Federal Agencies
Digitization Guidelines Initiative

Raster Still Images for Digitization
A Comparison of File Formats

Part 3. Narrative and Summary Table

Revised, August 29, 2014

The FADGI Still Image Working Group

http://www.digitizationguidelines.gov/still-image/

Raster Still Images for the Digitization of Textual Materials
A Comparison of File Formats

Part 3. Narrative and Summary Table

Introduction: File Format Sub-Group

Since its inception, the Federal Agencies Digitization Guidelines Initiative (FADGI) Still Image Working Group's work has mainly focused on guidelines related to image quality (e.g., resolution, sharpening, color encoding). As a supplement to these guidelines, the group identified a need to develop a set of recommendations for file encoding recommendations for archival and derivative renditions of digitized content, as the selection of format directly affects an implementer's options in terms of compression, color encoding, and metadata support. Equally important are the costs associated with implementation, integration with workflows, and ongoing support.

Over time, a variety of organizations have adopted what might be called "de-facto standards" for file formats[1] for digitization output. While these de-facto guidelines have served the digitization community well in the past, the FADGI group has recognized the need to take a fresh look at this topic to ensure that recommended file formats for digitization that come out of the FADGI group are in line with current best practices, standards, and research.

The intent of this sub-group is to analyze and compare file formats and their associated characteristics or properties in terms of the various objectives and uses for digitized content. The analyses and recommendations from this group will provide input to the ongoing updating of FADGI's Technical Guidelines for Digitizing Cultural Heritage Materials as digital still images.[2]

The bulk of the work completed by the sub-group was accomplished by a core team of five, with representatives from the Library of Congress (LOC), Government Printing Office (GPO), and National Archives and Records Administration (NARA).

The sub-group has varying levels of confidence about its findings and hopes to benefit from the experience and wisdom of colleagues. We know that we are not alone in parsing this topic; members of the digital library community discuss the pros and cons of various still image target formats from time to time. This revision takes advantage of comments received after the

[1] This document takes a broad view of the term file format, adhering to the definition spelled out in the FADGI glossary, located at: www.digitizationguidelines.gov/term.php?term=fileformat. In part, this definition states that the term names a "set of structural conventions that define a wrapper, formatted data, and embedded metadata . . . . The wrapper component on its own is often colloquially called a file format. The formatted data may consist of one or more encoded binary bitstreams for such entities as images or waveforms, and/or textually-encoded data, often marked up with XML or HTML, for texts."

[2] See http://www.digitizationguidelines.gov/guidelines/digitize-technical.html.

initial posting on the FADGI Web site in April 2014. Additional comments are always welcome and interested parties are encouraged to use the FADGI comment page.[3]

Guiding Principles and Selection of File Formats

This sub-group did not seek to recommend a single format for all digitization and preservation master creation, but rather to characterize and compare a set of viable formats widely available in the current environment. The output of the sub-group is intended to provide a resource that can be used by federal agencies considering a digitization initiative to compare and contrast the various attributes, characteristics, advantages, and disadvantages of each format to assist in making decisions on formats to be used for preservation and access copies.

Although a wide variety of formats might be compared, the team analyzed a subset that represent formats commonly used in large scale digitization projects, as well as one or two others that are not so widely employed but warranted consideration. The following formats were selected for this comparison project:

1. TIFF. For many digitization projects, the TIFF wrapper with encodings that include uncompressed, LZW compressed, or bitonal-Group 4 compression, has been the format of choice for the cultural heritage community. Some added information in the section *Often preferred: uncompressed raster data in a TIFF wrapper*.
2. JPEG 2000. A newcomer in the field, offering lossless and lossy compression and thus yielding smaller files, warmly embraced by some and the subject of anxiety by others. Some added information in the section *JPEG 2000 and its adoption*.
3. PDF. A format that has been especially attractive in commercial circles, typically for new born digital creations, occasionally employed in reformatting projects.[4] Some related information in the section *PDF as a master when initial raster scans are not retained*.
4. PNG. A format especially designed for Web environments and infrequently used as a master format in digitization projects.[5]
5. JPEG. A format of long standing, used in most digital cameras, and very widely deployed for pictorial content. Rarely used for masters in digitization.

[3] http://www.digitizationguidelines.gov/contact/comments.php

[4] The group's analysis of PDF included consideration of PDF/A, the name for a set of PDF subtypes that have special features to support archiving and preservation. Features like the requirement for device-independent representation of color space make a good fit for raster images. However, features like the requirement that all fonts be embedded and the ban on JavaScripts have no impact on PDF as a carrier of bitmapped images. Overall, the group concluded that PDF/A did not confer any significant preservation benefit in our context and therefore we evaluated all types of PDF together.

[5] The W3C specification for PNG (http://www.w3.org/TR/PNG/) includes a number of features beyond those required for ease of use online. For example, the standard includes features that support color management, such as a group of metadata tags under the heading *Colour Space Information* that could document an image's primary chromaticities and white point, image gamma, and carry an embedded ICC profile. In addition, PNG offers lossless compression with excellent results. It is not clear to the compilers of this document, however, how well supported these features are in the PNG tools in the marketplace today. We would also be glad to hear from libraries or archives that use PNG as a mastering file.

As can be seen in the associated matrixes, these formats were also split up into sub-categories if there were significant distinguishing characteristics that could/should be pointed out about each version. For JPEG 2000, for example, the matrix's division into columns on JP2 (core encoding and basic wrapper) and JPX (extended encoding and wrapper) permitted reporting that JPX provides better support for geospatial metadata (potentially important for scanned maps) than JP2. For TIFF, to take another example, the team divided its report in order to highlight differences between the various encodings permitted within the TIFF wrapper, e.g., uncompressed and losslessly compressed, or difference of capacity or function, e.g., BigTIFF or GeoTIFF. Further subdivisions would be possible but the compilers felt that this would make the matrix excessively complex and inhibit comprehensibility.

Factors and terminology

The sub-group did not attempt to apply precise rankings for each factor. The rough yardsticks we employed are described by the questions and comments in the *Questions to Consider* column. The terminology employed for the sustainability factors has been taken from the Library of Congress format sustainability Web site.[6] Some of the terminology for the Settings and Capabilities factors has been taken from the quality and functionality factors provided at the same site.[7]

Often preferred: uncompressed raster data in a TIFF wrapper

Today, the most frequently used *encoding* employed by memory institutions is uncompressed, barely an encoding at all. With uncompressed data, the raster (aka *bitmapped*) data is stored in a straightforward manner, one sample point after another in a grid. Specialists call the sample points where the grid lines intersect *picture elements* or *pixels*.

The values stored in the file on a pixel-by-pixel basis may represent grayscale or color information in varying degrees of precision, depending on how many bits are allocated to each pixel. An uncompressed data structure has one powerful strength: it is relatively *transparent.* This pertains to the sustainability factor of *transparency*: it would not be difficult to build a tool to read the wrapper information and also unpack the rasterized data in order to present the image. To be sure, there is a correlative weakness: the lack of compression makes for big files.

Uncompressed TIFF files consume a lot of storage space, and each time one is summoned from storage, it takes a bit of time to read back from media and travel thru the network to a display device or printer. Although not extensively used at institutions like the Library of Congress, TIFF does support the use of the LZW compression algorithm,[8] which will generally cut the size of grayscale or color bitmap in half, with a corresponding decrease in transparency.

[6] Sustainability factors are discussed here: http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml.

[7] The still image quality and functionality terms are discussed here: http://www.digitalpreservation.gov/formats/content/still_quality.shtml.

[8] http://www.digitalpreservation.gov/formats/fdd/fdd000135.shtml

The TIFF wrapper specification was developed by the Aldus Corporation, with some Microsoft connections, in the 1980s, and moved to Adobe in the 1990s more or less when Adobe bought Aldus. The most recent complete specification, version 6, dates from 1992. It is a very open and well documented *industry standard*, i.e., not a capital-S standard from a Standards Developing Body like the International Organization for Standardization (ISO). Specialists react to the 1992 specification date in two ways: some saying that it is being left behind by changing times while others argue that its endurance is a strength, especially considering the wide array of applications that can read it. It is the case, however, that the application array is not as deep as one might wish: TIFF files cannot be read natively in most browsers (although there are several plug-ins). Apple's Safari is notable exception.

JPEG 2000 and its adoption

One of the motivations for this format comparison is an interest in the JPEG 2000 format as an option for archival master files.[9] This was the focus of FADGI's JPEG 2000 Summit[10] in 2011 and has been a topic for discussion ever since. Some federal agencies produce extensive numbers of digital images each year and seek ways to reduce the cost for digital storage and network support (the smaller relative size of a JPEG 2000 file supports this goal). Other agencies have arrangements with outside entities that yield hundreds of thousands of JPEG 2000 images for their collections: ought these be retained as delivered and, if so, what are the issues attendant to their long-term management?

JPEG 2000 files, when combined with server software in an access system, offer the ability to tile images and proved facile scaling ("zooming") for end-users. This functionality has made JPEG 2000 attractive as a derivative service format even in institutions that were not ready to embrace the format for archival masters. The Library of Congress, for example, has made extensive use of JPEG 2000 in its online access applications for maps and scanned newspaper pages. These are both large-raster content forms that benefit from JPEG 2000's capability to tile and scale the previously notes dependency upon a commercial server application that zooms and tiles the underlying data to meet the end-users requests, delivering the imagery to the browser as cropped-to-order "old" JPEG files. Meanwhile, the archival master files for the Library's maps and newspapers are uncompressed TIFFs.[11]

[9] See the FADGI glossary entries for archival master files (http://www.digitizationguidelines.gov/term.php?term=archivalmasterfile), production master file (http://www.digitizationguidelines.gov/term.php?term=productionmasterfile), and derivative file (http://www.digitizationguidelines.gov/term.php?term=derivativefile).

[10] http://www.digitizationguidelines.gov/resources/jpeg2000.html

[11] The current practices for still image reformatting may be contrasted with those for moving images. At the Library, for example, JPEG 2000 encoded picture wrapped in MXF files is employed as the archival master target format when reformatting videotapes. The video content is for the most part protected by copyright and access is limited to the Library's premises, where end-user delivery is provided by MPEG files produced at the same time as the MXF masters. Regarding the JPEG 2000 component, this application uses single-tile imagery and (thus far) has not taken advantage of scalability features. The embrace of JPEG 2000 is relatively widespread in moving image applications, none more widespread than the digital cinema standard for theatrical distribution.

The matrix highlights some other appealing features of JPEG 2000. Both the wrapper and the encodings are proper *capital-S standards* from the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). The family of JPEG 2000 standards includes three encodings, with the main *core* encoding understood to be free of patent issues. One key JPEG 2000 compression process employs wavelet transforms to provide a very clean image, even in a lossy mode. The encoding includes a number of "resiliency" features that add a bit of error-protection absent in most other encodings. The JPEG 2000 wrapper provides a bit more help with color documentation than TIFF, and it has a "box" that can carry XML-encoded metadata.

The matrix also highlights some features that have been received mixed reviews from some observers. Although there has moderate-to-wide adoption overall, adoption to date in the cultural heritage community has been moderate at best. The format has not been implemented in two applications that would trigger wider use: digital cameras (where "old" JPEG prevails) and browsers. (As noted above, TIFF has also received very poor support in browsers). Regarding the sustainability factor of *transparency*, the output from the compression algorithms yields data with reduced transparency (when compared to an uncompressed bitmap). However, JPEG 2000's loss of transparency is mitigated by a set of *resiliency* elements. Meanwhile, regarding the use of the features that support tiling and scaling, as well as related features including those called *quality layers* and *progression order*, some users have found that files created in different applications may not interoperate. Readers will also note that cost factors for JPEG 2000 yield a mixed outcome: implementation and tool costs tend to be higher that for other formats, while the smaller file sizes provided by compression often reduce storage costs.

PDF as a master when initial raster scans are not retained

The focus for this sub-group is on raster-scanned images, more or less as produced, and when the archiving organization intends to retain them for the long term. This practice makes a good fit for collection types where the bitmapped representation has high intrinsic value, for example pictorial and other graphic materials, manuscripts, and rare books. The comparisons in the matrix are framed against this focal category, and our analysis indicated that PDF does not make a perfect fit.

Some sub-group members, however, have participated in the discussions in their agencies in which an alternate PDF-based practice for classes of multi-page printed matter has been conceptually explored. These are classes of printed matter in which the bitmapped representations have moderate value and end-users requirements stress the importance of legible typography and its successful conversion via OCR. These classes include such items as contemporary foreign newspapers for which an institution may have very extensive holdings and where reduced digital storage costs would be significant.

The alternate practice sketched in the preceding paragraph could be accomplished in a workflow in which page-level raster scans are made and subsequently assembled into issue-by-issue (multi-page) PDF files. After the PDFs have been produced and evaluated, in order to reduce long-term storage requirements, the initial page scans would be discarded. This alternate practice came under discussion after the main round of work on this comparison project had been completed and, since still hypothetical, it has not yet been analyzed by this subgroup.

Sub-Group Deliverables: Summary Table and Detailed Matrix

Two tables represent the team's output. The summary table in this document presents key findings that have been extracted from the larger, detailed matrix. The detailed matrix compares the formats in terms of attributes that are important to consider when selecting a file format for digitization. These attributes are grouped into four main categories: Sustainability Factors, Cost Factors, System Implementation Factors, and Settings and Capabilities. The detailed matrix takes two forms: a large unified table (part 1 of this trio of documents) and the same data organized as multiple pages for ease of printing (part 2).

In the detailed matrix's analysis, the categories of Sustainability Factors; Cost Factors System Implementation Factors, and Settings and Capabilities are divided into a number of sub-categories; readers are encouraged to scroll down column A in the matrix the see the list. Since the nuanced meaning for each subcategory may not be obvious, sets of questions and/or scoring conventions are listed in column B. These indicate how each attribute was interpreted for each format and provide the convention used in scoring for purposes of comparison between formats. Additional detail and notes from the sub-group supporting a particular score are made in columns where appropriate.

Findings and Next Steps

The summary table presents the team's main findings. These can be further summarized as follows:

1. There is little variation between the formats studied on Sustainability Factors. All formats have viable sustainability.
2. Regarding Cost Factors:
    a. TIFF offers the advantage of low implementation cost, but cost for storage tends to be medium to high depending on level of compression. Larger file sizes usually require that derivative images be produced to support access, adding to the overall implementation costs.
    b. JPEG 2000 offers the advantage of low to medium storage and network costs due to the nature of compression offered by the format, but implementation cost tends to be medium to high due to the high cost of toolsets available and the need for further development of tools to meet implementation needs.
    c. JPEG and PNG offer the advantage of relatively low implementation and access cost, and low to medium storage and network costs.
    d. PDF offers low to medium implementation and storage cost, but is generally used as an access format, not for raster-image preservation. (PDF is widely encountered as a format for born digital works, not the subject of this study.)
3. Regarding System Implementation Factors:
    a. Some disadvantages of JPEG 2000 lie in this area. Limited tools are available, and the ones that are available are complex and often lack the ability to implement

advanced features. Files can have a complex structure and some organizations have encountered interoperability problems where "legal" files will not open correctly when tested in multiple software applications.

4. A wide variety of tools exist for TIFF, PNG, JPEG, and PDF. There is modest variation in settings and capabilities between formats as far as clarity, color maintenance, etc. However, JPEG's lossy compression often yields undesirable visual artifacts.

We hope that both the findings and the comparison matrix itself ("the factors") will be useful to our colleagues in the digitization and preservation fields. We ask our readers to send us suggestions and corrections so that we can improve the matrix and summary.

Meanwhile, as noted earlier, the Working Group continues to refine its general guideline for still image digitization (http://www.digitizationguidelines.gov/guidelines/digitize-technical.html), and the findings from this format-comparison activity will inform that process.

Summary Table: Raster Still Images for Digitization: A Comparison of File Formats

Attribute Category	TIFF	JPEG 2000	JPEG	PNG	PDF
Sustainability Factors	-High level of sustainability related to disclosure, adoption, migration, and transparency. -Acceptable self documentation, offers less capability than other formats for entering metadata, embedded metadata limited to header tags.	-Good disclosure, core encoding widely adopted, acceptable transparency and migration -Robust resiliency -Good self-documentation, metadata entry and embedding capabilities -Possible patent impact for JPX (coding extensions)	-Good disclosure and migration, widely adopted, acceptable transparency -Self documentation acceptable: native metadata is only technical, descriptive requires XMP -Ubiquitous	-Good disclosure and migration, widely adopted, acceptable transparency -Self documentation good, can use XMP, no native support for EXIF	-Good disclosure and migration, widely adopted, acceptable transparency -Self documentation acceptable -Good embedded and native embedded metadata capabilities
Cost Factors	-Low implementation cost, cost of software and equipment needed is low. -High storage cost for uncompressed images, medium storage cost for compressed. -Not supported in most browsers for access	-Initial implementation cost medium-high due to cost of best toolsets available -Low to medium storage and network costs -Not supported in most browsers for access	-Low implementation cost -Low-medium storage and network cost -Low cost of providing access	-Low implementation cost -Medium storage and network cost -Low cost of providing access	-Initial implementation cost medium due to cost of best toolsets available -Low to medium storage & network cost with compression -Generally used as an access format, not for preservation
System Implementation Factors (Full Lifecycle)	-Low complexity -Wide availability of tools -Good compatibility, ease and accuracy of validation	-Medium-high in both technical and toolset complexity -limited tool availability -low compatibility	-Low complexity -Wide availability of tools -Good compatibility, ease and accuracy of validation	-Low complexity -Wide availability of tools -Good ease and accuracy of validation -Compatibility uncertain	-Medium complexity -Wide availability of tools -Good compatibility, ease and accuracy of validation
Settings and Capabilities	-Good on clarity, multi-page capability. -Acceptable on color maintenance - Searchable Text Embedding not natively supported	-Good on clarity, color maintenance -Multi-page capability and searchable text embedding not supported	-Clarity is good, but slightly less than other formats -Acceptable on color maintenance -Multi-page capability and searchable text embedding not supported	-Good on clarity and color maintenance -Multi-page capability and searchable text embedding not supported	-Clarity potentially good, but default settings generally yield reduced clarity -Acceptable on color maintenance, multi-page capability and searchable text embedding not supported