Page Image Compression for Mass Digitization

Stephen Chapman, Harvard University Library, USA; Laurent Duplouy, Bibliothèque nationale de France, France; John Kunze, California Digital Library, USA; Stuart Blair, Internet Archive, USA; Stephen Abrams, Harvard University Library, USA; Catherine Lupovici, Bibliothèque nationale de France, France; Ann Jensen, California Digital Library, USA; Dan Johnston, University of California at Berkeley, USA

Abstract

A number of the world's major libraries are embarking on large-scale projects to digitize books, journals, newspapers, and other printed materials. The main purpose of these mass digitization initiatives is to make published content visible to text indexing engines and accessible online for viewing and printing. The projects also present significant archiving challenges: capabilities must be developed within libraries to manage the hundreds of millions of files comprising their master volumes.

Even with the cheap disks and fast networks available today, projects of this scale must implement space-efficient imaging strategies to minimize long-term storage costs and maximize efficiencies for processing tasks such as file transfer, dynamic generation of deliverables, and migration. Libraries have had to accept that page image masters must be compressed, and that lossless compression of grayscale and color data will not achieve the efficiencies they are seeking in mass digitization.

In this paper, we present findings from studies coordinated by the California Digital Library, the Internet Archive, the Harvard University Library, and the Bibliothèque nationale de France to evaluate relationships between file size and perceived image quality for lossy compressed JPEG 2000 (JP2) images. We employed similar, but not identical, methods to create small test suites of source page images, which were then processed by four command-line JP2 codecs to produce images that observers rated from "perfect" to "unacceptable."

We present viable technical profiles for lossy JP2 encoding of page image masters, with recommended settings for selected command-line codecs. We are maintaining test suites of digitized book pages and invite others to use them to extend efforts to develop robust image processing algorithms that balance quality and file size in a variety of page image products.

Imaging Requirements in Mass Digitization

Under private-public and consortial partnerships, massive digitization efforts are ramping up in some of the world's largest libraries. [1] These industrial-scale "mass digitization" initiatives present management and technology challenges that raise familiar and new questions of tradeoffs between quality and production. [2]

At Archiving 2005, Phil Michel and Carl Fleischhauer presented a sound rationale for the benefits of high-resolution, high bit depth image capture in high-throughput digitization projects. By instituting a "fourth factor," *get more data*, in setting project requirements for reproduction quality, they reasoned that quality tolerances in digitization could be relaxed and that large master images would be amenable to supporting a wide range of current and future uses. [3]

Mass digitization efforts are driven by a combination of goals. Perhaps the most common goal is to increase access to materials without requiring physical proximity. An important component of increased access is the creation of searchable text that can be exploited by modern search engines to make discoverable that which was previously invisible to online search systems. The creation of digital surrogates also provides a backup copy to help safeguard against loss of the original. Participating institutions have not expressed a plan to replace books with digital files, but rather to augment access with digital files.

Thus, quality requirements for scans of text pages ("page images") in mass book digitization are largely being driven by a fifth factor, *get just enough data*. While page image masters in these projects would ideally have enough quality to serve the broadest possible range of uses over time, potentially including replacement of original printed volumes, not all uses can be accommodated. Three overarching functional requirements of mass text digitization favor the *get just enough* rationale over the *get more*. Digitization procedures in these projects must:

- enable very fast scanning of bound volumes (to reduce costs associated with human handling time);
- yield page image masters adequate for OCR and production of images with readable (legible) content when rendered as soft- and hard-copy outputs; and
- result in small file sizes, with the two-fold benefit of speeding up online transfer (ingest and access) and minimizing per unit (page/volume) storage costs.

Research Questions

In the fall of 2006, participants in three initiatives—Open Content Alliance, Google Library Project, and Gallica—were at similar points in defining technical profiles (of format, metadata, and codec settings) for page image masters. We agreed to undertake collaborative investigations so that we could evaluate a greater range of tools across a wider range of page types during the pre-production phases of our projects.

Key differences and similarities among these mass text digitization projects influenced the designs and starting points of our investigations into the use of lossy compression for page image masters.

The projects use different technologies for image capture, and the libraries (principals or partners) in each project exercise varying levels of control over image production. There is great heterogeneity in the populations of source material: in size and format (books versus newspapers), information content, and condition of bindings and pages. As evidenced in "traditional" library digitization projects, rendering objectives vary for historic text. Some imaging strategies are designed to render images that reproduce the item in hand, while others are organized to deliver uniformly "cleaned," high-contrast representations to monitors and printers.

We concurred that functional requirements for mass digitization dictate that all (library-copy) page image masters must be compressed. The prospect of saving uncompressed page images for our entire collections was out of the question; even at 60:1 compression, storage associated with mass digitization alone will exceed our current digital holdings by a factor of ten. Although a workable solution was in place for bitonal imaging, greyscale and color imaging presented challenges.

Motivated by the possibility of further reducing files sizes, each project is considering requiring two or more technical profiles for page images designed to optimize visual appearance while minimizing stored size. These include profiles for bitonal (1-bit), greyscale (8-bit), and color (24-bit) images. For some libraries, the best balance may be struck by composing a volume from a mixture of bitonal and multi-bit images. Others may elect to adopt a single 1-bit, 8-bit, or 24-bit specification for all page image masters. We have raised but not laid to rest concerns that the more divergent individual library's formats become, the more difficult it will likely be for our institutions to share usable content with each other.

For all cases in which bitonal imaging proves viable to meet image performance (OCR) and quality requirements, each project had specified the same technical profile for page image masters: 600 ppi TIFF bilevel (Class B) profile, lossless compression (CCITT T.6 aka "Group 4"). We have found that within any given volume, the file sizes from this profile average 105-120 KB per page, an average compression ratio of 187:1, depending upon the original volume's page size and density of printed information.

As for greyscale and color imaging, each project had targeted 300 ppi as a nominal (but not always minimal) sampling rate, and mandated adopting *some* type of lossy compression. Prior to undertaking the investigations described below, each project had produced JPEG images, at varying levels of compression, for evaluation. As would be expected in using JPEG for textual and other edge-defined content, we observed that benefits of meaningful reduction of file size come at the expense of visible image degradation.

From this point we investigated alternatives to JPEG with two operating premises. The first was our belief that, all other factors being equal in a given production imaging production workflow, lossy JPEG 2000 JP2 (ISO/IEC 15444-1) would outperform JPEG. Our second was that key librarian stakeholders in mass digitization projects would serve as appropriate subjects to evaluate image quality.

The main questions we wished to answer were, "For a variety of source page types, will JP2 produce better-than-JPEG quality with consistently smaller file sizes?" More importantly, we wished to understand, "What variables in JP2 production meaningfully contribute to perceived image quality of page images?"

A decision to select JP2 as an archival format would not be without risk. Each participating institution understands that JP2 is not renderable natively with contemporary web browsers. Moreover, the advanced JP2 compression algorithms that make JP2 so appealing have only recently been standardized and leave considerable room for codec- and vendor-specific interpretation. Nonetheless, faced with multi-million-dollar upfront storage costs, research libraries find the JP2 hypothesis compelling.

CDL/OCA Tests

The University of California Libraries and the California Digital Library (CDL) are principal players in the Open Content Alliance (OCA) and Google Library projects respectively. The University of California (UC), Berkeley Library is an early contributor of books to both initiatives.

Books in the OCA project are scanned with "Scribe" workstations built and operated by the Internet Archive (IA), also a founding member of the OCA. Google scans volumes for its Library project with an undisclosed technology.

CDL/OCA Round One

The CDL group assembled a sample group of 11 books representing many of the production techniques used in the late 19th and early 20th centuries, the period from which the majority of books would be selected for mass digitization. Text and illustration types are comparable to those evaluated in the Library of Congress Illustrated Book Study in 1999. [4] The sample books were scanned by the IA using their standard workflow at resolutions of either 500 or 300 ppi, and a subgroup of page image files, totaling 23 images from the 11 books, was selected for further study using different levels of JP2 compression.

Where the Illustrated Book Study required an observer to have the original book in hand to determine whether the digital reproduction fully or partially represented the full "structure" of the printing technology used (e.g., mezzotint, gravure), the evaluation methodology for the CDL test suite called upon observers to evaluate a series of page images relative to each other. The CDL methodology does not address the question of how the camera master images compare to the original print. It focuses instead upon identifying points where observers can detect visible artifacts of lossy compression, and when the artifacts become so pronounced that they marginalize the usability of the page images.

The JP2s were generated with LuraTech's LuraWave JP2 C-SDK toolkit (version 1.26, released 24 July 2005), and the following settings (for quality level Q50 in this case):

```
SetResolution(v-ppi, h-ppi)
SetProperty(File_Format = JP2)
SetProperty(image height in pixels)
SetProperty(image width in pixels)
SetProperty(bits per sample = 8)
SetProperty(colorspace = RGBa)
SetProperty(Rate_Quality = 50)
SetProperty(WaveletFilter = 9-7 filter)
SetProperty(QualityStyle = PSNR)
SetProperty(SpeedMode = Speed_Fast)
```

In the first of two test rounds, each reviewer evaluated five sets of six images on the LCD monitor he or she typically uses at work (library office or photography studio). Each set of images contained one baseline uncompressed TIFF RGB camera master and 5 lossy JP2 files made from the same TIFF master. Compression ratios in the first test group ranged from 4:1 to 340:1, and were chosen separately for each page image to give a wide range of visible compression effects for each page. [5]

Reviewers were drawn from organizations participating in mass digitization projects. Data were gathered from real-world rather than controlled settings: no effort was made to replicate

viewing environments or to assemble reviewers of comparable visual literacy. Reviewers compared images on screen and were asked to rate each JP2 rendering in one of four categories:

- **Perfect** – no discernable difference from baseline image;
- **Acceptable** – discernable, but not significant differences: image meets current and foreseeable needs;
- **Marginal** – discernable, relatively significant differences: image may not meet all current and foreseeable needs;
- **Unacceptable** – discernable, significant differences: image artifacts/degradation severely limit usability.

CDL encountered surprising consistency in reviewers' ratings of visible acceptability. Only in the case of the most heavily compressed test files did the number of sub-acceptable responses exceed the number of acceptable or better responses. Reviewers seemed generally tolerant of gradually diminishing quality up to the point where compression ratios reached 100:1 or greater. Files with compression rations of 25:1 or less were rated as "marginal" by 15% or fewer of the reviewers, and received no ratings of "unacceptable." Files in the range from 40:1 to 60:1 were mostly rated "acceptable", with sub-acceptable ratings from 30-40% of responses (mostly "marginal").

CDL/OCA Round Two

Shortly after CDL conducted this first test, IA used a subset of the test suite to create an on-line survey. [6] In this second round, reviewers were asked to examine a sequence of image pairs, rendered in the web browser, to rate the quality of each image using CDL's perfect/acceptable/marginal/unacceptable scale, and to select the version that "looks better or more readable to you." Each pair showed the same page image with different levels of compression. In many cases, image details were magnified (by 2x, 3x, 4x, or 5x).

In the second test, the effective compression ratios ranged from 20:1 to 78:1, but the compression ratio was the consequence rather than the objective of the encoding routine. Codec settings were used to aim for a consistent level of quality among a range of page types in a volume (or among millions of volumes) and to let file sizes fall where they may, not to compress image data to a target size and let quality fall where it may.

For both rounds, CDL and IA relied heavily on the Luratech codec's ability to compress files to a specified "quality level" instead of to a specified file size. When a quality target is specified, files sizes vary dramatically according to the complexity of the original image while the visual experience of the image remains at a roughly consistent level. Luratech quality settings range from a low of Q1 to a high of Q100.

In this second round, quality settings of Q70, Q50, and Q30, along with uncompressed TIFF, were compared. Consistent with the first round, IA's reviewers had a tolerance for highly compressed files, even as low as Q30. When compared with Q30, Q70 was preferred by a margin of 43% to 8% (33% expressed no preference). Preferences between Q30 and Q50 were expressed by nearly identical percentages. At the screen resolutions displayed in the survey, users were unable to discern quality differences between uncompressed and Q70, and a majority was indifferent to quality differences between uncompressed and Q50.

An additional test of OCR performed on these files showed slightly better OCR performance on compressed images over originals, peaking at about Q50.

CDL and IA have decided for now to select a variable quality measure, Q50, for all images from the OCA project, which corresponds to a mean page size of 620KB, for the entire 23-page suite of text and image pages, and about 500KB for the text-only pages. This will provide OCA partners with visually uncompromised images, good OCR, manageable processing and transfer sizes, and economical file sizes (in terms of disk capacity) across the spectrum of pages within the collection. Based on the Harvard/Google Partner tests, CDL and IA are now considering adding extra assistance to compression by recognizing pages without significant color content and storing them as greyscale instead of color.

These tests reveal that while judgments of quality do correspond to file size, only a professional digital photographer could detect artifacts of lossy JP2 compression at Q70, and users were generally very satisfied with Q50. These results correspond closely with the findings of the first round CDL/OCA survey.

Harvard/Google Partner Tests

The Harvard University Library is one of the first-round group of five partners in the Google Library project. Like OCA participants, the Google partners are collaborating to develop technical profiles for page images and metadata, envisioning that common specifications will yield short- and long-term benefits.

For its investigations, Harvard replicated the CDL strategy to produce a test suite. They selected five books from their collections and digitized 31 pages, again with the objective of identifying representative pages of the size, contrast, information content, and condition that would be encountered frequently in book digitization. David Remington, lead photographer in Harvard College Library Imaging Services, produced 300 ppi baseline RGB images with the production workflow that HCL has implemented for other book scanning projects. Volumes were scanned at a Zeutschel bookscanner and, as shown at the test suite site, camera masters were left uncorrected and batch-processed with color correction, tonal adjustment, and sharpening scripts to create "processed RGB" files. [7]

Harvard Round One

For the first round of evaluations, Harvard generated a large number of images to examine how variables among input source type and JP2 codecs and settings would affect quality. Three source types were used: (unprocessed) 300 ppi 24-bit RGB → (processed) 300 ppi 24-bit RGB → 300 ppi 8-bit greyscale. From each of these, two command-line codecs were used to create four lossy JP2 versions. Harvard used the Aware version 3.11.2 command-line codec to generate lossy JP2s at four quality levels specified in terms of Peak Signal to Noise Ratio (PSNR) of 45, 40, 35, and 30 dB (highest to lowest quality). PSNR $= 20 \cdot \log_{10}((2^b - 1)/\text{RMSE})$, where RMSE is the root mean square error between the uncompressed and compressed images. With the Kakadu version 5.2.2 command-line codec, Harvard produced lossy JP2s with a fixed-bit-rate bits-per-sample (BPS) formula at four quality levels: 5, 3, 0.5, and 0.25. BPS = *size*/(*height***width*).

None of the 15 reviewers in this round—from the OCA, Google Library, and Microsoft Live Books projects—evaluated all renderings of all pages. Each reviewed page image received an average of 2.97 ratings, with a low of 1, high of 6, and median of

3. In total, reviewers submitted 558 ratings, using the CDL perfect/acceptable/marginal/unacceptable scale.

The quality assessments in this round point to several key variables that bear upon perceived quality:

- **Color management** and other processing of camera-output data improve ratings. 60% of processed RGB images received scores of Acceptable or Perfect; 55% of unprocessed RGB achieved this standard. (Distribution =
  25% P, 35% A, 20% M, 20% U for processed RGB;
  15% P, 40% A, 25% M, 20% U for unprocessed RGB.)
- **Color JP2s** produce higher quality than greyscale, with no size penalty. (Distribution =
  25% P, 35% A, 20% M, 20% U for processed RGB;
  28% P, 27% A, 20% M, 25% U for unprocessed RGB.)
- **Choice of JP2 codec/compression method** (e.g., Kakadu BPS versus Aware PSNR) has a stronger bearing upon quality than effective compression ratio. In other words, quality per file size is codec/code-setting dependent. The smaller Kakadu-generated JP2s (avg. 275 KB) were generally rated higher than the Aware images (avg. 293 KB). (Distribution =
  20% P, 38% A, 24% M, 18% U for Kakadu;
  24% P, 32% A, 20% M, 25% U for Aware.)

Harvard Round Two

Applying the lessons learned from the first round, Harvard planned a second round that would refine their understanding of the impact of encoding methods upon quality and file size, as well as establish with greater precision the dividing points between unacceptable and usable (e.g., "marginal") quality for text and "non-text" pages. Evaluations from the initial Harvard round did a better job of distinguishing dividing lines between "perfect" and "acceptable" quality than at the lower-end of the scale, which is more pertinent to the *get just enough data* mandate of mass digitization. Therefore Harvard re-processed a subset of 11 pages (with text, image, and text+image content) from the test suite, using only the preferred processed RGB images as sources to generate two new sets of Kakadu-generated JP2 page images.

One set (4 versions per page) of Kakadu fixed bit rate images was again generated, but with much finer tuned settings for quality versions of 0.8, 0.6, 0.4, and 0.2 BPS. The second set was generated with variable-rate encoding, a method that minimizes overall distortion to a particular threshold value specified in terms of the slope of the distortion function. Slope values of 51492, 51748, 52004, and 52516 were used to produce the corresponding quality versions for evaluation.

The 26 reviewers in Round 2, who submitted 520 ratings for the 11-page test suite, preferred variable-rate encoding for 73% of the pages. (In this round, each reviewed page image received an average of 5.91 ratings, with a low of 1, high of 6, and median of 5.) Overall ratings were 12% P, 34% A, 26% M, 28% U. Tables 1-3 correlate file sizes, quality ratings, and content type for the variable-rate encodings—divided coarsely into categories of text and non-text.

Table 1. File Size Averages, by Quality Rating

	Text pages	Non-text pages
Perfect	232 KB	509 KB
Acceptable	236 KB	360 KB
Marginal	194 KB	268 KB
Unacceptable	123 KB	215 KB

Table 2. Text Pages, Quality Ratings, by Size

Avg. size	slope value	P	A	M	U
317 KB	51492	28%	65%	5%	2%
225 KB	51748	14%	49%	28%	9%
181 KB	52004	5%	19%	56%	20%
96 KB	52516	0%	5%	14%	81%

Table 3. Non-Text Pages, Quality Ratings, by Size

Avg. size	slope value	P	A	M	U
571 KB	51492	14%	73%	13%	0%
372 KB	51748	9%	54%	23%	14%
278 KB	52004	5%	18%	41%	36%
90 KB	52516	0%	5%	14%	81%

Summary findings from both rounds of the Harvard/Google Partners investigations may be generalized to the following recommendations:

- As a single method of encoding for mass digitization, variable-rate encoding performs better than other methods to balance perceived image quality and file size.
- To meet the *get enough data* mandate—i.e., to assume that a given technical profile for JP2 will consistently yield "marginal" or better-quality page images—assume that file sizes for text pages will average 181-225 KB, and that file sizes for non-text pages will average 268-372 KB. Further narrowing these tolerances depends upon how liberal or conservative space planning estimates must be for a given mass text digitization project. (*Caveat*: subsequent evaluations of the same variable-rate technical profile used for JP2 production from source page images generated in the Google workflow indicates the potential to reduce these sizes by as much as 20% if cropping and additional pre-processing steps are intelligently applied in the pre-compression workflow.)

JP2 Profiles for Mass Digitization

For now, the Google partners have agreed in principle to the following profile recommendations for use of lossy JP2 for mass digitization. The baseline component of the technical profile is both content- and codec independent:

- JPEG 2000 JP2 (ISO/IEC 15444-1) with lossy compression, sRGB or greyscale color space, and built-in error resilience.
- avoid use of fixed compression ratio or fixed output size methods to compress page image data
- use the RLCP (Resolution→Layer→Component→Position) progression order

These specifications are believed to optimize the images for the fastest decoding speed on the widest range of codecs.

Recommendations for tile size, quality layers and decomposition levels are, to a certain extent, delivery application dependent. Harvard's JP2 technical profile, for example, opts for 1024x1024 tiles, one quality layer, and decomposition levels based on the maximum pixel dimension to facilitate dynamic generation of arbitrarily sized and zoomed images. The number of levels m = ceiling(ln(p/150) / ln(2)), where p is the image's maximum pixel

dimension, p = max(*height*,*width*). Other settings are both content- and codec-dependent. For spatial resolution, we recommend 300 ppi, without upsampling in the image production workflow prior to applying compression.

The following settings for compression level should achieve "marginal" legibility. Within a given codec, one could adjust upward in quality and file size. To specify compression in terms of PSNR with the Aware codec the PSNR should not be lower than 35 dB. To specify compression in terms of variable bit-rate with the Kakadu codec, the distortion function slope value should be no lower than 52004 (yielding "marginal" quality images averaging 171 KB).

The Kakadu codec has options to minimize the MSE (mean square error) while giving equal weight to each of the RGB components, or using a "perceptual" weighting of these components. The results presented in Tables 2-3 are based on MSE weighting. However, user preferences for perceptual versus MSE weighting is fairly evenly split. To map between the two choices, the slope values in Table 4 should produce images of roughly equivalent size and quality.

Table 4. Slope Values for MSE and Perceptual Weighting

	MSE	Perceptual
Perfect	51492	52068
Acceptable	51748	52324
Marginal	52004	52580
Unacceptable	52516	52900

BnF Evaluation of JP2 for Newspaper Images

The Bibliothèque nationale de France (BnF) has been digitizing textual materials since 1992. Thirty million pages have been scanned to date, with 18.6 million to be available as full-text resources in the BnF's Gallica collections in mid-2007. Initially averaging 5,000-6,000 volumes per year, the library raised the annual production quota to 30,000 volumes in 2006 and 100,000 volumes targeted for 2007, 2008, and 2009.

BnF instituted a newspaper digitization program in 2005. Pages are digitized to produce 300ppi uncompressed 8-bit greyscale TIFF archival masters. (Given the 30 MB per page average file size, there is concern regarding how long this practice can be sustained.) These archival masters are processed to generate two outputs: a master page image optimized for Internet dissemination, and OCR-generated text encoded in ALTO (Analyzed Layout and Text Object) to map image and text coordinates. In addition, the TIFF image was envisioned to function as the master from which delivery versions could be generated dynamically for Internet display.

Gallica's users have low- to high-bandwidth connections. To accommodate these users and to manage storage costs, BnF sought an alternative to uncompressed TIFF as the delivery master. These large images performed too slowly with Gallica's applications, that work well for page images of books, to generate PDFs on the fly from greyscale TIFF masters.

BnF evaluated three options according to the following constraints:

- file sizes of entire pages could not exceed 500 KB;
- when rendered on screen, newspaper text must be readable (without perceptible degradation);
- the delivery format must be web browser compatible.

Results of JPEG, PDF MRC (Mixed Raster Content, bitonal TIFF CCITT T.6 layer for text), and JP2 encodings of pages that fall below the 500KB/page threshold are illustrated in Figures 1-3.

Figure 1. JPEG: unacceptable degradation for file size <500 KB.

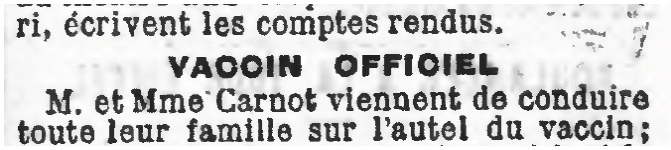

Figure 2. PDF MRC rendering: good compression ratio, but unacceptable degradation.

Figure 3. JP2: good compromise between size and quality.

BnF judged lossy JP2 as the best format to meet their size and quality requirements for renderings of full pages. Because online modes for newspaper delivery (e.g., article search and retrieval) deliver portions of page images, BnF's next question was to evaluate JP2 encoding options to tile the image.

Optimized Progression Order for Delivery

For newspaper pages, BnF concluded it is not possible to pre-determine whether the optimal progression order for interleaving should be based on quality or resolution. For example, for resolutions R1 and R2 and qualities Q1 and Q2, one may define two different progression orders:

(R1,Q1/R2, Q1/R1,Q2/R2,Q2) or
(R1,Q1/R1,Q2/R2,Q1/R2,Q2)

So, in the first case to reach the resolution R1 with the quality Q2, the packet R2, Q1 is also sent but not useful to the user. In the second case the useless packet R1,Q2 must be delivered to render an image of resolution R2 with the quality Q2.

Thus, to optimize performance, BnF elected to generate two JP2 files for each tile and to use a classical pyramid. With the JJ2000 codec, JP2s are constructed with each tile having a predefined definition square of 512x512 pixels, and the size of each tile resulting to approximately 50 KB. [8] Tiles of the best resolution layer have no reduction. Subsequent layers have a reduction of 4 in relation to the preceding layer. This process is run until an entire page image is completely contained in a square of 512x512 pixels. For example, a three-layer JP2 image will be

generated for a 5,285 by 7,009 pixel page image (corresponding to an A2 newspaper page scanned at 300 ppi). The JP2 file has 154 tiles for the best resolution, 12 tiles for resolution corresponding to a zoom of 25% and 1 tile for the entire image corresponding to a zoom of 6%.

BnF continues its evaluation of JP2, considering whether to institute a practice where a lossy compressed JP2 technical profile will be substituted for the current uncompressed TIFF profile. Storage costs will be considered, as will other libraries' adoption of JP2. If the BnF decides to adopt JP2 as the single format for page images of newspaper mass digitization, they will explore several TIFF-to-JP2 migration scenarios.

Conclusion

Technological and economic constraints effectively mandate that page image compression for mass book digitization must be lossy. Findings from the CDL/OCA, Harvard/Google, and BnF investigations indicate that JP2 codecs from three manufacturers can be used to generate page image masters of consistent usable quality at manageable file sizes.

Meaningful variables to maximizing quality for a given file size (or minimizing size for a given level of perceived quality) include selection of encoding method (PSNR and variable rate preferred for Aware and Kakadu respectively), selection of slope values, and choice of codec and version. Variables meaningful to processing and performance (e.g., decoding speed and use behaviors) include quality layers and decomposition levels. Recommendations for these may be application-dependent.

CDL, IA, and Harvard are maintaining test suites of images and making them available to the community for additional testing. We invite and welcome members of academe and industry to consult, or even "adopt," these suites and to augment the 2006 findings. We look forward to continuing investigations into the use of compression for mass digitization and the ways in which we can tailor masters to meet the widest possible range of uses while minimizing the size of massive datasets that will require periodic large-scale processing.

References

[1] Notable examples include: Google Books Library Project, http://books.google.com/googlebooks/library.html; Open Content Alliance, http://www.opencontentalliance.org/; Gallica, http://gallica.bnf.fr/; Internet Archive, http://www.archive.org/texts; Microsoft Live Books, http://books.live.com.

[2] Karen Coyle, Mass Digitization of Books (preprint available at http://www.kcoyle.net/jal-32-6.html), published in the Journal of Academic Librarianship, 32, 6 (2006).

[3] Phil Michel and Carl Fleischhauer, High Spatial and Bit Depth in Reformatting Projects: Supporting Varied Outputs and High Volume Throughput, Proc. Archiving 2005 (IS&T, Springfield, VA, 2005) pg. 101.

[4] Anne R. Kenney and Louis H. Sharpe II, Illustrated Book Study: Digital Conversion Requirements of Printed Illustrations. Report to the Library of Congress Preservation Directorate (1999).

[5] CDL Page Images Test Suite, available at http://www.cdlib.org/inside/massdig/cdl_survey/.

[6] Internet Archive Page Images Test Suite, available at http://www.cdlib.org/inside/massdig/ia_study/.

[7] Harvard University Library, Page Images Test Suite, available at http://preserve.harvard.edu/massdig/hul_study/.

[8] JJ2000 is the Java reference implementation (part 5) of JPEG 2000, available at http://jpeg2000.epfl.ch/.

Author Biography

Stephen Chapman is Preservation Librarian for Digital Initiatives in the Weissman Preservation Center, Harvard University Library (HUL). He advises the Harvard community about approaches to collections digitization, and is a member of the technical team managing the HUL Digital Repository Service. Stephen served as General Co-Chair for the Archiving 2006 Conference.

Laurent Duplouy is IT engineer for the IT department at the Bibliothèque nationale de France (BnF). He manages digitization workflows for Gallica and other projects, and contributes to the management of the library's OAIS preservation system. He previously served as the BnF digitization product manager for six years.

John Kunze is a Preservation Technologies Architect at the California Digital Library currently focusing on book digitizing and web archiving. He has created the workflow to scan, OCR, manually index, validate, and deliver an 8-million page tobacco document collection, as well as identifier tools such as the Archival Resource Key (ARK), Dublin Kernel metadata, and the NOID minter/resolver.

Stuart Blair was the Imaging Engineer for the Internet Archive's Scribe book scanning initiative, where he helped found the Open Content Alliance. He now advises Libraries Without Walls.

Stephen Abrams is the Digital Library Program Manager at the Harvard University Library, where he provides technical leadership for strategic planning and coordination for 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).

Catherine Lupovici is Digital Library Department director and preservation team manager at the Bibliothèque nationale de France where she provides strategy and coordinates digital library activities. She has many years experience in digitization projects.

Ann Jensen is formerly the Mathematics librarian at the University of California at Berkeley. Ann now works part-time for the California Digital Library, shepherding facets of UC's participation in the OCA project.

Dan Johnston heads the Digital Imaging Lab in the Preservation Department at the Library, University of California at Berkeley. He has worked on technical and production management of digital image capture for the Library's imaging projects since 1995.