Still Image File Format Comparison

Document I. Narrative Introduction

Version of July 30, 2013
For review by the FADGI Still Image Working Group

Background. The two FADGI Working Groups are exploring file formats for still images and video. The explorations are using similar, matrix-based tools to make comparisons relevant to preservation planning. The matrixes compare a limited number of formats in terms of roughly forty factors, grouped under the following general headings:

- Sustainability Factors
- Cost Factors
- System Implementation Factors (Full Lifecycle)
- Settings and Capabilities (Quality and Functionality Factors)

The still image effort is led by the Government Printing Office and it is comparing formats suitable for reformatting (digitization). The formats being compared include JPEG 2000, JPEG (DCT), TIFF, PNG, and PDF, and several subtypes. The findings from this project will be integrated into the Working Group's continuing refinement of its general guideline for raster imaging.

Document I - Narrative Introduction - Table of Contents

Other documents in the set

Document II. Summary Table

Document III. Detailed Matrix

FADGI Still Image Group: File Format Sub-Group Narrative Summary

Introduction: File Format Sub-Group

Since its inception, the 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.

Readers should note that 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."

Over time, a variety of organizations have adopted what might be called "de-facto standards" for file formats 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 develop guidelines for file formats and associated characteristics or properties for the various objectives and uses for digitized content. The guidelines will be developed through an evidence-based methodology. As noted below, the object is not to identify the *one* format that fills all purposes, but to indicate which formats should be considered for a given project or organizational workflow. The recommendations from this group will also result in candidate language and materials that will update existing FADGI Still Image documentation (located at: http://www.digitizationguidelines.gov/guidelines/digitize-technical.html).

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 specific 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. A proven warhorse.
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.
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 (e.g., JPEG 2000 was split up into JP2 and JPX) if there were distinguishing characteristics that could/should be pointed out about each version. In some cases, e.g., TIFF, the splitting permitted the team to highlight differences in encodings within the wrapper (uncompressed, lossless compression) or difference of capacity or function (GeoTIFF, BigTIFF)..

Sub-Group Deliverables: Summary Table and Detailed Matrix

Two tables represent the team's output. A simple overview is provided in the Summary Table. It rolls up the findings from the detailed matrix in summary form, providing a sketch of the findings for the high-level categories that are analyzed in more detail in the matrix.

The second table is the detailed matrix that attempts to compare each format in detail relative to a set of attributes that could be deemed important when considering a file format for digitization. These attributes are grouped into three main categories: Sustainability Factors, Cost Factors, System Implementation Factors, and Settings and Capabilities.

These three factors are broken down 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 tend to be medium to high depending on level of compression.
    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 offers 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 lies in this area. Limited tools are available, and the ones that are available are complex and often lack the ability to implement advanced features
4. A wide variety of tools exist for TIFF, PNG, JPEG, and PDF. There is little variation in settings and capabilities between formats as far as clarity, color maintenance, etc.

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. We are circulating this draft to the FADGI Still Image Working Group. Once we have received the Working Group's comments, we will revise the document and place it on the FADGI Web site for open public review. We anticipate that this will also lead to some revisions.

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.