Federal Agencies
Digitization Guidelines Initiative

Raster Still Images for Digitization
A Comparison of File Formats

Part 3. Narrative and Summary Table

April 17, 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 need has been identified to develop a set of recommendations for file encoding standards 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 standards 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).

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

[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.

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.
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.
3. PDF. A format that has been especially attractive in commercial circles, typically for new born digital creations, occasionally employed in reformatting projects.[3]
4. PNG. A format especially designed for Web environments and infrequently used as a master format in digitization projects.
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.

As can be seen in the attached matrixes, these formats were also split up into sub-categories if there were 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.

One of the motivations for this format comparison is an interest in the JPEG 2000 format as an option for archival master files.[4] This was the focus of FADGI's JPEG 2000 Summit[5] 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. 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

[3] 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.

[4] 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).

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

so, what are the issues attendant to their long-term management? Although not conclusive, representatives from those agencies were reassured to see that JPEG 2000 remains a plausible option in this comparison project.

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 preservation.
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 yield 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.	-Initial implementation cost medium-high due to cost of best toolsets available -Low to medium storage and network costs -Not supported in all browsers for access (requires added S/W layer)	-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