XEROX® GLOBAL SERVICES

Technology | Document Management | Consulting Services

JPEG 2000 Profile for the National Digital Newspaper Program

Library of Congress
Office of Strategic Initiatives

Date: April 27, 2006

Prepared by: Robert Buckley, Research Fellow, Xerox Innovation Group 585.422.1282

Roger Sam, Business Development Executive, 202.962.7708

Real People. Real World Solutions.

Table of Contents

1 EXECUTIVE SUMMARY

This report documents the JPEG2000 file and codestream profiles for use in production masters during Phase 1 of the National Digital Newspaper Program (NDNP). NDNP, a joint collaborative program of the National Endowment for the Humanities and the Library of Congress, is intended to provide access to page images of historical American newspapers. Web access will be provided through the use of JPEG2000 production masters. For these masters, this report recommends using a visually lossless, tiled JPEG2000-compressed grayscale image, with multiple resolution levels and multiple quality layers, encapsulated in a JP2 file with Dublin Core-compliant metadata.

This report was prepared for the Office of Strategic Initiatives at the Library of Congress by Xerox Global Services, with inputs from the Imaging and Services Technology Center of the Xerox Innovation Group. A preliminary version of this report was available in April 2005 and its results were shared with awardees at a technical review meeting in May 2005. The profile described here has been in use since then.

Client Information

LOC Contact:

Name	Michael Stelmach
Company	Library of Congress
Address	101 Independence Ave SE, MS LA-G05 Washington, DC 20540-1310
Telephone	202-707-5726
Fax	
E-Mail	mste@loc.gov

Xerox Contact:

Name	Roger Sam
Company	Xerox Global Services
Address	1301 K Street NW Suite 200 Washington, DC 20005
Telephone	202-962-7708
Fax	202-962-7926
E-Mail	roger.sam@xerox.com

Change History

Version	*Date*	*Comment*
1.0	Jan 18, 2006	Version 1.0 of the report issued
2.0	Apr 27, 2006	Clarified format use in Section 3.1; added text on coding efficiency in Section 3.5; clarified the effect of code-block size on coding efficiency; added this Change History

2 INTRODUCTION

The National Digital Newspaper Program (NDNP) is a collaborative program of the National Endowment for the Humanities and the Library of Congress. It is intended to enhance access to historically significant American newspapers. NDNP will provide web access to a national directory of US newspaper holdings and to millions of page images.

According to the Technical Guidelines for Applicants [1], each newspaper page image will be supplied in two raster formats:

- An uncompressed grayscale TIFF 6.0 file, usually at 400 dpi
- A compressed JPEG2000 (JP2) file

JP2 is the image file format defined in Part 1 of the JPEG2000 standard [2]. A JP2 file contains a JPEG2000 codestream along with the image parameters, file information and metadata needed to interpret and render the JPEG2000-compressed image data. Appendix A contains an overview of the JP2 file format.

According to the Library of Congress and descriptions in the Technical Guidelines, the uncompressed TIFF file will be the master page image and the JP2 file will serve as the surrogate for day-to-day use and client access. The JP2 file will provide a high-quality, low-bandwidth, scalable version of the same image that is stored in uncompressed form in a large TIFF file. During the initial phase of NDNP, JP2 file access will use services created with a software development kit from Aware Inc.

Since the TIFF files are the archival masters, the JP2 production masters will be derived from them by means of a conversion process. This process will include image compression and file export using the compression options and file format parameters described in this report. The JP2 files that each awardee institution will provide should be compatible with the profile defined in this report.

This profile was derived with reference to TIFF file samples provided by the Library of Congress. The content and characteristics of typical files from this set are given in Appendix B.

The profile defined in this document covers:

1. Codestream: JPEG 2000 coding parameters, such as wavelet filter, number of decomposition levels and progression order. These are the parameters specified when applying JPEG 2000 compression to the image.
2. File: Image coding parameters, such as color space, spatial resolution and image size, which are independent of JPEG 2000. To the extent that the conversion will start with existing TIFF image files, these parameters in many cases are already defined and not design parameters
3. Metadata: Metadata, such as image description and keywords for search. These are based on current practices.

The next three sections describe these aspects of the NDNP JPEG2000 profile. They are followed by a section that describes the tools used in the development of the profile. Appendices A through F provide supporting information on the file format and sample images.

3 JPEG2000 CODESTREAM PARAMETERS

This profile defines a lossy compressed image with the goal of no objectionable visual artifacts and acceptable OCR performance, based on results obtained using the sample images. The JPEG2000 codestream in the production master shall contain a single-component 8-bit image with the same image size as the corresponding TIFF archival master. The compression ratio shall be eight to one. The codestream shall have 6 decomposition levels and multiple layers; in particular, this profile recommends 25 layers. The progression order shall be RLCP, which is resolution major. The codestream shall be tiled. The codestream does not use precincts or contain regions of interest.

Table 1 gives the codestream restrictions and characteristics for the JPEG2000 Production Master with reference to the markers defined in Annex A of Part 1 of the JPEG2000 standard [2]. This profile corresponds to Profile-1 of the JPEG2000 standard and would require a Cclass 2 decoder to process. The remaining parts of this section explain the rationale for the choices made.

Table 1: JPEG 2000 Production Master Codestream Profile

Parameter	Value
SIZ marker segment	
Profile	Rsiz=2 (Profile 1)
Image size	Same as TIFF master
Tiles	1024 x 1024 (Section 3.4)
Image and tile origin	XOsiz = YOsiz = XTOsiz = YTOsiz = 0
Number of components	Csiz = 1
Bit depth	Ssiz = 8
Subsampling	XRsiz = YRsiz = 1
Marker Locations	
COD, COC, QCD, QCC	Main header only
COD/COC marker segments	
Progression Order	RLCP
Number of decomposition levels	N_L = 6 (Section 3.2)
Number of layers	Multiple (Section 3.3)
Code-block size	xcb=ycb=6
Code-block style	SPcod, SPcoc = 0000 0000
Transformation	9-7 irreversible filter
Precinct size	Not used (Section 3.4)
Compressed File size	About one-eight of TIFF master or 1 bit per pixel (Section 3.1)

3.1 COMPRESSED FILE SIZE

One of the first choices to be made is how much compression should be used in the production master. Since the uncompressed TIFF file from the scan will be retained as the archival master during Phase 1 of the National Digital Newspaper Program, there is no a priori requirement for the JP2 production master to be lossless; visually lossless may be sufficient. Visually lossless means that while it is not possible to exactly reconstruct the original from the compressed image, the differences are either not noticeable, not significant or do not adversely affect the intended uses of the production master. In

this case, the intended uses of the production master are viewing, printing and possibly text recognition.

To judge the effect of compression on visual screen appearance, a series of images was generated by applying different compression ratios to selected test images. In particular, representative image (halftone), line art and text areas were selected from agreed upon test images. For each area, ten images were provided for viewing and evaluation: an uncompressed original and nine variants compressed to 2, 1.33, 1, 0.75, 0.67, 0.5, 0.4, 0.32 and 0.25 bits per pixel, corresponding to compression ratios of 4, 6, 8, 10.67, 12, 16, 20, 25 and 32 to 1[1]. Appendix C shows the images of the test areas.

The compressed image samples were delivered to the Library of Congress for their review and to establish quality thresholds in terms of their application. The 4:1 and 6:1 compressed images were judged visually lossless when viewed on a screen; only experienced viewers could locate compression artifacts in these images. The image quality exhibited by 8:1 and 10.67:1 was judged preferable. Image quality at 16:1 was acceptable, although it was felt that the artifacts and loss of resolution could make extended reading uncomfortable. Even the image quality at 32:1 was judged useable for some purposes.

The evaluations focused on text quality; halftone quality was not judged to be as important. Very little difference was noted in the printouts over the varying quality levels. The conclusion was that print quality was adequate, pending further analysis, but that it was less important than visual screen presentation quality.

As a result of these observations, the decision was made to use 8:1 compression in the profile for the JP2 production masters. This was judged a good compromise between file size and image quality. Since it was noted that higher compression ratios may be acceptable for some purposes, layers were introduced to make it possible to easily obtain images at a range of compression ratios between 8:1 and 32:1 (bit rates from 1 to 0.25 bits per pixel).

While screen viewing is the most important application, applying OCR to the production master may be an option. Some simple OCR studies were performed on text areas from the sample images. Differences were found when comparing the results of OCR applied to uncompressed and compressed images. In most cases, the OCR results from the 8:1 compressed images were *better* than were those obtained using the uncompressed images. The results are reported in Appendix D. More definitive conclusions require a study using a wider range of sample images and especially the same OCR tools as the Library of Congress uses or recommends.

3.2 Progression Order and Resolution Levels

The Library of Congress viewer was built assuming resolution-major progression. Further, it was assumed the codestream would be organized to make it easy to extract low resolution images that could be used for thumbnails or as a navigation pane. The two resolution-major progression orders defined in the JPEG2000 standard are RLCP (Resolution level-layer-component-position) progression and RPCL (Resolution level-position-component-layer) progression. Of these two, the profile specifies

[1] Compressing an area extracted from a page image to a target bit rate will not give the same results as compressing the page image to the same target bit rate and then extracting the area for evaluation. The differences however were found to be negligible at 8:1 compression for the selected areas and much less than the differences between the 8:1 and either the 6:1 or 10.67:1 compression ratio images.

RLCP progression so that within a resolution layer, the codestream is progressive by layer, i.e. by quality.

The number of resolution levels was selected so that the lowest resolution level gives a thumbnail of the desired size for a typical-size page image. This profile assumes that the lowest resolution level will generate QVGA-sized or smaller image: a QVGA or Quarter VGA image is 320 pixels wide by 240 pixels high. For a sample image that is 6306 by 8997, like the sample image in Appendix B.2, specifying five resolution levels means that the smallest resolution level image will be about 280 pixels high by 200 pixels wide. Because there can be page images larger than this sample, this profile specifies six resolution levels.

3.3 LAYERS

The Library of Congress observed that higher compression ratios may be acceptable for some purposes. Using layers makes it possible to obtain reduced-quality, full-resolution versions of the production master with compression ratios between 8:1 and 32:1, equivalent to bit rates between 1 and 0.25 bits per pixel. In particular, layers are introduced that correspond to compressed bit rates of 1, 0.84, 0.7, 0.6, 0.5, 0.4, 0.35, 0.3, 0.25 bits per pixel, which are equivalent to compression ratios of 8, 9.5, 11.4, 13.3, 16, 20, 22.9, 26.7 and 32.

At maximum quality and maximum resolution, all layers would be decompressed. However, at lower resolutions, higher compression ratios are possible without objectionable visual artifacts. Additional layers optionally allow higher compression ratios at lower resolutions.

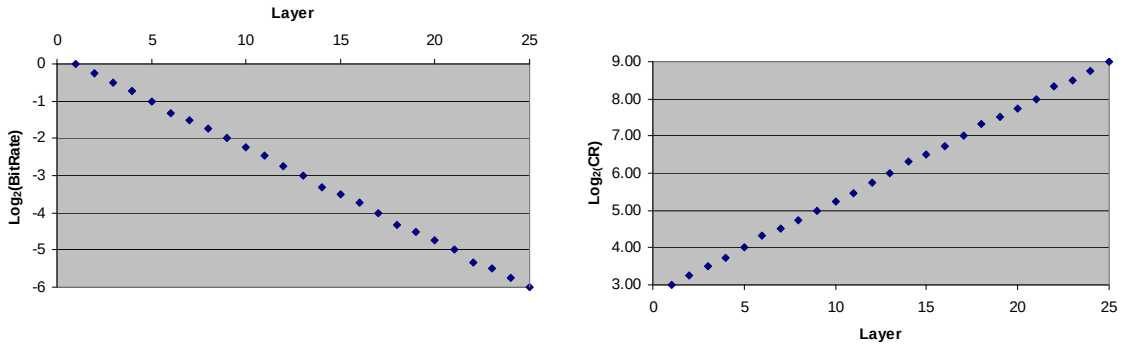
Altogether, this profile specifies 25 layers that cover the range from 1 to 0.015625 bits per pixel, or the equivalent compression ratio range of 8:1 to 512:1. The layers are specified in terms of bit rate; the bit rates and corresponding compression ratios (CR) are given in Table 2.

Table 2: Layer Definition

Layer	Bit rate	CR
1	1.0	8.0
2	0.84	9.5
3	0.7	11.4
4	0.6	13.3
5	0.5	16.0
6	0.4	20.0
7	0.35	22.9
8	0.3	26.7
9	0.25	32.0
10	0.21	38.1
11	0.18	44.4
12	0.15	53.3
13	0.125	64.0
14	0.1	80.0
15	0.088	90.9
16	0.075	106.7
17	0.0625	128.0
18	0.05	160.0

19	0.04419	181.0
20	0.03716	215.3
21	0.03125	256.0
22	0.025	320.0
23	0.0221	362.0
24	0.01858	430.6
25	0.015625	512.0

The bit rates for the layers were selected so that the logarithms of the bit rates (or the compression ratios) are close to being uniformly distributed between the maximum and minimum values. Figure 1 plots the bit rates and compression ratios in Table 2 against layers.



(a) (b)

Figure 1: Plots of (a) Log_2 of Bit Rate and (b) Log_2 of Compression Ratio (CR) vs. Layer

The exact bit rate values or compression ratios are not critical. What is more important is the range of values and there being sufficient values to provide an adequate sampling within the range.

Within a fixed bit budget, the more layers there are, the more bits are needed to signal the layer structure and the fewer bits are available to represent the compressed image data. As a result, the greater the difference between the original and the image decompressed from the compressed image. However, for the sample images in Appendix B, the differences between using 1 layer and using 25 were negligible. Compressing with 25 layers instead of 1 layer was about the same as using a compression ratio of a little less than 8.1 instead of a compression ratio of 8. This was judged a relatively small price to pay to obtain the advantages of quality scalability.

3.4 TILES AND PRECINCTS

Tiles and precincts are two ways of providing spatial addressability within the codestream. With spatial addressability, it is possible to extract and decompress a portion of the codestream, corresponding to a region in the image. In effect, this means an application can crop and decompress the codestream, which is more efficient than decompressing the codestream and then cropping out the desired region from the decompressed image. Along with resolution and quality scalability, spatial addressability is an important feature of JPEG2000.

Test images were generated with tiles and with precincts. In the images that had tiles, the tile size was 1024x1024. In the images that had precincts, the precinct size was 256x256 for the two highest resolution levels and 128x128 for the remaining levels.

In tests conducted by the Library of Congress, it was found that the Aware codec decoded images with tiles significantly faster than images with precincts. As a result, tiling was judged to be the preferred solution for decoding with Aware. Therefore this profile specifies the use of 1024x1024 tiles. The tile X and Y origins as well as the image X and Y origins are set to 0. The main header contains the Coding style default (COD), Coding style component (COC), Quantization default (QCD) and Quantization component (QCC) marker segments. Because these marker segments do not occur in the tile-part headers, the quantization and coding parameters are the same for all tiles. Also, the progression order is the same for all tiles.

3.5 CODING EFFICIENCY

Coding efficiency is a measure of the ability of a coder or coder option to compress an image. The more efficient a coder or option is, the smaller the file size for a given quality, or the higher the quality for a given file size. An objective measure of quality is PSNR, the peak signal-to-noise ratio, which is typically reported in dB[2]. Noise in this case is the error or difference between the original uncompressed image and the decompressed image.

Code Blocks: The quantized wavelet coefficients are coded independently by code block. The JPEG2000 standard limits the maximum size of a square code block to 64x64. There is nothing to recommend smaller code-block sizes, which are less efficient since smaller code blocks mean more overhead information in the file and less opportunity for the adaptive arithmetic coder to adapt to the statistics of the signal it is compressing. Therefore, this profile specifies 64x64 code blocks.

Bypass mode: The JPEG2000 coder uses an arithmetic coder that operates bit plane by bit plane, making multiple coding passes over each bit plane. For some images, the statistics of the least significant bit planes are such that there is little compression to be had with the arithmetic coder. In these cases, bypassing the arithmetic coder for some coding passes in less significant planes can speed up the coding (and decoding) with little loss in coding efficiency. However, for the sample images in Appendix B, the loss in efficiency was between 0.16 and 0.20 dB, which is up to twice the loss that comes from using 25 layers instead of 1. This profile does not specify the use of Bypass mode for coding, although the use of bypass mode can be revisited after more experience has been gained during the initial phases of NDNP.

Wavelet Filter: Part 1 of the JPEG2000 standard [2] defines two transformation types: a 9-7 irreversible filter and a 5-3 reversible filter. While the 5-3 filter is simpler to implement than the 9-7 filter, it is also less efficient. For the sample images in Appendix B, the loss in efficiency is a little over 1 dB. This was judged too high a cost in quality for only a slight gain in decoding speed. Therefore, this profile specifies the use of the 9-7 filter.

[2] PSNR in dB is 20 times the logarithm base 10 of the ratio of the peak signal (255 in this case) to the root mean squared error, where the error is the difference between the original and decompressed images.

4 JP2 FILE PARAMETERS

In a JP2 production master, the codestream specified in the previous section is embedded in a JP2 file. The JP2 file format is specified in Annex I of the JPEG2000 standard [2]. Appendix A contains an overview of the file format. A JP2 production master shall contain a JPEG2000 Signature box, a File Type box, a JP2 Header box and a Contiguous Codestream box. Table 3 gives the values for the data fields of these JP2 file boxes. The file shall also contain at least one metadata box, whose contents are described in the next section.

Table 3: JPEG 2000 Production Master File Profile	
JP2 Box/Data Field	**Value**
JPEG2000 Signature Box	<CR><LF><0x87><LF>[3]
File Type Box	
Brand	JP2
Version	0
Compatibility	'jp2 '
JP2 Header Box	
Image Header Box	
HEIGHT	TIFF ImageLength
WIDTH	TIFF ImageWidth
NC	1
BPC	Unsigned 8 bit
C	JPEG2000
UnkC	0
IPR	0
Colour Specification Box	
Method	Enumerated Colour Space (1) or Restricted ICC Profile (2)
PREC	0
Approx	0
EnumCS	17 (Greyscale) if Method = 1
Profile	Monochrome Input Profile if Method = 2
Resolution Box	
Capture Resolution Box	
Vertical	TIFF YResolution, converted to pixels/m
Horizontal	TIFF XResolution, converted to pixels/m
Contiguous Codestream Box	Codestream specified in Section 3

The JP2 production master shall contain an image with the same size and resolution as the image in the corresponding TIFF archival master. In particular, the JP2 production master file will be prepared after any image processing or clean-up and will correspond with the image that is used for OCR.

The image data in the JP2 production master is expected to have the same photometry (TIFF photometric interpretation) as the corresponding scanned TIFF file.

[3] In hexadecimal notation, the value of this field is 0x0D0A 870A.

To assess the photometry used in scanned images, the Library of Congress provided the image of a test target scanned on a microfilm scanner. A separate source provided the nominal densities of the patches on the target. The results of analyzing the scanned image of the test target are given in Appendix E. They show that the plot of target reflectance against pixel value is a straight line up to a pixel value of 250[4]. Encoding this relationship in a JP2 file requires using an ICC Monochrome Input Profile, as defined in Section 6.3.1.1 of the ICC Profile Format Specification [3].

The Monochrome Input Profile uses a one-dimensional lookup table to relate the input device values, in this case, the pixel values, to the luminance value Y of the XYZ Profile Connection Space. The JPEG2000 standard references the 1998 version of the ICC Profile Format Specification. There have been four major and minor revisions to the specification since, although there has been little change to the definition of the Monochrome Input Profile.

If gamma-corrected image data is available or can easily be generated, then a preferred method for representing the gray values in the JP2 production master is the enumerated grayscale color space defined in Annex I.5.3.3 of the JPEG2000 standard [2]. This grayscale color space is the gray analog of sRGB and codes the luminance using the same non-linearity as sRGB. The 8-bit pixel value D is converted to luminance Y using the following equations:

$$Y' = D / 255$$

$$\begin{aligned} Y &= Y' / 12.92 && Y' \le 0.04045 \\ &= ((Y' + 0.055) / 1.055)^{2.4} && Y' > 0.04045 \end{aligned}$$

These equations assume that a pixel value of 255 corresponds to a luminance value of 1.0, which is white. On this scale, a luminance value of 0.4 corresponds to a pixel value of about 170, which would provide a facsimile of the somewhat gray microfilmed image. If some scaling is used to make use of the full pixel value range, then the scaled signal would not follow the definition of the enumerated grayscale color space, but nevertheless could be represented by a Monochrome Input Profile.

5 METADATA

This profile requires that a JP2 production master contain metadata identifying the image content and provenance. This metadata would be in an XML box and would use Dublin Core elements to identify:

- File format
- Title of the newspaper
- Location of publication
- Date of Publication
- Page Label
- Description, including LCCN (Library of Congress Control Number)

[4] All the pixel values of the sample images in Appendix B are less than 250; in the case of the image in Appendix B.1, they are less than 240.

The template for the data field of the XML box with this metadata is shown below. This template was proposed by the Library of Congress for images scanned from microfilm. The application that writes the JP2 production master would supply the information between the sharp (#) characters in the template. An example of the use of this template is given in Appendix F.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdfsyntax-ns#">
  <rdf:Description rdf:about="urn:library-
ofcongress:ndnp:mets:newspaper:page://#The normalized LCCN#/#Date of publication
in CCYY-MM-DD#/#Edition order#/#Page sequence number#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
       <dc:format>image/jp2</dc:format>
       <dc:title>
          <rdf:Alt>
             <rdf:li xml:lang="en">#The title of the newspaper#.(#Location of
publication#) #Date of publication in CCYY-MM-DD# [p #page label#]</rdf:li>
          </rdf:Alt>
       </dc:title>
       <dc:description>
          <rdf:Alt>
             <rdf:li xml:lang="en">Page from #The title of the newspaper#
(newspaper). [See LCCN: #The normalized LCCN# for catalog record.]. Prepared on
behalf of #responsible organization#. </rdf:li>
          </rdf:Alt>
       </dc:description>
       <dc:date>
          <rdf:Seq>
            <rdf:li xml:lang="x-default">#Date of publication in CCYY-MM-DD#
            </rdf:li>
          </rdf:Seq>
       </dc:date>
       <dc:type>
          <rdf:Bag>
             <rdf:li xml:lang="en">text</rdf:li>
             <rdf:li xml:lang="en">newspaper</rdf:li>
          </rdf:Bag>
       </dc:type>
       <dc:identifier>
          <rdf:Alt>
              <rdf:li xml:lang="en">Reel number #The reel number#. Sequence
number #The sequence number#</rdf:li>
          </rdf:Alt>
       </dc:identifier>
  </rdf:Description>
</rdf:RDF>
```

It would be useful if a JP2 production master also contained a reference to the TIFF archival master from which it was derived.

This profile also recommends using technical elements from the NISO Z39.87 standard [4]. Besides the mandatory elements defined in that standard, the metadata should include JPEG2000-specific information. This metadata would be contained in an XML box using the MIX schema [5].

The current draft of the NISO Z39.87 standard defines a container with JPEG2000 format-specific data. The JPEG2000 information is comprised of two containers of data elements: CodecCompliance and EncodingOptions.

- CodecCompliance
  - codec: Specific software implementation of JPEG2000 used to compress the file or codestream
  - codecVersion: version of codec used
  - codestreamProfile: P1 (Profile 1)
  - complianceClass: C2 (Cclass 2)
- Encoding Options
  - tiles: 1024x1024
  - qualityLayers: 25
  - resolutionLevels: 6

At the time this was written, the draft NISO Z39.87 standard referred to here was in the process of being reballoted.

6 TOOLS

The JP2 files used to develop this profile were generated using Kakadu Version 4.2. While Kakadu was also used for decompression, some decompression and analysis were performed using LuraWave SmartDecompress Version 2.1.05.02.

The Kakadu command line that generates a JP2 file with the codestream described in Section 3 is:

```
kdu_compress -i in.pgm -o out.jp2 –rate
1,0.84,0.7,0.6,0.5,0.4,0.35,0.3,0.25,0.21,0.18,0.15,0.125,0.1,0.088,0.
075,0.0625,0.05,0.04419,0.03716,0.03125,0.025,0.0221,0.01858,0.015625
Clevels=6 Stiles={1024,1024} Corder=RLCP Cblk={64,64} Sprofile=1
```

APPENDIX A: JP2 FILE FORMAT OVERVIEW

The JP2 file format is defined in Annex I of Part 1 of the JPEG2000 Standard [2]. JP2 is a file format for encapsulating JPEG2000-compressed image data. Applications that can read the JP2 format and access the JPEG2000 codestream it contains are in a position to take advantage of the features and capabilities of JPEG2000, which enable progressive display, scalable rendering and "Just-In-Time" imaging.

A JP2 file consists of a series of elements called "boxes." Each box has 3 fields: a length field, a type field and a data field, which is interpreted according to the value of the type field. A JP2 file compatible with the profile defined in this report contains the following boxes:

- JPEG2000 Signature, which identifies the file as a member of the JPEG2000 file format family; it has a fixed-value data field.
- File Type, which identifies the file as a JP2 file and contains the version number along with any applicable compatibility and profile information.
- JP2 Header, which specifies image parameters, such as image size, bit depth, spatial resolution and color space. The JP2 Header box is a superbox, a box whose data field consists of the following boxes:
    - Image Header, which gives the height, width, number of components and bits per component of the image and identifies the compression type.
    - Colour Specification, which defines how an application should interpret the color space of the decompressed image.
    - Resolution, which it itself a superbox, whose data field contains the Capture Resolution box, which specifies the resolution at which the image was captured.
- Contiguous Codestream, which contains a single JPEG2000 codestream, compliant with Part 1 of the JPEG2000 standard
- XML box, which contains XML-encoded metadata

The structure and order of boxes in the JP2 file documented in this report is shown in Figure A-1.

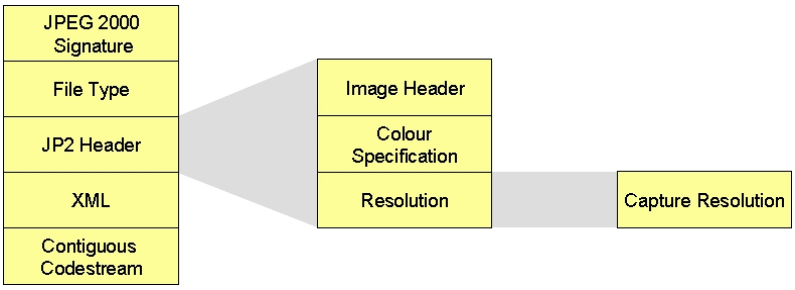


Figure A-1. JP2 File Structure

The JPEG2000 standard (Annex I.2.2 of [2]) requires that the JPEG 2000 Signature box be the first box in the JP2 file and that the File Type box immediately follow it. It also requires that the JP2 Header box come before the Contiguous Codestream box. This profile recommends that, for faster access to metadata, XML boxes come before the Contiguous Codestream box as well.

APPENDIX B: SAMPLE FILES

96 sample TIFF files were provided for testing purposes. They represented two sample sets of 24 newspaper pages each, scanned from microfilm by two vendors. From this image set, a subset was selected for developing the profile described in this report. Two representative images of this subset, one from each vendor, are shown here, along with the listings of the original TIFF files.

B1. SAMPLE FILE 1

Sample: \oclc\sn82015056\00000013.TIF

Tag	Value
41728 (10 ASCII)	microfilm
42016 (13 ASCII)	00000013.TIF
SubFileType (1 Long)	Zero
ImageWidth (1 Short)	5231
ImageLength (1 Short)	6861
BitsPerSample (1 Short)	8
Compression (1 Short)	Uncompressed
Photometric (1 Short)	MinIsBlack
DocumentName (27 ASCII)	Reel 00100493068
Make (9 ASCII)	NextScan
Model (8 ASCII)	Eclipse
StripOffsets (6861 Long)	8, 5239, 10470, 15701, 20932, 26163, 31394,...
Orientation (1 Long)	TopLeft
SamplesPerPixel (1 Short)	1
RowsPerStrip (1 Short)	1
StripByteCounts (6861 Long)	5231, 5231, 5231, 5231, 5231, 5231, 5231,...
XResolution (1 Rational)	300
YResolution (1 Rational)	300
PlanarConfig (1 Short)	Contig
ResolutionUnit (1 Short)	Inch
Software (20 ASCII)	Fusion Version 1.22
DateTime (20 ASCII)	2004:06:09 17:52:44
Artist (57 ASCII)	Library of Congress; OCLC Preservation Servic...

B2. SAMPLE FILE 2

Sample iarchives\sn82015056\00000003.tif

Tag	Value
ImageWidth (1 Long)	6306
ImageLength (1 Long)	8997
BitsPerSample (1 Short)	8
Compression (1 Short)	Uncompressed
Photometric (1 Short)	MinIsBlack
DocumentName (41 ASCII)	shington-evening-times-19050717-19050826
Make (9 ASCII)	NextScan
Model (24 ASCII)	Phoenix Rollfilm Type 2
StripOffsets (1125 Long)	9458, 59906, 110354, 160802, 211250, 261698,...
Orientation (1 Short)	TopLeft
SamplesPerPixel (1 Short)	1
RowsPerStrip (1 Long)	8
StripByteCounts (1125 Long)	50448, 50448, 50448, 50448, 50448, 50448,...
XResolution (1 Rational)	400
YResolution (1 Rational)	400
ResolutionUnit (1 Short)	Inch
Software (31 ASCII)	iArchives, Inc. imgPrep v3.001
DateTime (20 ASCII)	2004:10:12 11:18:33
Artist (16 ASCII)	iArchives, Inc.
41728 (10 ASCII)	microfilm
42016 (25 ASCII)	SN82015056/1905/00000003

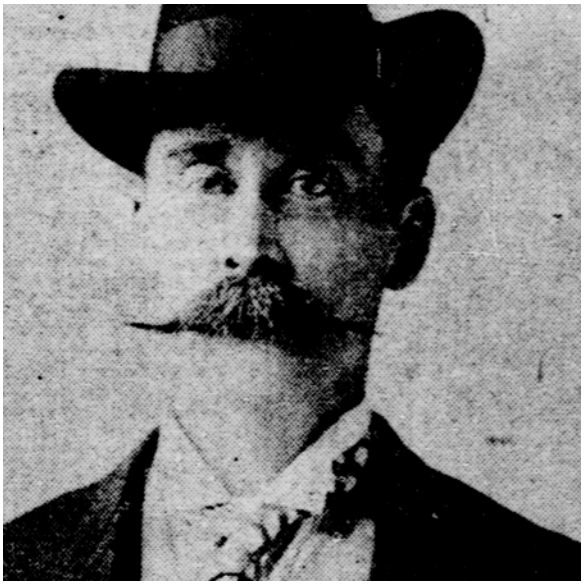
APPENDIX C: SAMPLE CONTENT AREAS

Quality assessments were made with respect to the image (halftone), line art and text areas, selected from the sample images in Appendix B and shown in this appendix.

Image (Halftone) Area

Source: iarchives\sn82015056\00000003.tif
Crop coordinates:

Left	3088
Top	752
Right	4112
Bottom	1776



Line art Area

Source: iarchives\sn82015056\00000003.tif
Crop coordinates:

Left	2656
Top	7408
Right	3680
Bottom	8432

Text Area

Source: \oclc\sn82015056\00000013.TIF
Crop coordinates:

Left	1047
Top	1329
Right	2391
Bottom	3378

It is built high above the cultivated land on either side and rises gradually upward till the fields give place to thickets of dwarf oaks and holly and thorn, with an undergrowth on the stony soil of bracken and bramble. Behind rolls a swelling stretch of weald, and out at one side, just below, lie the faint outlines, as shown by the height of the bushes, of a one time stupendous earthwork. Now it is a series of mere ridges, with dykes on the outer edges, reaching across the countryside; simply a long length of green mounds upon which penetrating roots and earth worms have persevered for centuries in their slow leveling process, to mark where once stood a famous Roman camp. But the white, straight road is as good as ever; it has outlived the camp, it will outlive many things of to-day—an eternal monument to the great road makers of history.

Down below on the lower modern road snorts a motor car; I can see it enveloped in a white cloud of dust; but up here I am alone with the shades of forgotten legions. There is no sense of solitude such as men experience in the unknown wilderness, but you find more a feeling of restfulness and satisfaction that you are following in the footsteps of men who, by building roads, shaped the destiny of nations.

Climbing steadily upward the edge of the ridge turns the square side of the camp and I wondered vaguely under what conditions those dykes were dug. Did our forebears hang onto the flanks of their conquerors, harassing them from every thicket of oak and thorn, or were they willing serving men? But they are dust again and the spring wind whispers of a new life as it rustles the dead fronds of the ruddy bracken, and there is the sharp blow of the axe on the green wood that tells of other life.

A little further on, in the midst of a clump of ash plants, I found it in the shape of an old hedger plying his bill hook with the vigorous skill that comes with experience and is not lost with years. I passed the time o' day

quired.

"No; I've heerd folks say as they never 'ud heer 'em any more. Us have it that Carter Maunders broke the spell. He wor plowin' back in a bit o' meadow land as lies on the edge o' the camp an' was goin' easy when his share hit up against sum't hard. The horses pulled in the collar a bit an' the plow went fair through it. When he looked down to see what it wor he'd plowed through a big brown pot cram full o' money, an' the coins lay all over the plow an' scattered 'round the furrows. He looked at it for a moment, an' he got that frit that he ran home. Ha, ha! Left his team standin' an' ran home, he did, an' put all the folks in a fluster by sayin' as how he'd plowed up pots o' money more'n he could lift. Us all went back wi' him, an' there they wor, thousands o' coins, little green bronze bits, all over the place. The pars'n took an' mended the pot an' sorted out the coins, an' now the squire has 'em. People as know say that they were the money o' the same folk as builded this gre't road, an' how mebbe the camp paymaster buried the money an' wor killed. Still it broke the spell, an' us'll never hear the clank o' armor and the marchin' any more. Young folks laugh at it, but there, they never believe owt but their own eyes an' ears."

He finished with an air of finality, a little contemptuous in his tone of the degeneracy of the rising generation.

Taking up his bill hook again, he moistened his palms in the primitive fashion of the laborer, and with the skill of an expert broke away the smooth barked staves crowned with the clustering buds. He cut carefully at their roots, saving the twists at the bottom for the handle crooks and leaving bare a hideous jagged stump.

"An ugly end for anything," I remarked.

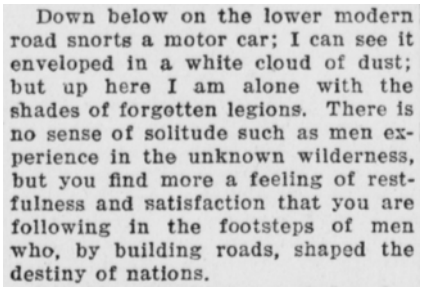
"Walkin' sticks for townsfolks," he snorted angrily, "to make 'em look

APPENDIX D: OCR RESULTS

D1. OCR SAMPLE 1

Source: \oclc\sn82015056\00000013.TIF
Crop coordinates:

Left	1050
Top	2200
Right	1710
Bottom	2640


Down below on the lower modern road snorts a motor car; I can see it enveloped in a white cloud of dust; but up here I am alone with the shades of forgotten legions. There is no sense of solitude such as men experience in the unknown wilderness, but you find more a feeling of restfulness and satisfaction that you are following in the footsteps of men who, by building roads, shaped the destiny of nations.

OCR results from uncompressed image[5]

Down below on the lower modern road snorts a motor car; I can see It enveloped <u>In</u> a white cloud of dust; but up here I am alone with the shades of forgotten legions. There Is no sense of solitude such as men experience In the unknown wilderness, but you find more a feeling of restfulness and satisfaction that you are following In the footsteps of men who, by building roads, shaped the destiny of nations.

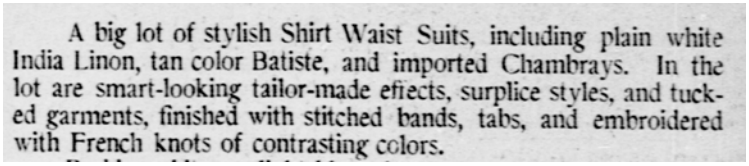
OCR results from 8:1 compressed image

Down below on the lower modern road snorts a motor car; I can see It enveloped in a white cloud of dust; but up here I am alone with the shades of forgotten legions. There Is no sense of solitude such as men experience In the unknown wilderness, but you find more a feeling of restfulness and satisfaction that you are following In the footsteps of men who, by building roads, shaped the destiny of nations.

D2. OCR SAMPLE 2

Source: iarchives\sn82015056\00000003.tif
Crop coordinates:

Left	4467
Top	5682
Right	6012
Bottom	6012


A big lot of stylish Shirt Waist Suits, including plain white India Linon, tan color Batiste, and imported Chambrays. In the lot are smart-looking tailor-made effects, surplice styles, and tucked garments, finished with stitched bands, tabs, and embroidered with French knots of contrasting colors.

OCR results from uncompressed image

A big lot of stylish Shirt <u>Vaist</u> Suits, <u>iivluding</u> plain white India Linon, tan color Batiste, <u>ani</u> imported Chamhravs. In th lot are smart-looking tailor-made efiects, surplice styles, and tucked garments. finished with stitched bands, tabs, and embroidered with French knots of contrasting colors.

OCR results from 8:1 compressed image

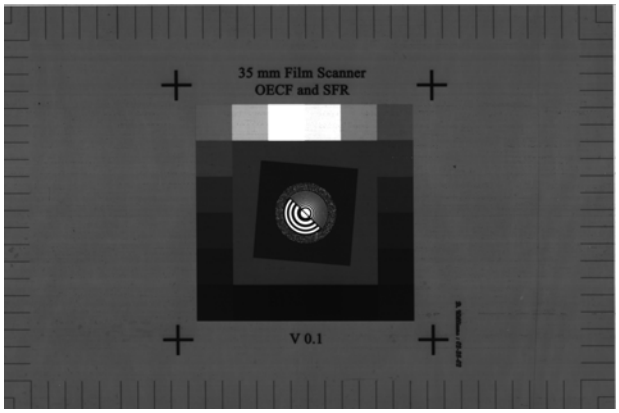
A big lot of stylish Shirt Waist Suits, including plain white India Linon, tan color Batiste, and imported Chamhravs. In th lot are smart-looking <u>railor</u>-<u>iiiade</u> efiects, surplice styles, and tucked garments, finished with stitched bands, tabs, and embroidered with French knots of contrasting colors.

The OCR results were obtained using Microsoft® Office Document Imaging Version 11.0.1897.

[5] Text results that are different from those obtained by OCR of the other image and that are incorrect are shown underlined.

APPENDIX E: TEST TARGET MEASUREMENTS

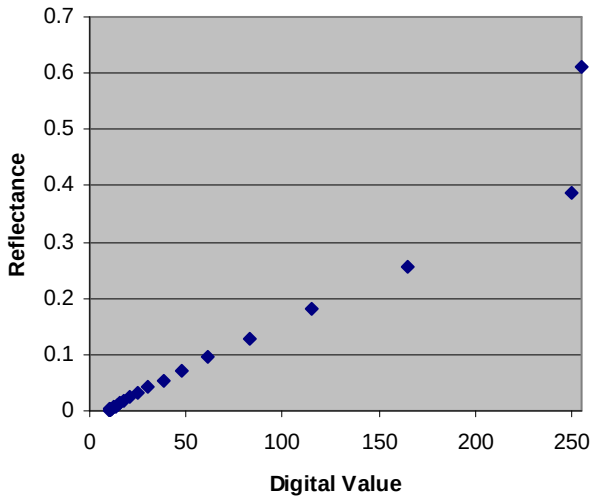
The test target image provided by the Library of Congress is shown below.



The following table shows the average digital values measured for the target's patches, the nominal Status A visual diffuse densities of the patches, and the corresponding reflectance values.

Digital Value	Density	Reflectance
255	0.214	0.611
250	0.411	0.388
165	0.591	0.256
115	0.74	0.182
83	0.888	0.129
61	1.023	0.0948
48	1.153	0.0703
38	1.265	0.0543
30	1.379	0.0418
25	1.506	0.0312
21	1.616	0.0242
18	1.734	0.0185
16	1.846	0.0143
14	1.961	0.0109
13	2.13	0.00741
12	2.24	0.00575
11	2.36	0.00437
10	2.535	0.00292
10	2.7	0.00200
10	2.87	0.00135

The following figure shows the plot of target reflectance against average digital value for the target's patches. The plot is a straight line up to a digital value of 250.



APPENDIX F: SAMPLE METADATA

Following is the data field of a metadata box based on the template in Section 4 for the sample image of Appendix B.1.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdfsyntax-ns#">
  <rdf:Description rdf:about="urn:library-
ofcongress:ndnp:mets:newspaper:page://sn82015056/1910-05-28/1/13"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
       <dc:format>image/jp2</dc:format>
       <dc:title>
          <rdf:Alt>
             <rdf:li xml:lang="en">The National Forum.(Washington, D.C.) 1910-
05-28 [p 13]</rdf:li>
          </rdf:Alt>
       </dc:title>
       <dc:description>
          <rdf:Alt>
             <rdf:li xml:lang="en">Page from The National Forum (newspaper).
[See LCCN: sn82015056 for catalog record.]. Prepared on behalf of Library of
Congress </rdf:li>
          </rdf:Alt>
       </dc:description>
       <dc:date>
          <rdf:Seq>
            <rdf:li xml:lang="x-default">1910-05-28
            </rdf:li>
          </rdf:Seq>
       </dc:date>
       <dc:type>
          <rdf:Bag>
             <rdf:li xml:lang="en">text</rdf:li>
             <rdf:li xml:lang="en">newspaper</rdf:li>
          </rdf:Bag>
       </dc:type>
       <dc:identifier>
          <rdf:Alt>
              <rdf:li xml:lang="en">Reel number 00100493068. Sequence number
1</rdf:li>
          </rdf:Alt>
       </dc:identifier>
  </rdf:Description>
</rdf:RDF>
```

BIBLIOGRAPHY

[1] The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants, Phase I, July 2004 *Note: This document was issued in connection with the Request for Proposals for the Phase I award competition, which is now closed. New guidelines are expected in Summer 2006.*

[2] ISO/IEC 15444-1:2004, *Information technology -- JPEG 2000 image coding system: Core coding system*

[3] International Color Consortium, Specification ICC.1:1998-09, File Format for Color Profiles

[4] National Information Standards Organization and AIIM International, Draft NISO Z39.87-2006/AIIM 20-2006, Data Dictionary – Technical Metadata for Digital Still Images, January 2006

[5] Library of Congress, Metadata for Images in XML Schema (MIX), http://www.loc.gov/standards/mix/, August 30, 2004