Life Beyond Uncompressed TIFF: Alternative File Formats for the Storage of Master Image Files

Robèrt Gillesse, Judith Rog and Astrid Verheusen; National Library of the Netherlands/Koninklijke Bibliotheek; The Netherlands

Abstract

The TIFF uncompressed file format is a widely accepted standard for storing master images of digitisation projects. As in the Koninklijke Bibliotheek, The National Library of the Netherlands (KB) the scale of these projects is rising rapidly, the need for an alternative, compressed file format is felt. This paper contains a summary of a research project in which four alternative file formats - JPEG 2000 Part 1 lossless and lossy, PNG, JPEG and TIFF LZW - have been described and tested. Four consequences of a choice for either of the formats have been described: consequences for storage, image quality, long term durability and functionality. In the final recommendations these consequences were weighed against three reasons the KB distinguishes for wanting to store master images:

1. *Substitution (JPEG 2000 Part 1 lossless or PNG)*
2. *Redigitalisation is no option (Visual lossless JPEG 2000 Part 1 lossy or JPEG)*
3. *Master file is the basis for access (JPEG and JPEG 2000 lossy with higher degrees of compression)*

Introduction/Background

The storage of image master files for digitisation projects as uncompressed TIFFs has become a good custom, but also a truism, and even a cliché. As mass digitisation projects are taking off in the Koninklijke Bibliotheek, The National Library of the Netherlands (KB), a need for a new strategy for storage of images has emerged. Millions of high resolution, RGB master image files will be made and will have to be (permanently) stored. If all these images – 40 million in total – are stored as uncompressed TIFFs, the estimate is that the KB will need no less than 650 TB of storage space by 2011. A capacity challenge indeed; and one that can only be mastered with a sound storage strategy.

Currently, the main question the KB faces revolves around *necessity*: do all master image files really need to be stored as uncompressed TIFF files and added to our long term repository, the e-Depot? Can we not make a distinction between digitisation projects in which images have to be stored for "eternity" (as originals are swiftly decaying and/or costs of digitisation have been so high that scanning anew is not an option) and projects which focus on access, thus allowing more pragmatic and economical storage choices?

In her IS&T Archiving 2007 presentation, author Judith Rog reevaluated the possibility of using compression on master images files. She concluded that in certain cases, compression might be used, although further research and testing was needed [1].

In October 2007, Judith Rog and Robèrt Gillesse of the KB began a research project focusing on finding alternative file formats for the storage of master image files. This has resulted in the final report *Alternative File Formats for Storing Master Images of Digitisation Projects* in March 2008 [2]. An earlier version of the report was reviewed by a large, international group of digital preservation-, digitisation- and image specialists and their comments have been included in the final version [3].

Although a need for further research was felt, the KB has, on the basis of the conclusions of this report, decided to stop creating and storing uncompressed TIFF files. For new digitisation projects, depending on the reason for storing master files (see below), JPEG 2000 Part 1 or JPEG - will be used.

This paper contains a summary of the report *Alternative File Formats for Storing Master Images of Digitisation Project*.

Definition and Exclusions

Master images are defined as followed: Raster images that are a high quality (in either colour, tonality or resolution) copy from the original source from which in most cases derivates are made for access purposes.

The following images were excluded from the scope of the study:

- Vector images
- 3D images
- Moving images
- Images in various editing layers (not identical to multiresolution images)
- Multipage files (PDF, multipage TIFF are dropped from consideration)
- Multispectral, hyperspectral images

A further exclusion was the focus on digitised low contrast material (e.g. older printed text, engravings, photographs and paintings). Higher contrast materials – read: (relatively) modern, non illustrated printed material – were out of the scope of this study.

Four Alternative Master Image Formats

In the research project four alternative file formats were selected:

1. JPEG 2000 Part 1 (lossless and lossy)
2. PNG 1.2
3. Basic JFIF 1.02 (JPEG)
4. TIFF LZW

The arguments for selecting precisely these four file formats resided in the following KB requirements for an alternative master file:

- Software support (very new or rarely used formats such as Windows Media Photo/JPEG XR and JPEG-LS are dropped from consideration).
- Sufficient bit depth: A minimum of 8 bits greyscale or 24 bits colour (bitonal, 1 bit, TIFF G4/JBIG files are dropped from consideration, as well as GIF due to 8 bits, limited colour palette).
- Possibility for lossless or high-end lossy compression (BMP excluded) [4].

Format descrition and consequences

In the research project, the four file formats were first described in general terms, concentrating on the history of the format, the standardization process (ISO or otherwise), the structure, and the encoding and decoding process.

Next, four consequences of a choice for either format were outlined:

1. *Consequences for necessary storage capacity*: how much storage space is saved in comparison to uncompressed TIFF storage? Lossless and lossy compression, and gradations of lossy compression were all taken into account.
2. *Consequences for image quality*: when using lossy compression, image degradation was evaluated by, among other things, measuring MTF [5].
3. *Consequences for long term durability*: for this purpose, a recent *quantifiable method for file format risk assessment* – developed by the KB – was used. This method ‘weighs’ file formats on the basis of seven, broadly accepted, durability criteria: Openness, Adoption, Complexity, Technical Protection Mechanism (DRM), Self-documentation, Robustness and Dependencies. Every file format receives a durability score (0-100). The working of, and the considerations behind, the File Format Assessment Method are given in Appendix three of the final report [6].
4. *Consequences for functionality*: what are the possibilities for multi-resolution, technical and descriptive metadata, simple and clear usage, effects of successive compression, and the Library of Congress’ quality and functionality factors for still images (normal rendering, clarity, color maintenance, support of graphical effects and typography and functionality beyond normal rendering) [7].

Three reasons for long term storage of Master Image Files

As said above the master file is a high quality copy from the original source from which in most cases derivates are made for access purposes. It is possible to delete the master files after the access derivates have been made. In that case, when other, more demanding use of the files is needed, digitisation will have to be performed again.

The KB distinguished three main reasons for wanting to store the master files for a long or even indefinite period:

5. Substitution (the original is susceptible to deterioration and another alternative, high-quality carrier – preservation microfilm – is not available).
6. Digitisation has been so costly and time consuming that redigitisation is no option.
7. The master file is the basis for access, or in other words: the master file is identical to the access file.

These three reasons form the basis on which the recommendations for the different file formats were made.

A comparison between the four file formats

Description of Formats

JPEG 2000 Part 1

- Standardization: JPEG 2000 Part 1 has been standardized since 2000 ISO/IEC. Other parts were standardized as well [The JPEG 2000 standard consists of twelve parts. A full description can be found at http://www.jpeg.org/jpeg2000/].
- Objective: Offer alternatives for the limited JPEG/JFIF format by using more efficient compression techniques, an option for lossless compression and multiresolution.
- Structure: The basis is a box structure which stores both the header as well as image information.
- Encoding: A six-step process. The most conspicuous is wavelet transformation (step 3) and packetizing (step 6) whereby the codestream is divided into packets and is sorted by either resolution, quality, colour or position.

PNG

- Standardization: PNG 1.2 has been ISO/IEC standardized since 2003.
- Objective: Follow up of the patented and limited GIF format, with a wealth of options as regards progressive structure, transparency, lossless compression and expansion of the standard.
- Structure: Chunks are the basis, which store both the header as well as image information.
- Encoding: A six-step process. What is notable is the option to apply separate filtering per scanline (thus increasing the effectiveness of the compression.

JPEG

- Standardization: The JPEG standard has been ISO/IEC (10918-1) standardized since 1994. An extension of Annex B of the standard – JFIF – has become the de facto standard and is simply designated as JPEG.
- Objective: To create a standard for the compression of continuous tone greyscale and colour images.
- Structure: Topic of investigation.
- Encoding: A five-step process. Most noteworthy is the use of the DCT compression technique.

TIFF LZW

- Standardization: The baseline TIFF 6.0 is not an ISO-IEC standard. The description of the baseline TIFF 6.0 (1992) is freely available on the Adobe website. LZW compression has been a part of the (extended) TIFF standard since version 5.0 (1988).
- Objective: Creation of a rich and extensible file format for raster images.
- Structure: The basis of the file format is formed by the so-called tags located both in the header (IFH) and in the image file directories (IFD).
- LZW is the compression algorithm embedded within the TIFF file [8].

Consequences for the Storage Capacity

On the storage test two limitations have been placed:

- Only 24 bit, RGB (8 bit per colour channel) files have been tested.
- Only two sets of (about 100) originals have been tested: a set of low contrast text material and a set of photographs.

File Format	Storage Gain Compared to the Uncompressed TIFF File
JPEG 2000 Part 1 lossless	52%
JPEG 2000 Part 1 lossy	Variable between 91% and 98%
PNG lossless	43%
JPEG lossy	Variable between 89% and 96%
TIFF LZW lossless	30%

Between the two sets of originals no obvious differences in storage gain were found. Is it clear however that high contrast, textual material will yield higher compression profits – this is part of further, future research.

JPEG 2000 Part 1 is obviously the most effective for lossless and lossy compression. However, JPEG is not really much inferior to lossy JPEG 2000 compression other than that compression artefacts occur earlier than with JPEG 2000 (see below).

Consequences for Image Quality

Naturally, no loss of image quality occurs with the lossless formats JPEG 2000 Part 1 lossless, PNG and TIFF LZW.

The lossy formats JPEG 2000 Part 1 lossy and JPEG degrade when compression levels are rising.

- The detail reproduction of JPEG degrades gradually when compression increases. In JPEG 2000, some loss of detail occurs only with extreme compression.
- No measurable loss of greyscale and colour (colour shift and Delta E) is observed for both JPEG and JPEG 2000. However, with increasing compression excessive "simplification" of the colour subtleties occurs which in the most extreme case results in unnatural tone and colour transitions (banding). This is caused by the quantification step in the encoding process.
- The artefacts that occur with increasing compression in JPEG 2000 and JPEG resemble each other a lot. What is important to note is that the visibility of these artefacts occurs much earlier in JPEG than in JPEG 2000. The following artefacts become visible with mounting compression:
  - Banding (rough colour or tone transitions)
  - Visible tiles (the tiles into which the files are divided become visible)
  - Woolly effect around elements rich in contrast.

A remaining topic of investigation is the expression of PSNR (Peak Signal-to-Noise Ratio) which gives an objective figure (expressed in dB) for the degradation that occurs during lossy compression.

Consequences for the Long-Term Sustainability

Application of the previously discussed File Format Assessment Method to the image formats discussed in this report, plus the uncompressed TIFF format that has been used until now, results in the following order from most to least suitable for long-term storage:

Ranking	Format	Score
1	Baseline TIFF 6.0 uncompressed	84,8
2	PNG 1.2	78,0
3	JP2 (JPEG 2000 Part 1) lossless	74,7
4	JP2 (JPEG 2000 Part 1) lossy	66,1
5	Basic JFIF (JPEG) 1.02	65,4
6	TIFF 6.0 with LZW compression	65,3

In appendix 2 of the final study the above scores are further specified.

The main thing is that from the perspective of long-term sustainability the choice for "Baseline TIFF 6.0 uncompressed" is the safest one. In practice it appears that this is not a viable option due to the large size of the files and the associated high storage costs.

The 'File Format Assessment Method" is still in its infancy. Feedback is being awaited from colleague institutions regarding this method. Additionally, there is not much experience with the application of this method in practice. Based on the experiences gained in the research project it appears necessary to adapt the method. It is therefore too early to entirely ascribe the choice of a durable format to this method. The results of the method will be tested against previous knowledge and experiences.

As the above table indicates, the choice for "Baseline TIFF 6.0 uncompressed" is the safest one from the perspective of long-term sustainability. If an alternative format has to be selected, we see that "PNG 1.2" and "JP2 (JPEG 2000 Part 1) lossless" – both lossless compressed formats – are the alternatives. Here we reach a point where the applied method may fall short. In the method, the characteristic "Usage in the cultural heritage sector as master image file" of the Adoption criterion makes a valuable contribution to the total score. However, what is not included in the method at the moment are the prospects for the future of this criterion. Although neither format is currently used on a large scale as a preservation master file in the cultural sector, JPEG 2000 has more potential. PNG has been in existence since 1996 and JP2 since 2000. The preference, for lossless formats, is thus for JPEG 2000.

Another issue that is neglected by the method is the loss of image quality caused by applying lossy compression methods. Although a file that is a qualitatively worse representation of the original can also be stored for the long-term, it is important – certainly if the original cannot be rescanned – to consider the use of the digitalized material in the long term. What must be considered in this respect is that a loss of quality which may be deemed acceptable today may no longer be acceptable in the future. For example, you might consider the use of alternative "display" hardware with a better resolution or different scope. From a long-term sustainability perspective, the use of lossy compression algorithms is discouraged. This certainly applies when the objective of digitisation is to replace the original (objective 1, substitution). If a lossy compression method is selected nevertheless, the use of "basic JFIF (JPEG) 1.02" is recommended due to the more certain future of this format as compared to the lossy JPEG 2000 Part 1 variant.

The ultimate advice, rendered exclusively from the perspective of long-term sustainability and the File Format Assessment Method, for an alternative image format for uncompressed TIFFs comes down to the following list, sorted from most to least suitable:

1. JP2 (JPEG 2000 Part 1) lossless
2. PNG 1.2
3. Basic JFIF (JPEG) 1.02
4. JP2 (JPEG 2000 Part 1) lossy
5. TIFF 6.0 with LZW compression

Consequences for the Functionality

Only the most relevant functions (for master storage) are listed in the table below.

Functionality	File Format
Lossless compression option	JPEG 2000 Part 1, PNG, TIFF LZW
Lossy compression option	JPEG 2000 Part 1, JPEG
Lossy and lossless compression option	JPEG 2000 Part 1
Option to add bibliographic metadata	JPEG 2000 Part 1, PNG, JPEG, TIFF LZW
Standard way to embed EXIF metadata	JPEG, TIFF LZW
Browser support	JPEG, PNG
Multiresolution options (suitability of the file as a high-resolution *access* master)	JPEG 2000 Part 1, TIFF LZW, to a very slight degree: JPEG
Maximum size	JPEG 2000 Part 1: unlimited (2^64). PNG: Topic of investigation. JPEG: Topic of investigation. TIFF LZW: 4 GB
Bit depths	JPEG 2000: 1 to max. 38 bits per channel. Compliance class 2: 16 bits per channel. PNG: 1 to 16 bits per channel. JPEG: 8 bits per channel. TIFF LZW: 1 to 16 bits per channel (theoretically to 32 bits per channel)
Standard support of colour spaces	JPEG 2000 Part 1: bitonal, greyscale, sRGB, palletized/indexed colour space PNG: bitonal, greyscale, sRGB, palletized/indexed colour space JPEG: greyscale, RGB TIFF LZW: Bitonal, greyscale, RGB, CMYK, YCbCR, CIEL*a*b
Option to use ICC profiles	JPEG2000 Part 1, PNG, JPEG, TIFF LZW (although not in a standard manner)
Multipage support	TIFF LZW

Summary

The table below summarizes all the above information in a matrix. The figures only indicate the order of success in the various parts.

	JP2 part 1 lossless	JP2 Part 1 lossy	PNG lossless	JPEG lossy	TIFF LZW lossless
Standardization	5	5	5	5	5
Storage Savings	3	5	2	4	1
Image Quality	5	4	5	3	5
Long-term Sustainability	5	2	4	3	1
Functionality	5	5	4	3	4
Score	**23**	**21**	**20**	**18**	**16**

It is noteworthy that JPEG 2000 comes out on top in both the lossless as well as the lossy versions.

The table above does not make a distinction between the three reasons for the long-term storage of master files as mentioned in the introduction. Some of the criteria on the left hand side of the table are less relevant depending on these reasons. In the recommendations below the importance of each of the five criteria are taken into account.

Recommendations

The recommendations for a an alternative master image follow the three reasons for long term storage of these files described above.

Reason 1: Substitution

The criteria “Long-term sustainability”, “Standardisation” and “Image Quality” are considered the most important when substitution of the original is the main reason for the long-term storage of the master file. JPEG 2000 Part 1 lossless, closely followed by PNG, are the most obvious choices from the perspective of long-term sustainability. When the storage savings (PNG 40%, JPEG 2000 lossless 53%) and the functionality are factored in, the scale tips in favour of JPEG 2000 lossless. The lossless TIFF LZW is not a viable option due to the slight storage gain (30%) and the low score in the File Format Assessment Method (especially due to patents, resulting in a low score on the “Restrictions on the interpretation of the file format” characteristic).

Due to the irreversible loss of image information, lossy compression is a much less obvious choice for this objective.

The creation of visual lossless images might be considered though. Both JPEG 2000 Part 1 (compression ratio 10, storage gain about 90%) and JPEG (PSD10 and higher, storage savings about 89%) offer options in this respect. In the latter case, it must be understood that visual lossless is a relative term – it is based on the current generation of monitors and the subjective experience of individual viewers. A big advantage of the JPEG file format is the enormous distribution and the comprehensive software support, including browsers.

Reason 2: Redigitisation Is Not Desirable

The criteria “Storage savings” and “Image Quality” are considered the most important when the main reason for the long-term storage of the master files is not wanting to do redigitisation. In this case lossy compression, in the visual lossless mode, is a more viable option. The small amount of information loss can be defended more easily in this case because there is no substitution. The above mentioned JPEG 2000 lossy and JPEG visual lossless versions are the obvious choices.

However, if absolutely no image information may be lost, then the above-mentioned JPEG 2000 lossless and PNG formats are the two recommended options.

Reason 3: Master File is Access File

The criteria “Storage savings” and “Functionality” are considered the most important when access is the main reason for the long-term storage of the master file. In this case a larger degree of lossy compression is self-evident. The two options are then JPEG 2000 Part 1 lossy and JPEG with a higher level of compression. The advanced JPEG 2000 compression technique enables more storage reduction without much loss of quality (superior to JPEG). When selecting the amount of compression, the type of material must be taken into account. Compression artefacts will be more visible in text files than in continuous tone originals such as photos, for example. However, the question is whether the more efficient compression and extra options of JPEG 2000 outweighs the JPEG format for this purpose, which is comprehensively supported by software (including browsers) and is widely distributed.

References

[1] Judith Rog, Compression and digital preservation: do they go together? (IS&T Archiving 2007 Final program and proceedings, 2007) pg 80-83.

[2] The final report is published on the KB website: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/Alternative%20File%20Formats%20for%20Storing%20Masters%202%201.pdf.

[3] Judith Rog and Robèrt Gillesse, Alternative File Formats for Storing Master Images of Digitisation Projects (KB 2008) pg 10.

[4] TIFF with lossless ZIP compression was dropped from the study out of sheer shortage of time.

[5] MTF (Modulation Transfer Function) is a measurement of detail reproduction of an optical system. Output is in reproduced line pairs (or cycles) per millimeter.

[6] Judith Rog, Caroline van Wijk, Evaluating File Formats for Long-term Preservation (KB 2007). http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/Alternative%20File%20Formats%20for%20Storing%20Masters%202%201.pdf#page=50 .

[7] http://www.digitalpreservation.gov/formats/content/still_quality.shtml.

[8] Other compression schemes that can be used within the TIFF file are ITU_G4, JPEG and ZIP.

Author Biography

Judith Rog (1976) completed her MA in Phonetics/Speech Technology in 1999. After working on language technology at a Dutch Dictionary Publisher she was employed at the National Library of the Netherlands/Koninklijke Bibliotheek (KB) in 2001. She first worked in the IT department of the KB for four years before joining the Digital Preservation Department in 2005. Within the Digital Preservation Department she participates in several projects in which her main focus is on file format research.

Robèrt Gillesse (1967) completed his MA in history at the University of Leiden in 1995. In 1996/7 he followed a postgraduate programme on history and computing at the University of Leiden. In 1998 he started working at the National Library of the Netherlands/Koninklijke Bibliotheek (KB) where he has been quality manager digitisation for the past six years. He sets up guidelines for digitisation projects, coordinates and performs quality control and conducts research in the area of image quality.

Astrid Verheusen (1967) completed her MA in history at the University of Leiden in 1992. In 1993 she followed a postgraduate programme on history and computing at the University of Leiden. She has been working at the National Library of the Netherlands since 2001. She has many years of experience in research and development projects in the field of digitisation and digital preservation.