Optimized Still Image Batch Processing of Special Collections Bound Monographs and Manuscripts Using DNG, JPEG 2000, and Embedded XMP Metadata

Michael J. Bennett; University of Connecticut Libraries; Storrs, CT/USA

Abstract

Batch still image processing is examined in the context of operational bound monographs and manuscripts reformatting. The scaling of overall workflows through the flexible use of Lightroom, Photoshop, VueScan, and Jhove on parametrically-edited raw DNG and batch-rendered JPEG 2000 files is surveyed. Potential gains in processing efficiency, in comprehensive device data capture and preservation, in adaptable master image repurposing capabilities, and in the smoother growth of the required large-scale digital storage capacities that surround such operational conversions are considered.

Introduction

Digital still image capture of archives and special collections' objects has often followed a traditional uncompressed TIFF archival copy > compressed JPEG access copy processing chain for many reformatting projects. Though this has operated well enough in most cases, newer image formats and metadata wrappers along with more powerful tools centered on such advances have allowed for novel image utilization and the re-evaluation of overall workflow efficiencies. In an ever-expanding electronic environment, users are in search of richer digital content and have come to expect greater image quality for innovative manipulations and enhanced study. Within this ecosystem the obligations of content creators towards coherent production, storage, management, preservation, and more flexible and finely-tailored output of their own quickly growing digital archives and special collections have become magnified as a product of increasing overall scale. In turn, it naturally follows that novel value-added enhancements in workflow design, using the inherent capabilities of new still imaging formats, metadata specifications, and the latest developments in image editing software are engineered.

DNG as RAW Safety Master File Format

When looking at raw image formats as the starting point of an overall digital imaging chain a number of scalable advantages over traditional TIFF-based archiving and raster processing become apparent. Though these are outlined in narrative depth elsewhere [1][2][3][4][5][6] a look at the current capture workflow of monographs and manuscripts employed at the University of Connecticut (UConn) Libraries may be pertinent.

Bound Monograph Workflow: DNG from Camera Color Filter Array (CFA) [7][8] Sensor Data

In this example, illustrated in http://digitalcommons.uconn.edu/libr_pubs/43/, Figures 1-21, page images of John Donne's 1611 *Conclave Ignati* are used. Proprietary Canon .CR2 camera raw files are first converted into a folder of DNG safety masters, segregated into left and right page Adobe Lightroom 3 Collections by either verso or recto page origin, and then losslessly rotated and cropped through synchronized Lightroom parametric [9] edits. Such DNG raw editing, particularly across large, homogeneous image groups, saves substantial processing time, overall CPU overhead, and required storage space against comparable raster image batch editing steps which, unless accomplished as unmerged layered TIFF or PSD files, are irreversible in final form. Raw DNGs can be losslessly compressed, can retain originally-captured sensor data even when parametrically edited, and in fact can quite easily be reversed back to their original latent, unedited state. In this manner, the format can adroitly serve as both a robust master and efficiently processed format.

At UConn, bound monographs are captured on Atiz BookDrive book cradles outfitted with dual Canon 5D II DSLR full-frame sensor cameras that shoot 3:2 aspect ratio images. As a result in order to minimize cropping (and the loss of maximum sensor sampling rate), recto and verso pages are shot in "landscape" orientation. In turn, they require either 90° clockwise or 90° counter-clockwise rotation to bring page text back into proper "portrait" reading alignment. To best facilitate batch processing, then, left and right images are captured with _L and _R file name suffixes respectively through Atiz BookDrive Capture software. Lightroom can then easily filter by file name suffix and segregate images into left and right image collections where batch clockwise or counter-clockwise rotation and cropping steps can be parametrically run on the DNG files in a quick, lossless manner. See http://digitalcommons.uconn.edu/libr_pubs/43/, Figures 5-21.

Loose Manuscript Page Workflow: DNG from Scanner Trilinear Array Sensor Data

DNGs can also be created directly from scanners through the use of VueScan software. In this way a measure of parametric editing workflow and image format continuity can be coordinated among a conversion lab's given range of capture devices. As a result, aspects of batch parametric processing need not be completely re-written from scratch for each equipment type but can be re-purposed and shared among a broader spectrum of cameras and scanners.

It bears noting that as opposed to color filter array (CFA) sensor devices like the majority of today's digital cameras, common flatbed scanners employ a trilinear array of RGB-filtered CCD sensor elements [14]. In turn, unlike CFA-based camera DNGs which contain mosaic sensor data, native scanner DNGs are linear encoded RGB files at inception. Such linear (gamma 1.0) DNGs, however, still enjoy many of the same lossless parametric editing efficiencies as camera-based DNGs when manipulated in

tools like Lightroom, Adobe Camera Raw, Bibble, etc. In addition, VueScan's default uncompressed DNGs can also be losslessly compressed when subsequently batch processed through such tools or Adobe's DNG Converter. The resulting storage savings of losslessly compressed DNGs (see chart in next section) scale favorably in terms of high volume conversion projects. Also, planned project capture standards may more easily sway towards higher resolution and/or greater bit depth aims since such choices can be less dictated by the elevated storage costs of traditional uncompressed TIFF creation and be more focused on the overall goal of high-quality imaging.

As previously illustrated and in the demonstration outlined in http://digitalcommons.uconn.edu/libr_pubs/43/, Figures 22-23, DNG can be flexibly leveraged across a broad array of project and operational aims. In contrast to proprietary raw specifications, DNG's openly documented architecture uniquely allows the format to be coherently preserved and predictably re-used across platforms and applications. Through the utilization of parametric signposts like "Snapshots," a variety of edited "states" along with various software processing versions can begin to be managed consistently through time.

Lossless JPEG 2000 as Raster Archival Master File Format Alternative to TIFF

One of the simpler ways to begin to explore the advantages of JPEG 2000 is to consider its losslessly compressed use as an archival raster format substitute to uncompressed TIFF. On average, a given lossless encoded JPEG 2000 file will be 1/3 the size of the same image saved as uncompressed TIFF all without loss of any image information. When factored into a given institution's total number of archival image files, substantial, scalable data storage savings can be readily achieved.

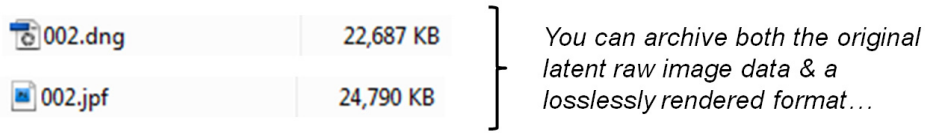
Lossless JPEG 2000 files can be batch-created directly from camera raw files or converted DNGs in the automated manner described in http://digitalcommons.uconn.edu/libr_pubs/43/, Figures 24-31.

The following illustration summarizes some of the scalable storage advantages of archiving both lossless JPEG 2000 [17] and raw DNGs for a given camera image vs. uncompressed TIF. By taking advantage of the lossless compression efficiencies of DNG and JPEG 2000, institutions not willing at this point in time to only save raw files can still reap the robust data preservation and processing gains of raw while maintaining the traditional benefits of rendered still image archiving. Significantly, this can all be achieved while taking up less storage space than a single uncompressed, rendered TIF.

If…

Name	Size	Date modified	Type	
002.CR2	26,644 KB	5/3/2011 2:11 PM	CR2 File	Camera raws
003.CR2	26,170 KB	5/3/2011 2:12 PM	CR2 File	
004.CR2	26,724 KB	5/3/2011 2:13 PM	CR2 File	
005.CR2	26,813 KB	5/3/2011 2:13 PM	CR2 File	
002.dng	22,687 KB	5/5/2011 2:15 PM	DNG File	DNG raws
003.dng	22,720 KB	5/5/2011 2:15 PM	DNG File	
004.dng	23,652 KB	5/5/2011 2:15 PM	DNG File	
005.dng	23,686 KB	5/5/2011 2:15 PM	DNG File	
002.jpf	24,790 KB	5/5/2011 2:22 PM	JPF File	Lossless JP2000
003.jpf	23,992 KB	5/5/2011 2:21 PM	JPF File	
004.jpf	21,699 KB	5/5/2011 2:21 PM	JPF File	
005.jpf	21,536 KB	5/5/2011 2:20 PM	JPF File	
002.tif	61,621 KB	5/5/2011 2:24 PM	TIF File	Uncompressed TIFF
003.tif	61,620 KB	5/5/2011 2:24 PM	TIF File	
004.tif	61,620 KB	5/5/2011 2:25 PM	TIF File	
005.tif	61,620 KB	5/5/2011 2:25 PM	TIF File	

Then…



…all while using less storage space than a single uncompressed TIFF

47,477KB (DNG + JPF) vs. 61,621KB (TIF)

Lossy JPEG 2000 Processed Master File Format

Through collaboration with software engineer, Hank Bromley, from the Internet Archive (IA) the author has tailored the UConn lab's monograph and manuscripts workflows to integrate with IA's batch ingest protocols. This has allowed the UConn Libraries' lab to function much like an IA scan center for online delivery of these material types. Part of this process is the creation of lossy (but visually lossless) JPEG 2000 processed master files, grouped into .tar files, one "tarball" of all page images per monograph volume. Lossy, irreversible JPEG 2000 is chosen because of its possible visually lossless compression and highly efficient storage savings that scale favorably across all aspects of the combined workflow (i.e. tarball upload, local and IA archiving, automated IA OCR, IA eBook format encodings, and interactive online "bookreader" interface generation). An example of the final results for one volume may be viewed at

http://www.archive.org/details/conclaveignati00donn.

DNG Safety Masters with "Processed_Master" Snapshots are the source for such rendered JPEG 2000 processed master images. The DNG Snapshots normally represent the source images parametrically rotated, cropped, with applied tonal adjustments best suited for high OCR success as described earlier. Lossy, but visually lossless, JPEG 2000s are then batch created along with embedded technical metadata through Photoshop from the DNGs. See http://digitalcommons.uconn.edu/libr_pubs/43/, Figures 32-35.

Leveraging Embedded Process Metadata in XMP

File-embedded XMP and its support for IPTC Core opens up new opportunities to create more robust still image files [18][19][20]. Such files can contain not only device-generated Exif information and parametric editing instruction tags (including Snapshots), but can also contain IPTC Core elements that can be edited either individually in Photoshop or in batches through Lightroom metadata presets and/or Adobe Bridge/Photoshop metadata templates.

The advantages of such additional embedded descriptive metadata are many. Individual still image files can be less dependent upon traditional external catalogs for their descriptions and can in essence be self-describing assets with sufficient descriptive information. This is of particular interest as images are exported and re-purposed beyond the institutional gates of their creation and become de-coupled from their original hosted settings.

Important file creation information or "process metadata" can also be efficiently embedded to include details of technical provenance and image editing [21]. Such particulars can greatly assist in future large-scale migrations and/or accurate file replications as hardware, workstation OS, and post-processing software versions change through time.

Finally, once embedded in all files, both descriptive and technical process metadata greatly assist in original digital asset management (DAM) system imports and/or future DAM platform migrations. As the vast majority of DAMs move toward fuller XMP compliance, catalog database migrations and their inherent problems may be made easier with more fully self-described source files that in essence become their own best record. Additionally, XMP is serialized in XML and stored using a subset of the W3C Resource Description Framework (RDF) [22]. As such, XMP's structure incorporates well when repurposed and leveraged through OAIS digital preservation technology stacks like Archivematica and repository frameworks such as Fedora.

Figures 36-38 from http://digitalcommons.uconn.edu/libr_pubs/43/ illustrate examples of how the UConn Libraries' lab has begun to embed and standardize such metadata into the various still image files examined throughout this study.

Conclusion

Today, recent developments in digital reformatting have included a growing movement toward making such conversions

more broadly operational, larger scale, and systemic [23][24][25]. Simultaneously, as the software and formats that surround still imaging evolve, a greater need for more robust and flexible digital objects is becoming apparent to meet novel repurposing needs [26][27]. In turn, decisions with regard to the scalable use of raw still image file archiving and processing, and data compression in general are important to consider when both quantity and quality are concurrent goals in today's reformatting ecosystem. Preserving the expertise of trained digital imaging technicians and the full sensitivities of the enlarging array of capture devices that they operate must be done now more than ever in both an efficient and extensible way to meet the requirements of feasible operational growth, new digital object use, and well managed storage over time. In so doing, institutions can more fully preserve and further utilize the fruits of their substantial investments in both digital conversion staff and equipment.

References

[1] Bennett, Michael J., and F. Barry Wheeler. "Raw as Archival Still Image Format: A Consideration." The Society for Imaging Science and Technology Archiving 2010 Final Program and Proceedings (2010): 185-193.

[2] Krogh, Peter. Non-Destructive Imaging: An Evolution of Rendering Technology, n.d. http://www.adobe.com/digitalimag/pdfs/non_destructive_imaging.pdf.

[3] Anderson, Richard. "Raw vs. Rendered | dpBestflow." dpBestflow, n.d. http://dpbestflow.org/node/264.

[4] Fraser, Bruce. "Understanding Digital Raw Capture", n.d. http://www.adobe.com/products/photoshop/pdfs/understanding_digitalrawcapture.pdf.

[5] Russotti, Patti, and Richard Anderson. Digital photography best practices and workflow handbook: a guide to staying ahead of the workflow curve. Burlington, MA: Focal Press, 2010.

[6] Krogh, Peter. The DAM book: digital asset management for photographers. 2nd ed. Beijing; Cambridge [Mass.]: O'Reilly, 2009.

[7] Allen, Elizabeth, and Sophie Triantaphillidou. The manual of photography. 10th ed. Oxford; Burlington, MA: Elsevier/Focal Press, 2011.

[8] Fraser, Bruce, and Jeff Schewe. Real world Camera Raw with Adobe Photoshop CS3. Berkeley, Calif.: Peachpit; London: Pearson Education [distributor], 2008.

[9] Krogh, Peter. Non-Destructive Imaging: An Evolution of Rendering Technology, n.d. http://www.adobe.com/digitalimag/pdfs/non_destructive_imaging.pdf.

[10] "Digital Collections | University of Connecticut Libraries: Standards", September 27, 2011.http:digitalcollections.uconn.edu/standards/standards.html.

[11] Friedl, Jeffrey. "Snapshotter Plugin | The Photo Geek." Snapshotter Plugin, The Photo Geek, n.d. http://thephotogeek.com/lightroom/snapshotter/.

[12] Arms, Caroline A., Carl Fleischhauer, and Jimi Jones. "Quality and Functionality Factors for Still Images", n.d. http://www.digitalpreservation.go/formats/content/still_quality.shtml.

[13] "ISO 22028-1:2004 - Photography and graphic technology -- Extended colour encodings for digital image storage, manipulation and interchange -- Part 1: Architecture and requirements", 2004. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37161.

[14] Peres, Michael R. The Focal encyclopedia of photography: digital imaging, theory and applications, history, and science. 4th ed. Amsterdam; Boston: Elsevier/Focal Press, 2007.

[15] Fraser, Bruce, and Jeff Schewe. Real world Camera Raw with Adobe Photoshop CS3. Berkeley, Calif.: Peachpit; London: Pearson Education [distributor], 2008.

[16] Bennett, Michael J. Jhove Audit Batch File, 2010.

[17] Lowe, David, and Michael J. Bennett. "A Status Report on JPEG 2000 Implementation for Still Images: The UConn Survey." The Society of Imaging Science and Technology Archiving 2009, Final Program and Proceedings (2009): 209-212.

[18] Adobe."XMP Specification Part 3, Storage in Files. July 2010." http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart3.pdf.

[19] Christensen, S. O., & Dunlop, D. "The Case for Implementing Core Descriptive Embedded Metadata at the Smithsonian." Proceedings of the International Conference on Dublin Core and Metadata Applications. (2010)

[20] Christensen, S. O., Dunlop, D., Pilsk, S., Snyder, R., Nguyen, D., Stauderman, S., Smith, S., et al. "Basic Guidelines for Minimal Descriptive Embedded Metadata in Digital Images." (2010) http://hdl.handle.net/10088/9719

[21] Metadata Working Group. "Guidlines for Handling Image Metadata, Version 2.0." http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf.

[22] TC 130. "ISO 16684-1:2012 - Graphic Technology -- Extensible Metadata Platform (XMP) Specification -- Part 1: Data Model, Serialization and Core Properties." http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57421.

[23] Erway, Ricky, and Jennifer Schaffner. Shifting Gears: Gearing Up to Get Into the Flow. Dublin, OH: OCLC, 2007. http://www.oclc.org/research/publications/library/2007/2007-02.pdf

[24] Smith, Stanley, Alan Newman, Chris Gallagher, Chris Edwards, and John ffrench. "Speed the Plow: Rapid Capture Digital Workflow Handout". Portland, OR, 2009. http://www.mcn.edu/conference/mcn2009/Smith_Getty_Rapid_Capture_Project_Photographs.pdf.

[25] Erway, Ricky. Rapid Capture: Faster Throughput in Digitization of Special Collections. Dublin, OH: OCLC, April 2011. http://www.oclc.org/research/publications/library/2011/2011-04.pdf.

[26] Mudge, Mark, Tom Malzbender, Carla Schroer, and Marlin Lum. "New Reflection Transformation Imaging Methods for Rock Art and Multiple-Viewpoint Display." Proceedings from the 7th VAST International Symposium on Virtual Reality, Archaeology and Cultural Heritage (2006).

[27] Mudge, M., Schroer, C., Earle, G., Martinez, K., Pagi, H., Toler-Franklin, C., Rusinkiewicz, S., et al. "Principles and Practices of Robust, Photography-based Digital Imaging Techniques for Museums." Proceedings from the 11th VAST International Symposium on Virtual reality, Archaeology and Cultural Heritage (2010).

Author Biography

Michael J. Bennett is Digital Projects Librarian & Institutional Repository Coordinator at the University of Connecticut. There he manages digital reformatting operations while overseeing the University's institutional repository. Previously he has served as project manager of Digital Treasures, a digital repository of the cultural history of Central and Western Massachusetts and as executive committee member for Massachusetts' Digital Commonwealth portal. He holds a BA from Connecticut College and an MLIS from the University of Rhode Island.