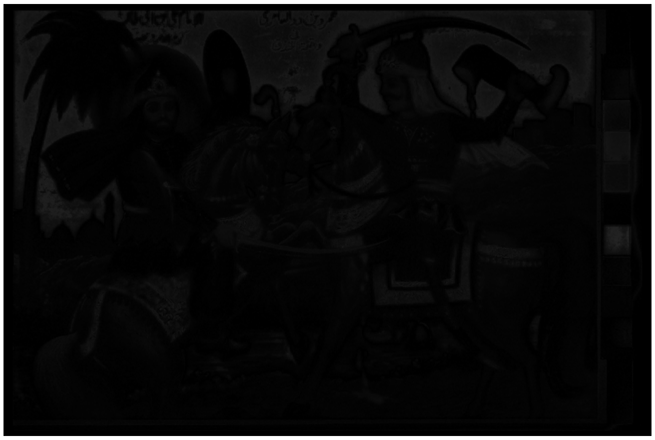

Figure 3. Comparison image from ImageMagick and Photoshop (grayscale reduction to 19 colors)

Secondary Migrations

In the second trial each application was used to migrate image data through a series of formats. The source PCD was migrated to TIFF, which was subsequently migrated to JPEG 2000.

ImageMagick and Photoshop both carried consistent RGB values across generations. Graphic Converter did not, due to the incomplete JPEG 2000 support, as noted above. The difference was limited to 42 colors, however, significantly smaller than the differences between software noted above.

In combining applications to carry color data forward, Graphic Converter was used to make an initial conversion from PCD to TIFF, based on its faithful rendering of this migration during the direct migration tests described above. Photoshop and ImageMagick were then used to make a conversion to JPEG 2000. The resulting files were compared and indicated identical RGB values.

Metadata

Inspection of the PPM files also showed mismatched image header data, copied from the metadata stored in the PCD files. TIFF and JPEG 2000 provide different levels of support for image metadata, and each application used for conversion has different facilities for maintaining this metadata as well. The PPM format allows commented lines which are used to varying degrees by conversion tools to store non-image data.

The Kodak Photo CD files provided for the following metadata elements in their native format:

- SCANTIME
- MODTIME
- MEDIATYPE
- SCANNERVENDOR
- SCANNERPRODID
- SCANNERFIRMREV
- SCANNERSERIAL
- PIW
- PHOTOFINISHER

None of the applications tested were able to completely carry this information forward, however, and in most cases the file formats provided no support for the metadata fields in Photo CD. By default, Adobe Photoshop and ImageMagick replaced all metadata with tags indicating the application version which was used, the date on which the file was last modified and the tags which were ignored in conversion. Photoshop, for example, replaced the PCD metadata with:

- SOFTWARE: Adobe Photoshop 7.0
- DATE: 2004:12:02 22:55:41
- Ignored Tags: $02BC, $A002, $A003

Graphic Converter preserved the PhotoCD Metadata as a comment in the TIFF file, but its incomplete support for JPEG 2000 led to loss of metadata on conversion.

Although it is desirable to package metadata within an image file to facilitate distribution of images and associated information, the varying levels of support and potential to lose metadata during image manipulation present obstacles to this.

It was also impossible to discern from a given file what its provenance might be. During the course of this study, we relied on a naming convention to keep track of the sequence of formats and software used to create them. File naming conventions, however, are subject to numerous possible human errors and lack the fixity that is desirable in an audit mechanism. A more formal and reliable means of recording and disseminating image provenance would be an essential component of any systematic migration plan.

Conclusions

Raster image data can be faithfully transferred across several generations of file formats. The discrepancies observed in the color interpretations between different applications indicate that long term image fidelity is dependent upon choosing an effective migration tool by carefully comparing source and migrated files for accuracy.

Unfortunately, no single application demonstrated the ability to consistently transfer color across all file formats. Graphic Converter performed best in the initial match of colors to the source file, but its incomplete implementation of the JPEG 2000 specification made it impossible to determine how successful this application would be in further migrations. ImageMagick and Photoshop were successful in performing a migration to

The value of open-source software should be emphasized as well. ImageMagick and the netpbm tools used in this study performed as well as the closed-source, and notably more expensive, commercial software. Although ImageMagick showed an initial mismatch in color from the PCD file, it was successful in faithfully transferring image data across subsequent formats.

The maintenance of metadata emerged as a central concern for image migration projects. The widely varying support provided by these tools made it a simple matter to lose metadata. It is advisable to replicate this information in an external system to insure that metadata is not lost during the migration process. Furthermore, none of the software tools or file formats provided the capacity to keep an audit trail of the various conversions made over the life of an image file.

Faithful migration of image data is possible to achieve, but the different color interpretations and levels of support for metadata provided by different software packages require careful evaluation of migration tools and suggest that external storage of metadata is necessary.

References

1. Paul Wheatley. “Migration - A CAMiLEON discussion paper. (2001). Ariadne, 29.
2. James Murray. Encyclopedia of Graphics File Formats. Sebastapol, CA: O'Reilly & Associates, 1996.
3. Eastman Kodak Co. Universal Film Terms for Reversal Films: Kodak Photo CD Information Bulletin PCD-043. Rochester, NY, 1996.
4. Graphic Converter: http://www.lemkesoft.com/en/graphcon.htm
5. ImageMagick: http://www.imagemagick.org/
6. Adobe Photoshop: http://www.adobe.com/products/ photoshop/
7. Netpbm: http://netpbm.sourceforge.net/

Biography

Jacob Nadal serves as the Head of the Craig Preservation Laboratory at the Indiana University Bloomington Libraries and as an Adjunct Lecturer in the Indiana University Bloomington School of Library and Information Science. He is responsible for managing the preservation activities for all classes of materials in the Indiana University Libraries, from rare books and manuscripts to electronic resources. He is an active teacher and consultant in areas related to the preservation of library collections and cultural heritage materials.