

Alternative File Formats for Storing Master Images of Digitisation Projects

Date: March 7, 2008
Authors: Robèrt Gillesse
Judith Rog
Astrid Verheusen
National Library of the Netherlands
Research & Development Department
Version: 2.0
Status: Final
File Name: Alternative File Formats for Storing Masters 2.0.doc

Management Summary

This document is the end result of a study regarding alternative formats for storing master files of digitisation projects of the Koninklijke Bibliotheek (KB) in The Hague, the Netherlands. The study took place in the context of reviewing the KB's storage strategy. The magnitude of digitisation projects is increasing to such a degree – the estimate is a production of 40 million images (counting master versions only) in the next four years – that a revision of this strategy is deemed necessary. Currently, master files are stored in uncompressed TIFF file format, a format used world wide. 650 TB of storage space will be necessary to store 40 million files in this format. The objective of the study was to describe alternative file formats in order to reduce the necessary storage space. The desired image quality, long-term sustainability and functionality had to be taken into account during the study.

The following four file formats were reviewed:

- JPEG 2000 part 1 (lossless and lossy)
- PNG 1.2
- Basic JFIF 1.02 (JPEG)
- TIFF LZW

For each file format, we examined what the consequences would be for the following if that format were to be selected:

1. The required storage capacity
2. The image quality
3. The long-term sustainability
4. The functionality

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

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

The recommendations for the choice of file format were made on the basis of these three reasons.

The study made use of existing knowledge and expertise at the KB's Research & Development Department. The *quantifiable file format risk assessment method* recently developed by the KB was employed for examining the long-term sustainability of the formats. The results of the study were presented to a selection of national and international specialists

in the area of digital preservation, file formats and file management. Their comments are incorporated into the final version of this document.

The main conclusions of this study are as follows:

Reason 1: Substitution

JPEG 2000 lossless and PNG are the best alternatives for the uncompressed TIFF file format 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.

Reason 2: Redigitisation Is Not Desirable

JPEG 2000 and JPEG are the best alternatives for the uncompressed TIFF file format. If no image information may be lost, then JPEG 2000 lossless and PNG are the two recommended options.

Reason 3: Master File is the Access File

JPEG 2000 lossy and JPEG with greater compression are the most suitable formats.

Table of Contents

MANAGEMENT SUMMARY	2
1 INTRODUCTION	6
1.1 CONSEQUENCES	7
1.2 THREE REASONS FOR LONG TERM STORAGE OF MASTER FILES.....	9
1.3 CONCLUSION.....	9
1.4 REVIEW BY SPECIALISTS.....	9
1.5 FOLLOW-UP STUDY	10
2 JPEG 2000	12
2.1 WHAT IS JPEG 2000?	12
2.1.1 General Information	12
2.1.2 JPEG 2000 Parts.....	12
2.2 HOW DOES IT WORK?	15
2.2.1 Structure	15
2.2.2 Encoding and Decoding	15
2.3 CONSEQUENCES FOR THE REQUIRED STORAGE CAPACITY.....	18
2.4 CONSEQUENCES FOR THE IMAGE QUALITY	18
2.5 CONSEQUENCES FOR THE LONG-TERM SUSTAINABILITY.....	19
2.6 CONSEQUENCES FOR THE FUNCTIONALITY	20
2.7 CONCLUSION.....	21
3 PNG	24
3.1 WHAT IS PNG?.....	24
3.2 HOW DOES IT WORK?	25
3.2.1 Structure	25
3.2.2 Encoding and Decoding/Filtering and Compression	25
3.3 CONSEQUENCES FOR THE REQUIRED STORAGE CAPACITY.....	25
3.4 CONSEQUENCES FOR THE IMAGE QUALITY	26
3.5 CONSEQUENCES FOR THE LONG-TERM SUSTAINABILITY.....	26
3.6 CONSEQUENCES FOR THE FUNCTIONALITY	26
3.7 CONCLUSION.....	27
4 JPEG	29
4.1 WHAT IS JPEG?	29
4.2 HOW DOES IT WORK?	30
4.2.1 Structure	30
4.2.2 Encoding and Decoding/Filtering and Compression	30
4.3 CONSEQUENCES FOR THE REQUIRED STORAGE CAPACITY.....	31
4.4 CONSEQUENCES FOR THE IMAGE QUALITY	31
4.5 CONSEQUENCES FOR THE LONG-TERM SUSTAINABILITY.....	32
4.6 CONSEQUENCES FOR THE FUNCTIONALITY	32
4.7 CONCLUSION.....	33
5 TIFF LZW	35
5.1 WHAT IS TIFF LZW?.....	35
5.2 HOW DOES IT WORK?	36
5.2.1 Structure	36
5.2.2 Encoding and Decoding/Filtering and Compression	36
5.3 CONSEQUENCES FOR THE REQUIRED STORAGE CAPACITY.....	36
5.4 CONSEQUENCES FOR THE IMAGE QUALITY	36
5.5 CONSEQUENCES FOR THE LONG-TERM SUSTAINABILITY.....	36
5.6 CONSEQUENCES FOR THE FUNCTIONALITY	37

5.7 CONCLUSION.....	38
6 CONCLUSION	40
APPENDIX 1: USE OF ALTERNATIVE FILE FORMATS.....	45
APPENDIX 2: FILE FORMAT ASSESSMENT METHOD – OUTPUT	48
APPENDIX 3 FILE FORMAT ASSESSMENT METHOD – EXPLAINED.....	50
APPENDIX 4: STORAGE TESTS	62
BIBLIOGRAPHY	63

1 Introduction

The study took place in the context of reviewing the Koninklijke Bibliotheek's (KB) storage strategy for digitisation projects. The magnitude of digitisation projects is increasing to such a degree – the estimate is 40 million images and 650 TB in uncompressed data storage by 2011, counting master versions only – that a revision of this strategy is deemed necessary. The central questions are whether all master files of digitisation projects actually have to be stored in the long-term storage system, what the costs are of long-term storage and what the alternatives are besides expensive, uncompressed, high-resolution storage in TIFF file format.

This study focuses on the last question. Its objective is to describe alternative file formats besides uncompressed TIFF for storing image master files.

In this study the context is that of digitized low contrast material – which means originals like for instance older printed text, engravings, photographs and paintings. Higher contrast materials – read: (relatively) modern, non illustrated printed material – are out of the scope of this study. The classification of different types of originals on the basis of their information value, the selection of a suited digitization quality connected to this value and, subsequently, the choice of using either lossy compression, lossless compression or no compression at all, are issues that have not been elaborated upon in this study. These two subjects will have to be part of a possible second version of this study.

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 derivatives are made for access use.

The following images are excluded from the scope of this study:

- Vector images
- 3D images
- Moving images
- Images in various editing layers (not identical to multiresolution images¹)
- Multipage files (PDF, multipage TIFF are dropped from consideration)
- Multispectral, hyperspectral images²

The following four file/compression formats will be reviewed:

1. JPEG 2000 part 1 (lossless and lossy)³
2. PNG 1.2

¹ Photoshop .psd or TIFF multilayer files, for example.

² This is because multispectral imaging has been no serious consideration for KB digitisation projects. This is not to say that this will not be case in the future. For sake of the argument now, multispectral images are not relevant.

³ A review of JPEG 2000 as an alternative file format was already conducted in large measure in 2007 by Judith Rog: *Notitie over JPEG 2000 voor de KB (Note regarding JPEG 2000 for the RL)*, version 0.2 (August 2007).

3. Basic JFIF 1.02 (JPEG)
4. TIFF LZW

The arguments for selecting precisely these four file formats reside in the following 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 consideration⁴, as well as GIF due to 8 bits, limited colour palette).
- Possibility for lossless or high-end lossy compression (BMP excluded).

TIFF with lossless ZIP compression is excluded from this study out of sheer shortage of time. In the second version of this study TIFF zip will have to be included.

1.1 Consequences

This report has an individual section for each of the four file formats listed above. A summary description of the format and how it works is followed by subsections describing the consequences of using the format in the following areas:

1. Consequences for the required storage capacity
2. Consequences for the image quality
3. Consequences for the long-term sustainability
4. Consequences for the functionality

Sub 1: This section provides an outline of the storage consequences of the format choice. The storage gain of the compressed compared to the uncompressed TIFF file is calculated in term of percentage: if necessary, a differentiation shall be made between lossy and lossless compression. Two sets comprising about one hundred scans are employed for the calculation... On these tests two limitations have been set:

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

These limitations were born out of fact that the great majority of KB files that were made and shall be made in the near future are of this nature: 24 bit RGB files of a low contrast nature. Of course higher (and maybe lower) bit depths and high contrast materials (modern print), which will yield other compression ratio's, will have to be included a in later versions of this study.

See Appendix 4 for the results of the text set.

Sub 2. This section attempts to outline the difference with regard to the uncompressed master file using as many quantifiable terms as possible (among other things by means of the Peak Signal-to-Noise Ratio – PSNR⁵ – and Modulation Transfer Function – MTF⁶).

⁴ Whether bitonal files actually fall outside the scope of image master files is not yet certain. It may be that the loss of brightness values is considered acceptable for some access projects (see below) of (relatively) modern, unillustrated material.

The following technical targets and tools are used to determine the possible decrease of image quality.

- Possible loss of detail is measured by means of the QA-62 SFR and OECF test chart.
- Possible loss of greyscale is measured using the Kodak Greyscale.
- Possible loss of colour is measured using the MacBeth ColorChecker.
- Artefacts are determined through visual inspection.

Sub 3: This section employs the *quantifiable file format risk assessment method* recently developed by Judith Rog, Caroline van Wijk and Jeffrey van der Hoeven for the KB. Using this method, file formats can be measured based on seven widely accepted sustainability criteria. The criteria are as follows: Openness, Adoption, Complexity, Technical Protection Mechanism, Self-documentation, Robustness and Dependencies. Each file format receives a sustainability score in this method. These seven main criteria are subdivided into measurable sub-characteristics. For example, the main criterion “Openness” is subdivided into the characteristics “Standardization,” “Restrictions on the interpretation of the file format” and “Reader with freely available source”. Each format receives a score between 0 and 2 for each characteristic. The method precisely defines how the score is determined. For example, a format will receive the maximum score of 2 for the “Standardization” characteristic if it is a “de jure standard”, a score of 1.5 if it is a “de facto standard” and so on down to a score of 0. The scores are subsequently multiplied with a weighing factor that is attributed to each main criterion or characteristic. The weights that are assigned to the criteria and their characteristics are not fixed. They depend on the local policy of an institution. A weight of 0 can be assigned if an institution chooses to ignore the characteristic. The weights that are used in the examples in this paper are the weights as assigned by the KB based on its local policy, general digital preservation literature and common sense. For example, the sub-characteristics “Standardization,” “Restrictions on the interpretation of the file format” and “Reader with freely available source” of the “Openness” criterion receive a weighing factor at the KB of 9 (Standardization), 9 (Restrictions on the interpretation of the file format) and 7 (Reader with freely available source), and all sub-characteristics of the criterion “Self-documentation,” which includes the option of adding metadata to files, receive a weighing factor of 1. The KB will initially not employ metadata that is stored in the files themselves. This is the reason for the relatively low weighing factor for this criterion. However, this may be different for other institutions. In this method, each file format ultimately receives a sustainability score between 0 and 100. The higher the score, the more suitable the format is for long-term storage and permanent access.

Appendix 2 of this document contains the interpretation of the method for the formats that are discussed in this report. Appendix 3 explains the method. For this study all discussed formats received a score of 0 for the “Support for file corruption detection” characteristic because the time and expertise to research this was lacking at this point in time. We are aware that PNG does provide a level of corruption detection in the file header, but lacked the time to research whether and to what level this is case for the other formats. Because all formats received the same score, this ultimately plays no role in the relative final scores of the formats with regard to each other.

Because the method was developed so recently and feedback is still awaited from colleague institutions, the ultimate choice for an alternative format is not solely based on the File

⁵ Cit.“[...] the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation.” Wikipedia: <http://en.wikipedia.org/wiki/PSNR>

⁶ MTF is a measurement of detail reproduction of an optical system. Output is in reproduced line pairs (or cycles) per millimetre.

Format Assessment Method. The results of the method are tested against other knowledge and experience in this area.

Sub 4: This section outlines the consequences for functionality. The section deals with questions of usage such as the following:

- Is the file format suitable as a high-resolution access master?
- Are there options for including bibliographic and technical (EXIF) metadata?
- Are the Library of Congress criteria regarding quality and functionality factors complied with? Normal display, clarity (support of colour spaces, possible bit depths), colour maintenance (option to include gamma correction and ICC colour profiles), support of graphical effects and typography (Alpha channel where transparency information can be stored) and functionality beyond the normal reproduction (animation, multipage and multiresolution support)⁷?

1.2 Three Reasons for Long Term Storage of Master Files

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

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

1. Substitution (the original is susceptible to deterioration and another alternative, high-quality carrier – preservation microfilm – is not available)
2. Digitisation has been so costly and time consuming that redigitisation is no option
3. 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 are made.

1.3 Conclusion

Each analysis is followed by a conclusion per format and after all analyses have been discussed, an overall conclusion is presented where the various formats are compared and recommendations are made for the selection of an alternative file format. The above mentioned reasons for long term storage will be included in this.

1.4 Review by Specialists

A panel of national and international specialists in the area of digital preservation, file formats and file management was asked to critically review this study and provide comments, where required. Their input is incorporated into this version of the report. The panel consisted of the following persons:

⁷ Sustainability of Digital Formats - Planning for Library of Congress Collections - Still Images Quality and Functionality Factors http://www.digitalpreservation.gov/formats/content/still_quality.shtml

- Stephen Abrams (Harvard University Library/University of California-California Digital Library)
- Caroline Arms (Library of Congress, US)
- Martijn van den Broek (Nederlands Fotomuseum [Netherlands Photo Museum], the Netherlands)
- Adrian Brown (National Archives, UK)
- Robert R. Buckley (Xerox Corporation)
- Aly Conteh (British Library)
- Carl Fleischhauer (Library of Congress)
- Rose Holley (National Library of Australia)
- Marc Holtman (City Archive of Amsterdam)
- Rene van Horik (DANS, the Netherlands)
- Dr. Klaus Jung (Luratech Imaging GmbH)
- Ulla Bøgvad Kejser (Kongelige Bibliotek Denmark)
- Rory McLeod (British Library)
- Andrew Stawowczyk Long (National Library of Australia)
- Boudewijn de Ridder (Nederlands Fotomuseum [Netherlands Photo Museum], the Netherlands)
- Brian Thurgood (Libraries and Archives of Canada)
- Thomas Zellmann (LuraTech Europe GmbH)

We would like to thank all experts for their very useful feedback that improved this report considerably. The amount of feedback that we received was overwhelming and showed us that the problem that was the immediate cause for this report is relevant at many other institutions as well.

1.5 Follow-up Study

All the feedback that we have received confirmed us in the idea that the report is far from complete as it is. We could easily have spent several more months on further, in-depth study on all the topics that are being addressed in the report. Unfortunately we lack the time to do so.

Among others, these items remain open for further study:

- A classification of different types of originals on the basis of their information value, the selection of a suited digitization quality connected to this value and, subsequently, the choice of using either lossy compression, lossless compression or no compression at all
- Further compression tests including:
 - High contrast, textual material
 - 16 bit files
 - Greyscale files
 - Using alternative compression software for JPEG2000 and PNG

- PSNR – Peak Signal To Noise Ratio
- Structure of the JPEG file
- Functioning of LZW compression
- Further work on the “File Format Assessment Method” that is being used in this report to assess file formats on their long-term sustainability. On the basis of the feedback from experts mentioned above, we have already adjusted and refined the method, but it will need our further and constant attention.

We are very much open to all input from others on one of these or all other topics from this study.

2 JPEG 2000

2.1 What is JPEG 2000?

2.1.1 General Information

JPEG 2000 is a standard (ISO/IEC 15444-1/ITU-T Rec. T.800) developed by the JPEG (Joint Photographic Experts Group) as a joint effort of the ISO, IEC and ITU-T standardization organizations. These groups are comprised of representatives of various commercial parties and academic institutes from the four corners of the globe.

The objective of the JPEG group was to develop a new image standard with the following basic principles:

- Complete openness of the format.
- An improved lossy compression algorithm compared to the current JPEG compression.
- An option for lossless compression.
- Comprehensive options for bundling metadata in the image file.
- Storage of several resolutions within one file.

These basic principles were implemented in the JPEG 2000 standard.

2.1.2 JPEG 2000 Parts

JPEG 2000 in the year 2007 is divided into twelve standards that are all more or less derivations of or supplements to the first standard: Part 1. This concerns still images (part 1 .jp2 and part 2 .jpx), documents (part 6 .jpm) and moving images (part 3 .mj2). The employed wavelet compression technology is the connecting element.

Only parts 1, 2, 4, 6 and 8 seem to be relevant for storing masters of still images.

The following contains an overview of the twelve parts in a summarized form⁸

Part 1

As its name suggests, Part 1 defines the core of JPEG 2000. This includes the syntax of the JPEG 2000 codestream and the necessary steps involved in encoding and decoding JPEG 2000 images. The later parts of the standard are all concerned with extensions of various kinds, and none of them is essential to a basic JPEG 2000 implementation. A number of existing implementations use only Part 1.

Part 1 also defines a basic file format called JP2. This allows metadata such as colour space information (which is essential for accurate rendering) to be included with a JPEG 2000 codestream in an interoperable way. JP2 uses an extensible architecture shared with the other file formats in the JPEG 2000 family defined in later parts of the standard.

Part 1 also includes guidelines and examples, a bibliography of technical references, and a list of companies from whom patent statements have been received by ISO. JPEG 2000 was developed with the intention that Part 1 could be implemented without the payment of licence

⁸ See for full description of the parts the JPEG 2000 homepage: <http://www.jpeg.org/jpeg2000/>.

fees or royalties, and a number of patent holders have waived their rights toward this end. However, the JPEG committee cannot make a formal guarantee, and it remains the responsibility of the implementer to ensure that no patents are infringed.

Part 1 became an International Standard (ISO/IEC 15444-1) in December 2000.

A second edition of Part 1 was published in 2004. Among other things, a standard colour spaces (YCC) was added.

Part 2

Part 2 defines various extensions to Part 1, including:

- More flexible forms of wavelet decomposition and coefficient quantization
- An alternative way of encoding regions of particular interest (ROIs)
- A new file format, JPX, based on JP2 but supporting multiple compositing layers, animation, extended colour spaces and more
- A rich metadata set for photographic imagery (based on the DIG35 specification)

Most of the extensions in Part 2 operate independently of each other. To assist interoperability, mechanisms are provided at both the codestream and the JPX file format level for signalling the use of extensions.

Part 2 became an International Standard (ISO/IEC 15444-2) in November 2001.

Part 3

Part 3 defines a file format called MJ2 (or MJP2) for motion sequences of JPEG 2000 images. Support for associated audio is also included.

Part 3 became an International Standard (ISO/IEC 15444-3) in November 2001

Part 4

JPEG 2000 Part 4 is concerned with testing conformance to JPEG 2000 Part 1. It specifies test procedures for both encoding and decoding processes, including the definition of a set of decoder compliance classes. The Part 4 test files include both bare codestreams and JP2 files.

Note that JPEG 2000 Part 4 explicitly excludes from its scope acceptance, performance or robustness testing.

Part 4 became an International Standard (ISO/IEC 15444-4) in May 2002.

Part 5

JPEG 2000 Part 5 (ISO/IEC 15444-5:2003) consists of a short text document, and two source code packages that implement JPEG 2000 Part 1. The two codecs were developed alongside Part 1 and were used to check it and to test interoperability. One is written in C and the other in Java. They are both available under open-source type licensing.

Part 5 became an International Standard (ISO/IEC 15444-5) in November 2001.

Part 6

Part 6 of JPEG 2000 defines the JPM file format for document imaging, which uses the Mixed Raster Content (MRC) model of ISO/IEC 16485. JPM is an extension of the JP2 file format defined in Part 1: it uses the same architecture and many of the same boxes defined in Part 1 (for JP2) and Part 2 (for JPX).

JPM can be used to store multi-page documents with many objects per page. Although it is a member of the JPEG 2000 family, it supports the use of many other coding or compression technologies as well. For example, JBIG2 could be used for regions of text, and JPEG could be used as an alternative to JPEG 2000 for photographic images.

Part 6 became an International Standard (ISO/IEC 15444-6) in April 2003.

Part 7

This part has been abandoned.

Part 8

JPEG 2000 Secured (JPSEC) or Part 8 of the standard is standardizing tools and solutions in terms of specifications in order to ensure the security of transaction, protection of contents (IPR), and protection of technologies (IP), and to allow applications to generate, consume, and exchange JPEG 2000 Secured bitstreams. The following applications are addressed: encryption, source authentication, data integrity, conditional access, ownership protection.

Part 8 became an International Standard (ISO/IEC 15444-8) in July 2006

Part 9

The main component of Part 9 is a client-server protocol called JPIP. JPIP may be implemented on top of HTTP, but is designed with a view to other possible transports.

Part 9 became an International Standard (ISO/IEC 15444-9) in October 2004.

Part 10

Part 10 is at the end of the Approval Stage (50.60). It is concerned with the coding of three-dimensional data, the extension of JPEG 2000 from planar to volumetric images.

Part 11

To address this issue, JPEG 2000 Wireless (JPWL) or Part 11 of the standard is standardizing tools and methods to achieve the efficient transmission of JPEG 2000 imagery over an error-prone wireless network. More specifically, JPWL extends the elements in the core coding system described in Part 1 with mechanisms for error protection and correction. These extensions are backward compatible in the sense that decoders which implement Part 1 are able to skip the extensions defined in JPWL.

Part 11 became an International Standard (ISO/IEC 15444-11) in June 2007.

Part 12

Part 12 of JPEG 2000, ISO/IEC 15444-12, has a common text with Part 12 of the MPEG-4 standard, ISO/IEC 14496-12. It is a joint JPEG and MPEG initiative to create a base file format for future applications. The format is a general format for timed sequences of media data. It uses the same underlying architecture as Apple's QuickTime file format and the JPEG 2000 file format.

Part 12 became an International Standard (ISO/IEC 15444-12) in July 2003.

Part 13 - An entry level JPEG 2000 encoder

Part 13 defines a royalty- and license-fee free entry-level JPEG 2000 encoder with widespread applications. There is no Final Committee Draft available yet.

2.2 How does it work?

2.2.1 Structure

A JPEG 2000 file is comprised of a succession of boxes. A box can contain other boxes and is then called a superbox.⁹ The boxes are of variable length. The length is determined by the first four bytes. Each box has a type that is determined by the second sequence of the four bytes.

Each file of the JPEG 2000 family begins with a JPEG 2000 signature box, followed by a file type box which determines, among other things, the type (e.g. JP2) and the version. This is followed by the header box, which contains various boxes in which the resolution, bit depth and colour specifications are set down, among other things. Optional are boxes in which XML and non-XML structured metadata can be determined about the file. This is followed by a “contiguous codestream” box which contains the image data.¹⁰

2.2.2 Encoding and Decoding

JPEG 2000 encoding takes place in six steps¹¹:

Step 1: Colour Component Transformation (optional)

First, the RGB colour space is changed to another colour space. This is an optional step, but mostly used and recommended for RGB-like colour spaces. Two options are possible for this:

1. Irreversible Colour Transform (ICT) to the YCbCr colour space
2. Reversible Colour Transform (RCT) to the YUV colour space

The first method is used for lossy compression and includes a simplification of the colour information and can bring about quantification errors.

Step: 2 Tiling

After the colour transformation the image is divided into so-called tiles. The advantage of this is that the decoder requires less memory in order to create the image. The size of the tiles can even be selected (if the encoding software offers this advanced option). If the tiles are made too small, or if the compression factor is very high, the same blocking effect can occur as with JPEG (this only applies to lossy compression). The size of the tiles has a minimal effect on

⁹ This box structure is related to the Quicktime and MPEG-4 format. Boxes are “atoms” in these formats.

¹⁰ For a comprehensive overview of the box structure of JP2, see the Florida Digital Archive description of JP 2 - section 1.14: http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgounds/jp2_bg.pdf.

¹¹ Wikipedia, [JPEG2000: http://en.wikipedia.org/wiki/JPEG_2000#Technical_discussion](http://en.wikipedia.org/wiki/JPEG_2000#Technical_discussion).

the file size: when a smaller tile is chosen, the file becomes a bit larger.¹² This step is optional as well, in the sense that you can use one single tile that has the same dimensions as the whole image. This would prevent the blocking effect/tiling artefacts, mentioned earlier.

Step 3: Wavelet Transformation

The tiles are then transformed with Discrete Wavelet Transformation (DWT).¹³

There are two possibilities for this:

1. Lossy (or visual lossless) compression by means of the 9/7 floating point wavelet filter.
2. Lossless compression by means of the 5/3 integer wavelet filter.¹⁴

Step 4: Quantification (for lossy compression only)

Scalar quantification of the coefficients in order to decrease the quantity of bits that represent them. The result is a set of whole numbers that must be encoded. The so-called quantification step is a flexible parameter: the larger this step, the greater the compression and the loss of quality.

Step 5: Encoding

Encoding includes a hierarchical succession of continually smaller “units”:

1. Sub-bands – frequency range and spatial area. These elements are split into:
2. Precincts – rectangular regions in the wavelet domain. These elements are split into the smallest JPEG 2000 element:
3. Code blocks: square blocks in a sub-band. The bits of the code blocks are encoded by means of the EBCOT (Embedded Block Coding with Optimal Truncation) scheme. The significant bits are encoded first and then the less significant bits. The encoding itself takes place in three steps (coding passes), whereby the less relevant bits are filtered out in the lossy version.

Step 6: Packetizing

This is the process whereby the codestream is divided into “packets” and “layers” that can be sorted by resolution, quality, colour or position within a tile.

Packets contain the compressed data of the code blocks of a specific position of a given resolution of a component of a tile.

The packets, in turn, are a component of a layer: a layer is a collection of packets, one of each position, for each resolution.¹⁵

By arranging these layers in a certain way, it is possible during decoding/access to stipulate that certain information be made available first and other information later. This particularly plays a role for access via the Web.

¹² Robert Buckley, *JPEG 2000 for Image Archiving, with Discussion of Other Popular Image Formats*. Tutorial IS&T Archiving 2007 Conference, p. 41, slide 81.

¹³ Instead of Discrete Cosine Transformation (DCT), which is used for JPEG. The DCT technique works in blocks of 8x8 pixels, which renders the image pixelated with higher compression.

¹⁴ Robert Buckley, *JPEG 2000 for Image Archiving, with Discussion of Other Popular Image Formats*. Tutorial IS&T Archiving 2007 Conference, p. 42, slide 83.

¹⁵ *Ibidem*, p. 32, slide 64.

For example, if you choose to arrange the decoding per resolution, then you can first offer a low-resolution image during access, with larger-resolution images becoming available as the decoding goes on. If you arrange the codestream by quality, then you can repeatedly offer more quality/bit depth. Arranged by colour channels, you can always offer various colours and arranged by location, you can show certain parts of the image first. For example, the codestream can be arranged so that access takes place first by Quality (L), then Resolution (R), then Colour Channel (C) and then Position (P). The order is then LRCP. Other possible orders are: RLCP, RPCL, PCRL and CPRL. A special option of LRCP (LRCP with Region of Interest Encoding) is constructing a certain part of the image first.¹⁶

The two illustrations below¹⁷ show how continually decoding more blocks results in a continuously higher resolution (RPCL).

Scalability: Progressive By Resolution

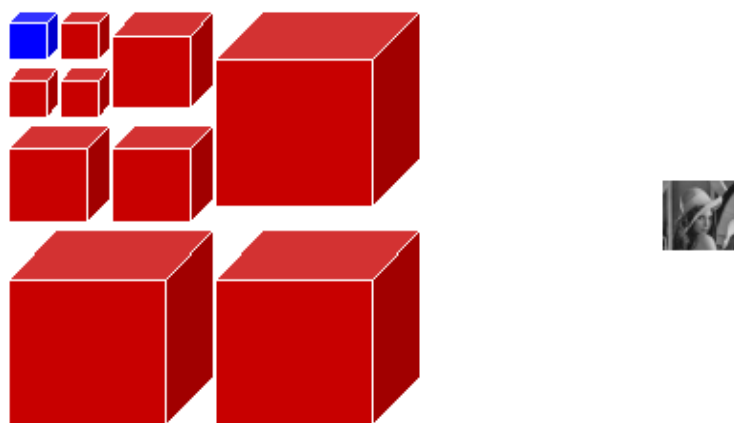


Figure 1: Resolution with one block decoded

¹⁶Buckley, *JPEG 2000 Image Archiving*, page 34, slide 68.

¹⁷ *Ibidem*, p. 28, slide 55, 56.

Scalability: Progressive By Resolution

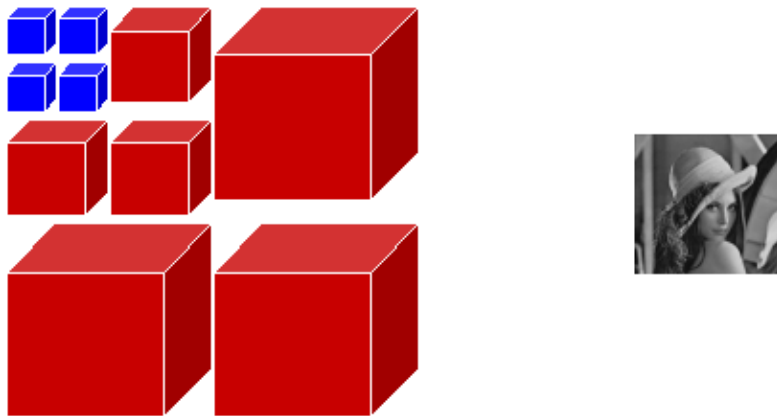


Figure 2: Resolution with three blocks decoded

2.3 Consequences for the Required Storage Capacity

Based on various test sets¹⁸, it appears that JPEG 2000 in lossless mode can yield a benefit of about 50% compared to an uncompressed file.

Based on the test material, it appears that the gain that can be achieved with JPEG 2000 part 1 *lossy* compression can vary – assuming the Lead Photoshop plugin compression ratio settings between 10 and 50 – between 91% and 98%¹⁹.

2.4 Consequences for the Image Quality

The lossless mode has no consequences for the image quality.

Lossy versions:

The quantity of compressions does degrade the image. Five versions were tested by means of the Lead JPEG 2000 Photoshop plugin (compression ratio): 25, 75, 100 and 500.

Detailed Loss – MTF

Original TIFF (QA-62 SFR and OECF test chart): MTF 5.91/5.91. File size 4.7 MB

Compression Ratio	MTF Horizontal and Vertical (three RGB channels on average)	File Size

¹⁸ See Appendix 4.

¹⁹ The Lead Photoshop plugin might not give optimal compression results. An alternative test with the native JPEG2000 plugin proved to give no great differences though. Lossless compression of the Photoshop plugin proved to be slightly less successful than that of the LEAD plugin: 53% for the former, versus 52% for the latter. Further testing of alternatives (like the Lurawave clt command compression tool <http://www.luratech.com/products/lurawave/jp2/clt/>) will have to be performed.

25	5.8 / 5.8	83 KB
75	5.8 / 5.8	62 KB
100	5.8 / 5.8	47 KB
500	3.9 / 3.1	10 KB

Greyscale and Colour Loss

There is no measurable loss of greyscale in Kodak Greyscale.

Delta E values remain the same with various compression values and no extra colour shift occurs.

Artefacts

In JPEG 2000 files, three clearly visible artefacts appear when the compression increases (tested based on various types of materials):

1. Posterizing or banding (coarse transitions in colour or grey tones). Clearly visible starting at an approximate compression ratio of 75 for text materials. For continuous tone images artefacts are becoming slightly visible at compression ratio 100, and well visible at 200..
2. Tiling effect: The tiles only become visible with extreme compression (compression ratio 200). This could be prevented by choosing a tile that has the same dimension of the image itself.
3. Woolly effect around elements rich in contrast. Visible starting at an approximate compression ratio of 75.

The last effect is particularly visible in text (around the letters). Continuous tone originals such as photos and paintings appear to be more suitable for strong JPEG 2000 compression than text materials do (or other materials with high-contrast transitions such as line drawings).

PSNR

Topic of investigation.

2.5 Consequences for the Long-Term Sustainability

In order to be able to make an accurate comparison between the JP2 format and the other formats that are either lossless compressed (“PNG 1.2” and “TIFF 6.0 with LZW compression”) or lossy compressed (“basic JFIF (JPEG) 1.02”), we divide the JP2 format into “JP2 (JPEG 2000 Part 1) lossless” and “JP2 (JPEG 2000 Part 1) lossy.”

The application of the “File Format Assessment Method” to the “JP2 (JPEG 2000 Part 1) lossless” format results in a score of 74,7 on a scale of 0-100. For the lossy compressed version, the method results in a score of 66,1. When the four formats that are compared in this report are sorted from most to least suitable for long-term storage according to the named method, “JP2 (JPEG 2000 Part 1) lossless” ends up in second place with this score, right after PNG 1.2 (with a score of 78). We then come to a point where the applied method possibly comes up 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 Adoption. The expectation is that – although JPEG 2000 and PNG are currently not used as master files to a large extent – JPEG 2000 does have potential as a master file. PNG has existed since 1996 and JP2 only since 2000.

“JP2 (JPEG 2000 Part 1) lossy” ends up in third place right above “basic JFIF (JPEG) 1.02” with almost the same score. For both the lossless as well as the lossy versions of JPEG 2000,

the score is primarily low due to the low adoption rate of the format. Adoption is a very important factor in the method. Despite the almost equal scores of “basic JFIF (JPEG) 1.02” and the lossy version of JPEG 2000, there is still a preference for “basic JFIF (JPEG) 1.02” due to the more certain future of this file. A report on the usage of JPEG 2000 as a Practical Digital Preservation Standard has recently been published on the DPC website²⁰.

2.6 Consequences for the Functionality

- Options for including bibliographic and technical (EXIF) metadata
 - Bibliographic metadata: It is possible to add metadata in three boxes: One for XML data, a limited IPR (Intellectual Property Rights) box²¹ and UUID (Universal Unique Identifier) based on ISO 11578:1996.
 - Technical metadata: There is as yet no standard manner for storing EXIF metadata in the JPEG 2000 header. Suggestions have been made to do this in an UUID box.²²
- Suitability of the format for offering it as a high-resolution access master
 - Browser support: very limited (only by Apple’s Safari browser)²³.
 - High-resolution image access: Because browsers do not yet support JPEG 2000, a JPEG generated on-the-fly is typically used as an intermediary image.
- Maximum size

Image dimensions width and height can be up to $(2^{32})-1$. File size can be unlimited with a special setup (take code stream box to the end of the file and signal “unknown” length). File format boxes can signal a length up to $2^{64}-1$ bytes = 16 million TB. These are of course theoretical file sizes as no existing program will support them²⁴.

LOC Quality and Functionality Factors: ²⁵

- Normal display
 - Screen display: Yes
 - Printable: Yes
 - Zoomable: Yes
- Clarity
 - High-resolution options: Yes. A lot of compression can damage detailing (see section 3.4 above).
 - Bit depths: The JPEG 2000 Part 1 core file can vary from 1 bit to 38 bits.²⁶

²⁰ Robert Buckley, JPEG 2000 – a Practical Digital Preservation Standard?, a DPC Technology Watch Series Report 08-01, February 2008: <http://www.dpconline.org/graphics/reports/index.html#jpeg2000>

²¹ This option is greatly expanded in Part 8 of JPEG 2000 standard.

²² Wikipedia, JPEG 2000, http://en.wikipedia.org/wiki/JPEG_2000. The Adobe XML based XMP standard – which makes use of UUID box - seems to provide a standard way of storing EXIF information in the header. <http://www.pctoday.com/editorial/article.asp?article=articles%2F2005%2F0304%2F44t04%2F44t04web.asp>
http://en.wikipedia.org/wiki/Extensible_Metadata_Platform

²³ <http://echoone.com/filejuicer/formats/jp2>

²⁴ Klaus Jung, email 13 february 2008 to Judith Rog

²⁵ http://www.digitalpreservation.gov/formats/content/still_quality.shtml.

²⁶ Buckley, *JPEG 2000 Image Archiving*, p. 45, slide 90. In Part 4, three different compliance classes can be indicated. Class 2 limits these options to 16 bits.

- Colour maintenance
 - Support of various colour spaces: Yes (though not via ICC profile).
 - Option for including gamma correction: No.
 - Options for including ICC colour profiles: JPEG 2000 part 1 offers the standard option of sRGB, greyscale and YCC. As an alternative, a limited form of ICC colour profiles²⁷ can be provided.²⁸
- Support of graphic effects and typography.
 - Vector image options: No.
 - Transparency information: Yes.
 - Option to specify fonts and styles: No.
- Functionality beyond normal display
 - Animation: No (this option is offered in JPEG 2000 Part 3 and 12).
 - Multipage support: No (this option is offered in JPEG 2000 Part 6).
 - Multiresolution: Yes. There is also an option to construct the image proceeding from colour, quality or position.

2.7 Conclusion

Format Description

- Standardization: JPEG 2000 Part 1 has been standardized since 2000 ISO/IEC. Other parts were standardized later or are not yet fully ISO standardized.
- 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 resolution, quality, colour or position.

Consequences for Storage Capacity

- Lossless: Storage gain is approximately 50%.
- Lossy: Storage gain is variable between 91% and 98%.

Consequences for Image Quality

- Lossless: None.
- Lossy:
 - Some loss of details while using strong compression.
 - No loss of greyscale/colour.
 - Artefacts: Posterizing, pixelation, woolly effect around elements that are rich in contrast with a large amount of compression.
 - PSNR: Currently being investigated.

²⁷ Definition according to the ICC Profile Format Specification ICC.1:1998-09

²⁸ Florida Digital Archive description of JP 2 – section 1.8:

http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgrounds/jp2_bg.pdf

Consequences for the Long-Term Sustainability

- Lossless: File Format Assessment Method score 74,7.
- Lossy: File Format Assessment Method score 66,1.
- Main problem: Low adoption rate.

Consequences for the Functionality

The most important advantages:

- Possibility of lossless and variable lossy compression.
- Very effective wavelet compression.
- Very comprehensive multiresolution options: It is possible to create the image based on quality, resolution, colour and position.
- Comprehensive metadata possibilities.
- Options for very diverse bit depths (1 to 38 bits).

The most important disadvantages:

- Low adoption rate on the consumer market.
- Low adoption rate in software support (image editing and viewing software).
- No browser support (other than through server side software that generates JPEG files on-the-fly JPEG).
- Compressing and decompressing takes a relatively large amount of computing power.
- No standard option for adding EXIF metadata.

Recommendation

Reason 1: Substitution

JPEG 2000 Part 1 lossless is a good alternative from the perspective of long-term sustainability. The most effective lossless compression (50%), no loss of image quality and the flexible nature of the file format (particularly due to the wealth of multiresolution options) are an extra argument that speaks in favour of JPEG 2000 lossless. The only real long-term worry is the low rate of adoption.

Due to the irreversible loss of image information, JPEG 2000 Part 1 lossy is a much less obvious choice for substitution. The creation of visual lossless images might be considered (i.e., images that cannot visually be differentiated from the original uncompressed file) (storage gain of approximately 90%). 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.

Reason 2: Redigitisation Is Not Desirable

In this case JPEG 2000 Part 1 lossy, in the visual lossless mode, is a viable option. The small amount of information loss can be defended more easily in this case because there is no substitution.

Reason 3: Master file is access file

In this case JPEG 2000 Part 1 lossy with a larger degree of compression is self-evident. 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 functionality options of JPEG 2000 outweighs the JPEG format for this purpose, which is comprehensively supported by software (including browsers) and is widely distributed.

3 PNG

3.1 What Is PNG?

PNG (Portable Network Graphics) is a datastream and an associated file format for a lossless compressed, portable, individual raster image²⁹ which was initially developed for transmission via the Internet. A large group of developers (PNG development group) began developing the format in 1995 under the supervision of the World Wide Web Consortium (W3C) as an alternative for the then-patented GIF format and associated LZW compression. The first official version, 1.0, came into existence in 1997 as a W3C Recommendation. PNG version 1.2 was revealed in 1999, and this version has been ISO standardized (ISO/IEC 15948:2003) since 2003, with the specifications being freely available via the W3C: <http://www.w3.org/TR/PNG/>³⁰.

The objectives of the developers of PNG were as follows³¹:

- a. Portability: Encoding, decoding, and transmission should be software and hardware platform independent.
- b. Completeness: It should be possible to represent true colour, indexed-colour, and greyscale images, in each case with the option of transparency, colour space information, and ancillary information such as textual comments.
- c. Serial encode and decode: It should be possible for datastreams to be generated serially and read serially, allowing the datastream format to be used for on-the-fly generation and display of images across a serial communication channel.
- d. Progressive presentation: It should be possible to transmit datastreams so that an approximation of the whole image can be presented initially, and progressively enhanced as the datastream is received.
- e. Robustness to transmission errors: It should be possible to detect datastream transmission errors reliably.
- f. Losslessness: Filtering and compression should preserve all information.
- g. Performance: Any filtering, compression, and progressive image presentation should be aimed at efficient decoding and presentation. Fast encoding is a less important goal than fast decoding. Decoding speed may be achieved at the expense of encoding speed.
- h. Compression: Images should be compressed effectively, consistent with the other design goals.
- i. Simplicity: Developers should be able to implement the standard easily.
- j. Interoperability: Any standard-conforming PNG decoder shall be capable of reading all conforming PNG datastreams.
- k. Flexibility: Future extensions and private additions should be allowed for without compromising the interoperability of standard PNG datastreams.
- l. Freedom from legal restrictions: No algorithms should be used that are not freely available.

²⁹ In contrast to the GIF format, PNG does not offer any animation options (animated GIF). The separate MNG format was created for animation objectives: <http://www.libpng.org/pub/mng/>.

³⁰ What is strange is that the ISO itself mentions the year 2004:

http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=29581.

³¹ These objectives are listed in the W3C specifications under the “Introduction” header: <http://www.w3.org/TR/PNG/> henceforth: *PNG specs*.

These objectives have been achieved in the last PNG standard.

3.2 How does it work?

3.2.1 Structure

The PNG datastream consists of a PNG signature³² which indicates that it is a PNG datastream, followed by a sequence of chunks (meaning a “component”). Each chunk has a chunk type that specifies the goal. A certain number of chunks is mandatory (*critical*), and a large part is inessential (*ancillary*). This chunk structure was developed with the idea of keeping the format expandable while simultaneously being backwards compatible.

3.2.2 Encoding and Decoding/Filtering and Compression

Encoding takes place in six steps³³:

1. Pass extraction: To allow for progressive display, the PNG image pixels can be rearranged to form several smaller images called reduced images or passes.
2. Scanline serialization: The image is serialized one scanline at a time. Pixels are ordered left to right in a scanline and scanlines are ordered top to bottom.
3. Filtering: Each scanline is transformed into a filtered scanline using one of the defined filter types to prepare the scanline for image compression.
4. Compression: Occurs on all the filtered scanlines in the image.
5. Chunking: The compressed image is divided into conveniently sized chunks. An error detection code is added to each chunk.
6. Datastream construction: The chunks are inserted into the datastream.

Only filtering and compression are described below.

Prior to compression, compression filters are used that order the bytes per scanline. A different filter can be used per scanline. This greatly increases the success of the compression. The PNG compression algorithm uses the lossless, unpatented inflate/deflate method (zlib/gzlib).³⁴

The success of the compression depends on the correct and complete implementation of the PNG encoding options. It can be useful to research software tools for limiting PNG file sizes.³⁵

3.3 Consequences for the Required Storage Capacity

Based on test sets, it appears that PNG in lossless mode can yield a benefit of about 40% compared to an uncompressed file. Further tests with more refined compression options must prove whether or not this can still be optimized.

³² See section 5.2 of the *PNG specs* and Wikipedia:

http://en.wikipedia.org/wiki/Portable_Network_Graphics#File_header.

³³ *PNG specs*, section 4.5.1: <http://www.w3.org/TR/PNG/#4Concepts.EncodingIntro>.

³⁴ See chapters 9 and 10 of the *PNG specs* and Wikipedia:

http://en.wikipedia.org/wiki/Portable_Network_Graphics#Compression.

³⁵ Wikipedia PNG lemma gives recommendations for various tools

http://en.wikipedia.org/wiki/Portable_Network_Graphics#File_size_and_optimization_software.

3.4 Consequences for the Image Quality

Because PNG filtering and compression is lossless there is no degradation of the image quality. However, the assumption is that the bit depth remains the same as that of the source file. Decrease of the bit depth – an option that the PNG format offers – must be viewed as a form of *lossy* compression.

3.5 Consequences for the Long-Term Sustainability

Applying the “File Format Assessment Method” to the “PNG 1.2” format results in a score 78 on a scale of 0-100. When the four formats that are compared in this report are sorted from most to least suitable for long-term storage according to the named method, “PNG 1.2” ends up in first place with this score, directly ahead of “JP2 (JPEG 2000 Part 1) lossless.” In the PNG case, too, the low adoption rate of the format has a negative effect on the final score. As mentioned earlier in section 1.5 (“Consequences for the Long-Term Sustainability” for JPEG 2000), despite the fact that PNG scores four point higher than JP2 (JPEG 2000 Part 1) lossless” we still prefer the latter on account of this format's better future outlook as regards adoption.

3.6 Consequences for the Functionality

- Options for including bibliographic and technical (EXIF) metadata
 - Bibliographic metadata: PNG offers the option of including content metadata in both ASCII as well as UTF-8 and offers a number of standard options (Title, Author, Description, Copyright, Creation Time, Software, Disclaimer, Warning, Source, Comment). It is possible to expand this set according to your own wishes.³⁶
 - Technical metadata: PNG does not (yet) support EXIF information (technical metadata that provides information about the camera and camera settings).
- Suitability of the format for offering it as a high-resolution access master
 - Browser support: Yes.
 - High-resolution image access: In theory, yes. Through lossless compression, PNG remains relatively large for this objective.
- Maximum size
 - Topic of investigation.

LOC Quality and Functionality Factors:³⁷

- Normal display
 - Screen display
 - Printable: Yes.
 - Zoomable: Yes.
- Clarity
 - High-resolution options: Yes.
 - Bit depths: Can vary from 1 to 16 bits per channel.
- Colour maintenance

³⁶ See section 11.3.4.2 of the *PNG specs*.

³⁷ http://www.digitalpreservation.gov/formats/content/still_quality.shtml.

- Support of various colour spaces: Yes (though not via ICC profile).
- Option for including gamma (brightness) correction: Yes (also chroma – colour saturation - correction).
- Options for including ICC colour profiles: PNG offers the option of using the sRGB colour space and including ICC colour profiles.³⁸
- Support of graphic effects and typography.
 - Vector image options: No.
 - Transparency information: Yes.
 - Option to specify fonts and styles: No.
- Functionality beyond normal display
 - Animation: No³⁹
 - Multipage support: No.
 - Multiresolution: No.

3.7 Conclusion

Format Description

- 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).

Consequences for Storage Capacity

- Storage gain is approximately 40%.

Consequences for Image Quality

- Lossless, so none.

Consequences for the Long-Term Sustainability

- File Format Assessment Method score 78.
- Main problem: Low adoption rate.

Consequences for the Functionality

The most important advantages:

- Lossless compression.
- Comprehensive support by image editing and viewer software and browsers.
- Comprehensive metadata possibilities.
- Options for very diverse bit depths (1 to 16 bits per channel).
- Comprehensive options for transparency.

The most important disadvantages:

³⁸See section 4.2 of the *PNG specs*.

³⁹ The related MNG format offers this option: <http://www.libpng.org/pub/mng/>

- No option for lossy compression (other than by decreasing the bit depth), so images remain relatively large.
- No multiresolution options.
- No standard option for adding EXIF metadata.

Recommendation

Reason 1: Substitution

PNG Part 1 lossless is a possible alternative from the perspective of long-term sustainability. Lossless compression is ideal for substitution objectives because no image information is lost. The compression is somewhat less effective than that of JPEG 2000 Part 1 lossless (40% versus 50%). The comprehensive software support is a plus but the low level of actual use (both on the consumer market as well as in the cultural heritage sector) is worrisome.

Reason 2: Redigitisation Is Not Desirable

PNG is also suitable for this goal, although the less effective, lossless compression is a minus.

Reason 3: Master file is access file

In this case, PNG is a less obvious choice due to the lack of a lossy compression option (and thus more storage gain).

4 JPEG

4.1 What is JPEG?

First and foremost, JPEG (Joint Photographic Experts Group) stands for the committee that was established to create a standard for the compression of continuous tone greyscale and colour images (as the name indicates).⁴⁰ The committee started this task in 1986, and in 1992 the first version of this standard was ready, which in 1994 was standardized as ISO 10918-1 and as ITU-T Recommendation T.81. The JPEG committee is also at the basis of the JBIG format (bitonal compression format) and the JPEG 2000 format.

The JPEG standard is more than a description of a file format: It both specifies the codec with which the images are compressed/encoded in a datastream as well as the file format that this datastream contains.

The JPEG standard consists of four parts⁴¹:

- Part 1 - The basic JPEG standard, which defines many options and alternatives for the coding of still images of photographic quality.
- Part 2 - Sets rules and checks for making sure software conforms to Part 1.
- Part 3 - Set up to add a set of extensions to improve the standard, including the SPIFF file format.
- Part 4 - Defines methods for registering some of the parameters used to extend JPEG.

The description of the JPEG file format – the *JPEG Interchange Format* – is included as annex B of the 10981-1 standard. The confusing part is that a stripped, or real-world, version⁴² of this description, JFIF (*JPEG File Interchange Format*), has become the de-facto standard with which applications work and which is generally designated as JPEG⁴³. JFIF simplified a number of things in the standard – among them a standard colour space – and thus made the JPEG Interchange Format usable for a range of applications and uses.

The following discusses the JFIF standard, which will be designated as JPEG.⁴⁴

⁴⁰ To cite the group: “This Experts Group was formed in 1986 to establish a standard for the sequential progressive encoding of continuous tone greyscale and colour images.” CCITT T.81 *Information Technology – Digital compression and coding of continuous-tone still images – requirements and guidelines* p. 1.

<http://www.w3.org/Graphics/JPEG/itu-t81.pdf>.

⁴¹ <http://www.jpeg.org/jpeg/index.html>.

⁴² <http://www.jpeg.org/jpeg/index.html>. The JFIF format was developed by Eric Hamilton of C-Cube Microsystems.

⁴³ JFIF standard: <http://www.jpeg.org/public/jfif.pdf>. The “real world” terminology was coined by the JPEG committee itself: “As well as the standard we created, nearly all of its real world applications require a file format, and example reference software to help implementors”. <http://www.jpeg.org/jpeg/index.html>.

⁴⁴ Five other extensions of the standard are worth mentioning:

- JPEG_EXIF (most recent version 2.2). This is an extension of the JPEG standard (based on the baseline JPEG) that is used en masse in digital cameras. EXIF information contains technical metadata about the camera and camera settings.
- Adobe JPEG. The version of JPEG as used by Adobe applications does comply with the JPEG standard but not with JFIF. An important difference with the JFIF standard is in the fact that Adobe can save JPEG files in the CMYK colour space. This version of JPEG is *not* publicly documented. Florida Digital Archive description of JFIF (p. 5, 6): http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgrounds/jfif.pdf.

4.2 How does it work?

4.2.1 Structure

Topic of investigation.

4.2.2 Encoding and Decoding/Filtering and Compression

Encoding (assuming a 24-bit RGB file) takes place in four steps⁴⁵:

1. Conversion of the RGB colour space of the source file to the YCbCr colour space (Y is the brightness component, Cb and Cr two colour or chroma components, blue and red).⁴⁶
2. The resolution of the colour data is decreased (also: downsampling or chroma subsampling), mostly with a factor of two. This is based on the fact that the human eye sees more details in the brightness component Y than in the colour components, Cb and Cr. This can already yield a 33 to 50% gain compared to the source file and is a lossy process.
3. The image is divided into 8 x 8 pixels (block splitting). For each block, for each of the Y, Cb and Cr components, so-called discrete cosine transformation (DCT) is applied (a break/conversion of the pixel values per component to a frequency-domain representation).
4. The amplitudes of the frequency components are quantified. Because the eye is less sensitive to high-frequency variations in brightness (than for small changes in colour or brightness in large/wide areas), high-frequency components are therefore stored with a lower degree of accuracy. The quality setting of the encoder determines how much of the high-frequency information is lost. In the case of extreme compression, this information is completely left out.
5. Ultimately, the 8 x 8 blocks are compressed even more by means of a lossless algorithm (a version of Huffman encoding).

Decoding simply runs in the opposite direction.

The most noticeable JPEG artefact – the pixelation – occurs in the quantifying step.

-
- SPIFF based on Annex F of the ISO10918-1 standard. This is an extension of the JPEG standard (Part 3) and is intended as the successor of the limited JFIF standard (JFIF inventor Eric Hamilton played an important role in the development of SPIFF) which can contain both the JPEG-DCT compression as well as JBIG bitonal compression. However, this appears to be rarely used. See: <http://www.fileformat.info/format/spiff/egff.htm> en <http://www.digitalpreservation.gov/formats/fdd/fdd000019.shtml>.
 - Lossless JPEG. Two formats of the lossless JPEG have been developed:
 - Lossless-JPEG (1993) as an extension of the JPEG standard, which uses a completely different compression technique. This has not gained popularity, other than in some medical applications.
 - JPEG-LS (1999). A “near-lossless” format that offers better compression than the lossless JPEG format and is much less complex than the lossless JPEG 2000 version. This also appears to have hardly gained any popularity.

Wikipedia: http://en.wikipedia.org/wiki/Lossless_JPEG. JPEG-LS homepage: <http://www.jpeg.org/jpeg/jpegls.html>.

⁴⁵ Wikipedia: <http://en.wikipedia.org/wiki/JPEG#Encoding>. The encoding described here is the most used method. This is thus not the *only* method.

⁴⁶ This step is sometimes skipped. This is the case for a high-quality JPEG, whereby the file is stored in sRGB “where each colour plane is compressed and quantized separately with similar quality steps.” Wikipedia http://en.wikipedia.org/wiki/JPEG#Color_space_transformation.

4.3 Consequences for the Required Storage Capacity

Based on the test material, it appears that the gain that can be achieved with JPEG compression can vary – assuming Adobe Photoshop compression values JPEG 10 to JPEG PSD 1 – between 90 and 98%.

4.4 Consequences for the Image Quality

Five variations of JPEG compression were tested in Photoshop (scale 0-12, designated as *PSD*): PSD 0, 3, 5, 8 and 10. 0 and 3 are extreme compression, 5 is average, 8 and 10 are slight.

Detailed loss – MTF

Original TIFF (QA-62 SFR and OECF test chart): MTF 5.91/5.91. File size 4.7 MB

Compression Ratio	MTF Horizontal and Vertical (three RGB channels on average)	File Size
JPEG PSD 10	5.9 / 5.8	204 KB
JPEG PSD 8	5.4 / 5.2	128 KB
JPEG PSD 5	4.9 / 4.8	84 KB
JPEG PSD 3	4.3 / 4.2	64 KB
JPEG PSD 0	3.8 / 3.5	57 KB

Greyscale and Colour Loss

No measurable loss of greyscale in Kodak Gray.

The delta E values remain the same at various compression values and no extra colour shift occurs (in the contrary = the RGB values are drawn to each other to one value)⁴⁷.

Artefacts

In JPEG files, three clearly visible artefacts appear the more the compression increases (tested based on various types of materials):

1. Posterizing or banding (coarse transitions in colour or greyscale). Somewhat visible starting at JPEG PSD 7/8. Clearly visible starting approximately at JPEG PSD 5.
2. Pixelation: Visible starting approximately at JPEG PSD 2.
3. Woolly effect around elements rich in contrast. Visible starting approximately at JPEG PSD 4.

The last effect is particularly visible in text (around the letters). Continuous tone originals such as photos and paintings appear to be more suitable for strong JPEG compression than text materials do (or other materials with high-contrast transitions such as line drawings).

PSNR

Topic of investigation.

⁴⁷ This is why delta E might not be a good tool for measuring colour differences between the compressed and uncompressed file. Colour differences *do* occur in distorting subtle colour changes (see “artefacts”).

Consequences of Repeated Compression

The image degrades when it is compressed several times. Tests have shown that degradation when applying JPEG PSD 10 compression doesn't really become visible until compression has been executed four times.

4.5 Consequences for the Long-Term Sustainability

Application of the "File Format Assessment Method" to the "basic JFIF (JPEG) 1.02" format results in a score of 65,4 on a scale of 0-100. When the four formats that are compared in this report are sorted from most to least suitable for long-term storage according to the named method, "basic JFIF (JPEG) 1.02" ends up in third place with this score, not much ahead of or almost equal with "TIFF 6.0 with LZW compression" with a score of 65,3 and just beneath "JP2 (JPEG 2000 Part 1) lossy," which scores 66,1 points. The lossy form of the compression and the fact that the format is little used as a master format in the cultural heritage sector both play an important role in the final score of the format. If a choice has to be made between "JP2 (JPEG 2000 Part 1) lossy" and "basic JFIF (JPEG) 1.02," preference is given to the latter due to the more certain future of this file.

4.6 Consequences for the Functionality

- Options for including bibliographic and technical (EXIF) metadata
 - Content-related metadata: Yes.
 - Technical metadata: The separate JPEG EXIF format was developed for the inclusion of EXIF information (see note 35).
- Suitability of the format for offering it as a high-resolution access master
 - Browser support: JPEG is supported by all standard browsers.
 - High-resolution image access: Often the high-resolution JPEG is used as a zoom file. This is done by creating separate resolution layers, as separate images. Sometimes these images again parted into tiles.⁴⁸
- Maximum size
 - Topic of investigation.

LOC Quality and Functionality Factors: ⁴⁹

- Normal display
 - Screen display: Yes.
 - Printable: Yes.
 - Zoomable: Yes.
- Clarity
 - High-resolution options: Yes. A lot of compression can damage detailing (see section 3.4 above).
 - Bit depths: Limited to 8 and 24 bits.⁵⁰

⁴⁸ See the Geheugen van Nederland (memory of the Netherlands) (<http://www.geheugenvannederland.nl/>) for the first and the solution by the image database of the Amsterdam City Archive (<http://beeldbank.amsterdam.nl/>) for the second.

⁴⁹ http://www.digitalpreservation.gov/formats/content/still_quality.shtml.

- Colour maintenance
 - Support of various colour spaces: Yes (though not via ICC profile).
 - Option for including gamma correction: No.
 - Options for including ICC colour profiles: Yes⁵¹
- Support of graphic effects and typography.
 - Vector image options: No.
 - Transparency information: Yes.
 - Option to specify fonts and styles: No.
- Functionality beyond normal display
 - Animation: No.
 - Multipage support: No.
 - Multiresolution: More or less. It is possible to store thumbnails with larger images⁵². However, this function is not or rarely supported by image editing and viewer software.

4.7 Conclusion

Format Description

- 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.

Consequences for Storage Capacity

- Storage gain is variable between approximately 89% and 96%.

Consequences for Image Quality

- Gradual loss of detail with increased compression.
- No measurable loss of greyscale/colour.
- Artefacts: Visible posterizing, pixelation, woolly effect around elements that are rich in contrast with a large amount of compression.
- PSNR: Topic of investigation.

Consequences for the Long-Term Sustainability

⁵⁰ A 12-bit JPEG is used in some medical applications. The 12-bit JPEG is a part of the JPEG standard but is rarely used and supported. Wikipedia JPEG:

http://en.wikipedia.org/wiki/JPEG#Medical_imaging:_JPEG.27s_12-bit_mode.

⁵¹ ICC profile 4.2.0.0. LOC description.

<http://www.digitalpreservation.gov/formats/fdd/fdd000018.shtml#factors>.

⁵² Starting with version 1.02 LOC description JFIF

<http://www.digitalpreservation.gov/formats/fdd/fdd000018.shtml#factors>.

- File Format Assessment Method score 65,4.
- Main problems: Lossy compression and slight use as a master format in the cultural heritage sector.

Consequences for the Functionality

The most important advantages:

- Comprehensive support by image editing and viewer software and browsers.
- Compression and decompression requires little computing power.
- Efficient, variable DCT compression.
- Standardized method for accommodating EXIF metadata (in JPEG EXIF format).

The most important disadvantages:

- No options for lossless compression.
- Limited bit depth options (8 bits greyscale, 24 bits colour).
- No multiresolution options.

Recommendation

Reason 1: Substitution

JPEG is not the most obvious file format choice for substitution purposes. In particular the irreversible loss of image information is not desirable in view of long-term storage. The relatively low File Format Assessment Method score (66) stems from this fact. One option to consider could be the creation of visual lossless images – JPEG PSD 10 and higher (storage gain approx. 89%). 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.

Reason 2: Redigitisation Is Not Desirable

In this case a visual lossless JPEG is a viable option. The small amount of information loss can be defended more easily in this case because there is no substitution. The comprehensive distribution and support of JPEG is an extra argument that speaks in favour of this file format.

Reason 3: Master file is Acces File

In this case JPEG with a larger degree of compression is self-evident. The JPEG compression technique enables a rather large decrease in storage without much loss of quality. When selecting the amount of compression, the type of material must be taken into account. Compression artefacts in text files will be visible before those in continuous tone originals such as photos, for example.

5 TIFF LZW

5.1 What is TIFF LZW?

Strictly speaking, TIFF LZW is not a separate file format. TIFF (Tagged Image File Format) 6.0 is the file format, LZW (Lempel-Ziv-Welch, the names of the developers) is the compression algorithm that is used *within* TIFF (in addition to LZW compression, TIFF offers the option of using ITU_G4, JPEG and ZIP compression). The following provides a brief description of the TIFF 6.0 format, with a more detailed discussion of the LZW compression method.

The first version of the TIFF specification (developed by Microsoft and Aldus, with the last version currently being a part of Adobe) appeared in 1986 and unofficially is called version 3.0. Version 4.0 was launched in 1987 and version 5.0 in 1988. The latter offered options for limited colour space (palette colour) and LZW compression. The baseline TIFF 6.0 standard dates from 1992, which included CYMK colour definition and the use of JPEG compression, among other things. Version 6.0 was followed by various extensions (see section 4.2.1 below) – the most important ones being: TIFF/EP (2001), TIFF/IT (2004), DNG (2005) and EXIF.⁵³

The baseline TIFF 6.0 is not ISO-IEC standardized.

The objective was to create a file format to store raster images originating from scanners and image editing software. The main objective “is to provide a rich environment within which applications can exchange image data. This richness is required to take advantage of the varying capabilities of scanners and other imaging devices”⁵⁴. The standard must also be expandable based on new imaging requirements: “A high priority has been given to structuring TIFF so that future enhancements can be added without causing unnecessary hardship to developers”⁵⁵. This option has been abundantly used. The disadvantage of this is that not all extensions are used by all image editing and viewer software.

The LZW compression algorithm dates from 1984 and is basically an improved version of the LZ78 algorithm from 1978. The name gives Jacob Ziv and Abraham Zempel developed the LZ78 format, and Terry Welch developed the faster, improved LZW. It was developed as a lossless data (thus not only for images) compression algorithm. In addition to being used in TIFF, LZW became famous largely due to its use in the GIF format. In addition, LZW is notorious due to the patent that Unisys claimed to have on the algorithm (via developer Terry

⁵³ TIFF/EP extension (ISO 12234-2) for digital photography (http://en.wikipedia.org/wiki/ISO_12234-2)
TIFF/IT (ISO 12369) extension for prepress purposes
(<http://www.digitalpreservation.gov/formats/fdd/fdd000072.shtml>).
DNG Adobe TIFF UNC extension for storing RAW images
(<http://www.digitalpreservation.gov/formats/fdd/fdd000188.shtml>).
EXIF technical metadata of cameras and camera settings
(<http://www.digitalpreservation.gov/formats/fdd/fdd000145.shtml>).

⁵⁴ TIFF Revision 6.0 June 1992. p. 4. Scope. <http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf>.

⁵⁵ Ibidem.

Welch⁵⁶). This patent expired in 2003 (US) and 2004 (Europe and Japan), although Unisys still claims to possess certain improvements to the algorithm.⁵⁷

5.2 How does it work?

5.2.1 Structure

The TIFF file begins with an 8-byte image file header (IFH) that refers to the image file directory (IFD) with the associated bitmap. The IFD contains information about the image in addition to pointers to the actual image data.⁵⁸

The TIFF tags, which are contained in the header and in the IFDs, contain basic geometric information, the manner in which the image data are organized and whether a compression scheme is used, for example. An important part of the tags belongs to the so-called baseline TIFF.⁵⁹ All tags outside of this are extended and contain things such as alternative colour spaces (CMYK and CIE Lab) and various compression schemes.⁶⁰

There are also tags called *private tags*. The TIFF 6.0 version offers users the option to use their own tags (and also to develop them through private IFDs⁶¹), and this is done quite a lot. The above-mentioned TIFF/EP, TIFF/IT make use of this option. Because the used tags are public, there is talk of open extensions. The LOC documentation contains a valuable overview⁶² of this extension:

http://www.digitalpreservation.gov/formats/content/tiff_tags.shtml.

5.2.2 Encoding and Decoding/Filtering and Compression

Topic of investigation.

5.3 Consequences for the Required Storage Capacity

Based on test sets, it appears that TIFF LZW in lossless mode can yield a benefit of about 30% compared to an uncompressed file.

5.4 Consequences for the Image Quality

Because LZW compression is lossless there is no degradation of the image quality.

5.5 Consequences for the Long-Term Sustainability

Applying the “File Format Assessment Method” to the “TIFF 6.0 with LZW compression” format results in a score 65,3 on a scale of 0-100. When the four formats that are compared in this report are sorted from most to least suitable for long-term storage according to the named method, “TIFF 6.0 with LZW compression” ends up in last place with this score, not far behind or almost equal with “basic JFIF (JPEG) 1.02,” with a score of 65,4.

⁵⁶ As an employee of the Sperry Corporation, Welch developed the algorithm, and that is what the patent was initially based on. Sperry Corporation later became a part of Unisys.

http://en.wikipedia.org/wiki/Graphics_Interchange_Format#Unisys_and_LZW_patent_enforcement.

⁵⁷ http://www.unisys.com/about_unisys/lzw/.

⁵⁸ A TIFF can contain several IFDs – this is then a multipage TIFF (not a baseline TIFF).

⁵⁹ Part 1 from the TIFF 6.0 specs: <http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf>.

⁶⁰ Part 2 from the TIFF 6.0 specs: *ibidem*.

⁶¹ The EXIF extension makes use of this option:

http://www.digitalpreservation.gov/formats/content/tiff_tags.shtml.

⁶² Which is strangely enough not maintained by Adobe itself.

This low score primarily stems from the possible patents that still exist on the LZW compression method (see http://www.unisys.com/about_unisys/lzw/) and the resulting low rate of adoption of this version of TIFF as a master archive format in the cultural sector. The patents that Unisys still claims to hold are different from the ones that were often referred to in the past and expired in 2003/2004. When we used the same evaluation method to assess a baseline TIFF 6.0, we see a much higher score because LZW compression is not used in this version. Therefore, from the perspective of long-term sustainability use of TIFF 6.0 with LZW compression is discouraged

5.6 Consequences for the Functionality

- Options for including bibliographic and technical (EXIF) metadata
 - Content-related metadata: Yes.
 - Technical metadata (EXIF): Yes.
- Suitability of the format for offering it as a high-resolution access master
 - Browser support: No.
 - High-resolution image access: TIFF LZW is very limited when it comes to exchangeability of high-resolution images via the Web. Because the format compresses in a lossless manner the files remain relatively large. TIFF is also not supported by browsers. JPEG thus becomes the more obvious choice.
- Maximum size
 - File size: 4 GB. There are proposals to enlarge this to 20 GB (BigTIFF)⁶³

LOC Quality and Functionality Factors: ⁶⁴

- Normal display
 - Screen display: Yes.
 - Printable: Yes.
 - Zoomable: Yes.
- Clarity
 - High-resolution options: Yes.
 - Bit depths: The TIFF 6.0 standard offers the options of 1 bit, 4 bits, 8 bits, 16 bits (and theoretically even 32 bits) per channel.
- Colour maintenance
 - Support of various colour spaces: Yes (though not via ICC profile). Standard: Bitonal, greyscale, RGB, CMYK, YCbCR, CIEL*a*b
 - Option for including gamma correction: No.
 - Options for including ICC colour profiles: Yes. ICC colour profiles can be included, although there does not appear to be a standard way for this. The TIFF/EP and TIFF/IT standards developed private tags that can also be

⁶³ <http://www.awaresystems.be/imaging/tiff/bigtiff.html>

Photoshop should be possible to open the 4 GB file

<http://kb.adobe.com/selfservice/viewContent.do?externalId=320005&sliceId=1>

⁶⁴ http://www.digitalpreservation.gov/formats/content/still_quality.shtml

included in regular TIFF 6.0 files. Adobe Photoshop, on the other hand, appears to use yet another method.⁶⁵

- Support of graphic effects and typography.
 - Vector image options: No.
 - Transparency information: Yes (through a so-called alpha channel).
 - Option to specify fonts and styles: No.
- Functionality beyond normal display
 - Animation: No.
 - Multipage support: Yes.
 - Multiresolution: TIFF offers the option of multiresolution (Image Pyramid). It is unclear whether this a subsequent addition to the private tags.
 - In any case, it is not a part of the TIFF 6.0 1992 standard (baseline and extended). It is also unclear to which extent this functionality is supported by viewers.

5.7 Conclusion

Format Description

- 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 encoding: Topic of investigation.

Consequences for Storage Capacity

- Storage gain is approximately 30%.

Consequences for Image Quality

- Lossless, so none.

Consequences for the Long-Term Sustainability

- File Format Assessment Method (lowest score): 65,3
- Main problem: Possible patents on the LZW compression method and the resulting low rate of adoption as a master archive format in the cultural sector.

Consequences for the Functionality

The most important advantages:

- Lossless compression
- Support of image editing and viewer software
- Comprehensive metadata possibilities
- Options for very diverse bit depths (1 to 16 bits per channel)
- Option for including EXIF information

⁶⁵ LOC TIFF docu: <http://www.digitalpreservation.gov/formats/fdd/fdd000022.shtml#factors>.

The most important disadvantages:

- No option for lossy compression, which leaves the images relatively large
- No browser software support

Recommendation

Reason 1: Substitution

TIFF 6.0 LZW is the least desirable option from the perspective of long-term sustainability (the lowest score in the File Format Assessment Method). The uncertainties regarding the patents that appear to exist on the LZW compression method render the choice of TIFF LZW unwise for this objective. Lossless compression LZW is in itself ideal for substitution objectives because no image information is lost. However, the compression is much less effective (30%) than that of JPEG 2000 Part 1 lossless and PNG (50% and 40%, respectively). The comprehensive software support is a plus but the low level of actual use (by both consumers as well as the cultural heritage sector) is worrisome.

Reason 2: Redigitisation Is Not Desirable

The patents and the less effective lossless compression do not make TIFF LZW an obvious choice for this objective.

Reason 3: Master File is Access File

The lack of a lossy compression option does not make TIFF LZW an obvious choice for this objective.

6 Conclusion

Description of Formats

Standardization: JPEG 2000, PNG and JPEG are ISO/IEC standardized. TIFF 6.0 is not, though the TIFF 6.0 standard is public and is made available by Adobe.

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 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 sharpness of JPEG degrades gradually when compression increases. In JPEG 2000, some sharpness deterioration 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 earlier in JPEG than in JPEG 2000.
 - Banding (rough colour or tone transitions)
 - Pixelation (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) of the degradation that occurs during lossy compression.

Consequences for the Long-Term Sustainability

Application of the previously discussed File Format Assessment Method (see the introduction and Appendices 2 and 3) to the image formats discussed in this report, plus the uncompressed TIFF format that has been used until now for the master images, results in the following order in these formats 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

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, not much experience has been had with the application of this method in practice. Based on the experiences gained in this study 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 in the long-term, it is important – certainly if the original cannot be rescanned – to not only consider the use of the digitalized material in the short term but also 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 for future, other uses of the material. 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 add EXIF metadata	JPEG, TIFF LZW
Browser support	JPEG, PNG
Multiresolution options (suitability of the file as a high-resolution <i>access</i> 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.

	JPEG 2000 part 1 lossless	JPEG 2000 part1 lossy	PNG lossless	JPEG/JFIF 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

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 using the master file as access file 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.

Appendix 1: Use of Alternative File Formats

The below list is by no means complete. Its sole objective is to provide an idea of the distribution of the various file formats.

JPEG 2000

Although there are many institutions that are using JPEG 2000 files as “access copies” and many institutions are investigating the use of JPEG 2000 as an archival format, only one cultural institution has been found to date that has definitively chosen JPEG 2000 as its sole archival format. A topic of investigation is the use of JPEG 2000 in the medical field.

Examples of institutions and companies that use JPEG 2000:

- The British Library is the only institution that has chosen JPEG 2000 as one of its archival format (TIFF is still uses as well): “The DPT have taken the view that since the budget for hard drive storage for this project has already been allocated, it would be impractical to recommend a change in the specifics as far as file format is concerned for this project. As such, we recommend retaining the formats originally agreed in MLB_v2.doc. These are:
 - Linearized PDF 1.6 files for access, with the “first page” being either the table of contents, or the first page of chapter one, depending on the specifics of the book being scanned.
 - JPEG 2000 files compressed to 70 dB PSNR for the preservation copy.”
 - METS/ALTO3 XML for metadata.

The JP2 files fulfil the role of master file but a lack of industry take-up is a slight concern from a preservation viewpoint. However, the format is well defined and documented and poses no immediate risk.”⁶⁶

The risk of “lack of industry take-up” is thus recognized but is not considered as a large enough threat to prevent a choice for JPEG 2000.

- Library of Congress: Uses JPEG 2000 accessing the American Memory website (<http://memory.loc.gov/ammem/index.html>).
- National Digital Newspaper program (NDNP) (<http://www.loc.gov/ndnp/>) uses uncompressed TIFF 6.0 as master and JPEG 2000 for all derivatives.
- At the National Archives of Japan you can choose between JPEG and JPEG 2000 for accessing objects in the Digital Gallery (http://jpimg.digital.archives.go.jp/kouseisai/index_e.html). The format in which the master images are stored is unclear.
- Google uses JPEG 2000 in Google Earth and Google Print.
- Second Life uses JPEG 2000.

⁶⁶ <http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresmicro.pdf>.

- Motion JPEG 2000 (MJ2) is used by the members of Digital Cinema Initiatives (DCI) as a standard for digital cinema. Some of the members of DCI are:
 - Buena Vista Group (Disney)
 - 20th Century Fox
 - Metro-Goldwyn-Mayer
 - Paramount Pictures
 - Sony Pictures Entertainment
 - Universal Studios
 - Warner Bros. Pictures
- The medical arena uses JPEG 2000 quite a lot - see DICOM (<http://medical.nema.org/>).
- Biometrics: e.g. the new German passport contains a chip with biometric data and an image in JPEG 2000.
- Video Surveillance Applications
- The Library and Archives Canada (LAC) conducted a feasibility study regarding the use of JPEG 2000 (<http://www.archimuse.com/mw2007/papers/desrochers/desrochers.html>). Up until now however, a copy in TIFF is being archived as well. This is done as an extra safety net.
- Internet Archive.
- University of Connecticut (http://charlesolson.uconn.edu/Works_in_the_Collection/Melville_Project/index.htm).
- University of Utah (<http://www.lib.utah.edu/digital/collections/sanborn/>).
- Smithsonian Libraries.
- J. Paul Getty.

PNG

- The National Archives of Australia uses PNG as archival format.
- No further cultural heritage institutions were found that use the PNG format as an archival master.

JPEG

- The masters of the newspapers of the Leids Archief (Archive in Leiden) are stored as JPEGs.
- The National Library of the Czech Republic uses high-quality JPEG (PSD 12) files as masters for the Memoria and Kramerius projects.
<http://www.ncd.matf.bg.ac.yu/casopis/05/Knoll/Knoll.pdf>.

TIFF LZW

- The National Archives and Records Administration (NARA) in the U.S. uses TIFF LZW as archival master for their internal digitisation projects.
- No other examples were found.

Appendix 2: File Format Assessment Method – Output

See Appendix 3 for a description of the method. The criteria, characteristics and weighing factors are not exactly the same as in the IPRES paper in Appendix 3. This is because after presenting the paper at IPRES and gaining more experience in applying the method, it has already appeared necessary to adapt the method.

Raster Images	Weight ⁶⁷	Baseline TIFF 6.0 uncompressed		basic JFIF (JPEG) 1.02		JP2 (JPEG-2000 Part 1) lossy compressed		JP2 (JPEG-2000 Part 1) lossless compressed		PNG 1.2		TIFF_LZW 6.0	
		Score	Total	Score	Total	Score	Total	Score	Total	Score	Total	Score	Total
Openness	3												
Standardization	9	1	3	1.5	4.5	2	6	2	6	2	6	1	3
Restrictions on the interpretation of the file format	9	2	6	1	3	1	3	1	3	2	6	1	3
Reader with freely available source	7	2	4.66667 ⁶⁸	2	4.66667	2	4.66667	2	4.66667	2	4.66667	2	4.66667
Adoption	2												
World wide usage	4	1	2	2	4	1	2	1	2	1	2	1	2
Usage in the cultural heritage sector as archival format	7	2	7	0	0	0	0	1	3.5	1	3.5	0	0
Complexity	3												
Human readability	3	0	0	0	0	0	0	0	0	0	0	0	0
Compression	6	2	4	0	0	0	0	1	2	1	2	1	2
Variety of features	3	1	1	1	1	1	1	1	1	1	1	1	1
Technical Protection Mechanism (DRM)	5												
Password protection	3	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2
Copy protection	3	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2
Digital signature	3	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2
Printing protection	3	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2

⁶⁷ The weights that are assigned to the criteria and their characteristics are not fixed. They depend on the local policy of an institution. The weights that are used in the examples in this paper are the weights as assigned by the KB based on its local policy, general digital preservation literature and common sense.

⁶⁸ $4,6667 = 2 \text{ (score)} * 7 \text{ (weight for the characteristic)} / 3 \text{ (normalisation factor because there are 3 sub-characteristics for the openness criterion)}$

Content extraction protection	3	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2	2	1.2
Self-documentation	2												
Metadata	1	2	1	1	0.5	2	1	2	1	1	0.5	2	1
Technical description of format embedded	1	1	0.5	0	0	0	0	0	0	0	0	1	0.5
Robustness	5												
Format should be robust against single point of failure	2	1	0.4	1	0.4	2	0.8	2	0.8	1	0.4	0	0
Support for file corruption detection	2	0	0	0	0	0	0	0	0	0	0	0	0
File format stability	2	2	0.8	2	0.8	2	0.8	2	0.8	2	0.8	2	0.8
Backward compatibility	2	2	0.8	2	0.8	2	0.8	2	0.8	2	0.8	2	0.8
Forward compatibility	2	2	0.8	0	0	0	0	0	0	0	0	2	0.8
Dependencies	4												
Not dependent on specific hardware	8	2	4	2	4	2	4	2	4	2	4	2	4
Not dependent on specific operating systems	8	2	4	2	4	2	4	2	4	2	4	2	4
Not dependent on one specific reader	8	2	4	2	4	2	4	2	4	2	4	2	4
Not dependent on other external resources (font + codecs)	8	2	4	2	4	2	4	2	4	2	4	2	4
Maximum score = 63,667		53.9667		41.6667		42.0667		47.5667		49.6667		41.5667	
Perct of 100		84.7644		65.445		66.0733		74.712		78.0104		65.2879	

Appendix 3 File Format Assessment Method – Explained

The following paper, concerning the File Format Assessment Method, was presented, in a slightly different form, on the IPRES Conference 2007 (<http://ipres.las.ac.cn/>) but is not yet published. Some changes have been made in the definitions of the criteria and characteristics after gaining more experience with applying the method and receiving feedback from others.

Evaluating File Formats for Long-term Preservation

Judith Rog, Caroline van Wijk
National Library of the Netherlands; The Hague, The Netherlands
judith.rog@kb.nl, caroline.vanwijk@kb.nl

Abstract

National and international publishers have been depositing digital publications at the National Library of the Netherlands (KB) since 2003. Until recently, most of these publications were deposited in the Portable Document Format. New projects, for example the web archiving project, force the KB to handle more heterogeneous material. Therefore, the KB has developed a quantifiable file format risk assessment method. This method can be used to define digital preservation strategies for specific file formats. The choice for a specific file format at creation time or later in the life cycle of a digital object influences the long-term access to the digital object. The evaluation method contains seven sustainability criteria for file formats that are weighed for importance. There seems to be consensus on the sustainability criteria. However, as the weighing of these criteria is connected to an institution's policy, the KB wonders whether agreement on the relative importance of the criteria can be reached at all. With this paper, the KB hopes to inspire other cultural heritage institutions to define their own quantifiable file format evaluation method.

Introduction

Over more than a decade, the Koninklijke Bibliotheek (KB) has been involved with the preservation of digital publications. In 1996, the first agreements were signed with Elsevier Science and Kluwer Academic, international publishers of Dutch origin, on the long-term preservation of their e-journals. In 2002 it was decided that the scope of the e-Depot would be broadened to cover the whole spectrum of international scientific publishing. The e-Depot, the electronic archive the KB uses for the long-term storage and preservation of these journals, became operational in 2003 (National Library of the Netherlands, 2007a). At this moment, the e-Depot holds over 10 million international e-publications. Up until now, the vast majority of the publications in the e-Depot consist of articles from e-journals. For all but a few of these articles the format in which they are published is the Portable Document Format (PDF), ranging from PDF version 1.0 to 1.6. For this reason, the research the KB has done to keep the articles preserved and accessible for future use, focused mainly on PDF. At this moment, however, the scope of the e-Depot is broadened. Apart from the ongoing ingestion of the electronic publications, in the coming five years, data resulting from ongoing projects such as web archiving (Digital Preservation Department KB, 2007b), DARE (Digital Preservation Department KB, 2007c), national e-Depot (KB, 2007d) and several digitisation projects (KB, 2007e) will be ingested in the e-Depot as well. The content from these projects

is very heterogeneous concerning file formats. Even the ‘traditional’ publications that the publishers are providing are getting more and more diverse. Articles can be accompanied by multi media files or databases that illustrate the research.

This more diverse content forces the KB to reconsider its digital preservation strategy. At the foundation of each strategy is the basic principle that the KB will always keep the original publication. The digital preservation strategy describes what actions (e.g. migration or emulation) the KB undertakes to ensure that these publications are preserved and remain accessible for future use. The strategy also describes which choices to make for specific formats during creation, ingest or at a later stage because choices at each of these stages can influence the sustainability of the file. The current strategy is mainly focused on preserving PDF files, but our strategy will need to cover a much wider variety of formats from now on. Whether preservation actions are needed and which actions are needed, depends among other things on the long-term sustainability of the file format of the publication. But what makes a file format suitable for long-term preservation? The criteria for evaluating file formats have been described by several authors (Folk & Barkstrom, 2002; Christensen, 2004; Brown, 2003; Arms & Fleischhauer, 2005; Library of Congress, 2007). But only very rarely though are these criteria applied to a practical assessment of the file formats (Anderson, Frost, Hoebelheinrich & Johnson, 2005). To apply the sustainability criteria we need to know whether all criteria are equally important or whether some are more important than others. And how do you measure whether, and to what degree the format meets the criteria? The application of the criteria should be quantifiable to be able to compare file formats and to give more insight into the preference for certain file formats for long-term preservation.

The KB has started to develop such a quantifiable file format risk assessment. The file format risk assessment facilitates choosing file formats that are suitable for long-term preservation. This paper describes the file format assessment method that the KB has developed and how it is applied in the preservation strategies at the KB. The KB invites the digital preservation community to start a discussion on sustainability criteria and the importance of each criterion by presenting its file format evaluation method.

File Format Assessment for Long-term Preservation

Methodology

The general preservation criteria used in the KB’s method originate from the aforementioned digital preservation literature. The KB’s assessment method does not take into account quality and functionality criteria such as clarity or functionality beyond normal rendering as defined in Arms & Fleischhauer (2005). The KB archives publications which are end products that for example do not need editing functionality after publishing. Also the KB archives the publications for long-term preservation purposes and is not the main point of distribution for these publications. Regular access to and distribution of publications is offered by publisher’s websites and university repositories etc. This reasoning might be very specific to the KB and it explains the choice for only applying sustainability criteria in the risk assessment method. In the next sections, the criteria, the weighing of the criteria and an example of the application of the method will be described.

The criteria on which classifications of suitability of file formats from the view point of digital preservation will be based are described below. The criteria form measurable standards by which the suitability of file formats can be assigned. The criteria are broken

down into several characteristics that can be applied to all file formats. Values are assigned to each characteristic. The values that are given differ among file formats. The sustainability criteria and characteristics will be weighed, as the KB does not attribute the same importance for digital preservation planning to all characteristics. The weights that are assigned to the criteria and their characteristics are not fixed. They depend on the local policy of an institution. The weights that are used in the examples in this paper are the weights as assigned by the KB based on its local policy, general digital preservation literature and common sense. The range of values that can be assigned to the characteristics are fixed.

The weighing scale runs from zero to seven. These extremes are arbitrary. Seven is the weight that is assigned to very important criteria from the point of view of digital preservation and zero is the score assigned to criteria that are to be disregarded. The values that are assigned to the characteristics range from zero to two. The lowest numerical value is assigned to the characteristic value that is seen as most threatening to digital preservation and long-term accessibility. This value is zero. The highest numerical value is assigned to the characteristic value that is most important for digital preservation and long-term accessibility. This value is two. The scale from zero to two is arbitrary. The criteria do not all have the same number of characteristics. The total score that is assigned to all characteristics is therefore normalised by dividing the score by the number of characteristics.

By applying the file format assessment method to a file format, the format receives a score that reflects its suitability for long-term preservation on a scale from zero to hundred. The higher the score, the more suitable the format is for long-term preservation. The score a format receives can vary over time. A criterion such as *Adoption* for example is very likely to change over time as a format gets more popular or becomes obsolete.

Criteria defined

The criteria that are used in this methodology are *Openness, Adoption, Complexity, Technical Protection Mechanism (DRM), Self-documentation, Robustness and Dependencies*.

Openness

The criterion *Openness* of a file format is broken down into the characteristics *Standardisation, Restrictions on the interpretation of the file format, Reader with freely available source*. These characteristics indicate the relative ease of accumulating knowledge about the file format structure. Knowledge about a file format will enhance the chance of successful digital preservation planning.

Adoption

The criterion *Adoption* of a file format has two characteristics: *World wide usage* and *Usage in the cultural heritage sector as archival format*. These characteristics indicate the popularity and ubiquity of a file format. When a specific file format is used by a critical mass, software developers (commercial, non commercial) have an incentive to sustain support for a file format by developing software for the specific file format such as readers and writers. However, as a cultural heritage institution, it is not only important to consider usage in general, but also, and more importantly even, the usage by other cultural heritage institutions that share the same goal of preserving the documents for the long-term.

Complexity

The characteristic *Complexity* of a file format is broken down into the characteristics *Human readability*, *Compression*, *Variety of features*. These characteristics indicate how complicated a file format can be to decipher. If a lot of effort has to be put into deciphering a format, and with the chance it will not completely be understood, the format can represent a danger to digital preservation and long-term accessibility.

Technical Protection Mechanism (DRM)

The characteristic *Technical Protection Mechanism* of a file format is broken down into the characteristics *Password protection*, *Copy protection*, *Digital signature*, *Printing protection* and *Content extraction protection*. These characteristics indicate the possibilities in a file format to restrict access (in a broad sense) to content. Restricted access to content could be a problem when the digital preservation strategy migration is necessary to provide permanent access to the digital object.

Self-documentation

The characteristic *Self-documentation* of a file format is broken down into the characteristics *Metadata* and *Technical description of format embedded*. These characteristics indicate the format possibilities concerning encapsulation of metadata. This metadata can be object specific or format specific. When a format facilitates the encapsulation of object specific information (such as author, description etc.) or format specific information in the header on how to read the format for example, the format supports the preservation of information without references to other sources. The more that is known about a digital object, the better it can be understood in the future.

Robustness

The characteristic *Robustness* of a file format is broken down into the characteristics *Robust against single point of failure*, *Support for file corruption detection*, *File format stability*, *Backward compatibility* and *Forward compatibility*. These characteristics indicate the extend to which the format changes over time and the extend to which successive generations differ from each other. Also, this characteristic provides information on the ways the file format is protected against file corruption. A frequently changing format could threaten continuity in accessibility for the long term. Large differences among generations of a file format could endanger this continuity equally. The values for *file format stability* ‘rare release of newer versions’, ‘limited release of newer versions’ and ‘frequent release of newer versions’ correspond to ‘release once in ten years’, ‘release once in five years’ and ‘release once a year’ respectively.

Dependencies

The characteristic *Dependencies* of a file format is broken down into the characteristics *Not dependent on specific hardware*, *Not dependent on specific operating systems*, *Not dependent on one specific reader* and *Not dependent on other external resources*. These characteristics indicate the dependency on a specific environment or other resources such as fonts and codecs. A high dependency on a specific environment or on external resources provides a risk for digital preservation and long-term accessibility. External resources could be lost over time and difficult to retain and a high dependency on a specific environment strongly ties the format to a specific time and space.

The full list of criteria, the weights as assigned by the KB, the criteria and their possible values can be found in Appendix I. An example of the file format assessment method applied to MS Word 97-2003 and PDF/A-1 can be found in Appendix II

Application of File Format Assessments

The KB has defined a digital preservation policy for the content of the e-Depot. This policy is the starting point for digital preservation strategies for the digital objects stored in the e-Depot. A digital preservation strategy starts at creation time of a digital object and defines preservation actions on the object at a later stage in the object's life cycle. The KB will not restrict the use of specific file formats for deposit. Any format in general use can be offered. However, KB does give out recommendations and uses the file format assessment method to define strategies.

During the last decade the KB has carried out many digitisation projects. The development of digitisation guidelines has been part of these projects. These guidelines not only make sure that specific image quality requirements are met. They also ensure that the created master files meet the requirements that the digital preservation department has set for metadata and technical matters such as the use of specific file formats and the use of compression (no compression or lossless compression). A file format evaluation method is essential for making well thought-out choices for specific file formats at creation time of digital objects.

The KB has had a lot of influence on the creation process as the owner of the digitisation master files. However, this is not the case for millions of digital publications that have been and will be deposited by international publishers. The KB does have deposit contracts that contain several technical agreements (e.g. file format in which the publisher chooses to submit the publications). Also, as most publications are deposited in PDF, guidelines for the creation of publications in PDF (Rog, 2007) have been created. The PDF guidelines are related to the standard archiving format PDF/A, but are easier to read for non-technical persons. They contain ten 'rules' for PDF functionality that describe best practices at creation.

As was mentioned before, the deposited publications have been quite homogenous concerning file formats. Most publications have been deposited in PDF version 1.0 to 1.6. The file format assessment method has been used to assess this main format stored for its digital preservation suitability. However, new projects will make the digital content of the archive more heterogeneous in the near future. This will require more elaborated file format evaluations.

One example of the use of file format evaluations for new e-Depot content is the evaluation of formats that are harvested for the DARE project. DARE publications are harvested from scientific repositories such as the Dutch university repositories. Most harvested publications are PDFs, however a small part of the articles are harvested in MS Office document formats such as MS Word and MS PowerPoint and in the WordPerfect format. The concrete result of the use of file format risk assessment at the KB is the decision to normalise MS Office documents and WordPerfect documents to a standard archiving format: PDF/A. MS Word documents score 22% if assessed by the assessment method. PDF/A's assessment score amounts to 89 %. The main difference between the formats can be found in the criteria *Openness*, *Adoption* and *Dependencies*. For these three criteria, MS Word does have a considerably lower score than PDF/A-1 has. In accordance with the preservation policy both original and normalised files are stored for long-term preservation purposes.

Interestingly enough, an archival institution that is partner in the National Digital Preservation Coalition (NCDD), does not consider PDF/A suitable for archiving its digital data for the long term. One of its valid arguments for not using PDF/A is that PDF/A does not offer the same editing functionality that is available in datasheets. It would be very interesting to compare the differences among cultural heritage institutions concerning the sustainability criteria and the importance of these criteria. This will be much easier if institutions make their file format evaluation quantifiable.

The biggest challenge for the application of the file format risk assessment in the near future will be the web archiving project. As websites contain many different file formats, this new type of content for the e-Depot will require quite different preservation strategies and plans from the current ones.

Conclusion and Discussion

This paper describes the file format assessment that was developed by the KB to assess the suitability of file formats for long-term preservation. The suitability is made quantifiable and results in a score on a scale from zero to hundred that reflects the suitability of the format for long-term preservation. Formats can easily be compared to each other. The criteria, characteristics and scores that the formats receive are transparent.

The KB hopes to receive feedback on the methodology from other institutions that have to differentiate between formats to decide which format is most suitable for long-term preservation. There seems to be consensus on the sustainability criteria. However, the KB would like to know whether these criteria are the right ones and whether the possible scores a format can receive on a characteristic offer practical options to choose from. The weighing that can be applied to a criterion is not fixed in the methodology. The weighing can be adjusted to the local policy. Therefore, the KB would like to invite other cultural heritage institutions for a discussion about and preferably a comparison of quantifiable file format risk assessments.

References

- National Library of the Netherlands (KB). (2007a). *The archiving system for electronic publications: The e-Depot*. Retrieved August 20, 2007, from: <http://www.kb.nl/dnp/e-depot/dm/dm-en.html>
- Digital Preservation Department National Library of the Netherlands (KB). (2007b). *Web archiving*. Retrieved August 20, 2007, from http://www.kb.nl/hrd/dd/dd_projecten/projecten_webarchivering.html;
- Digital Preservation Department National Library of the Netherlands (KB). (2007c). *DARE: Digital Academic Repositories*. Retrieved August 20, 2007, from http://www.kb.nl/hrd/dd/dd_projecten/projecten_dare-en.html
- National Library of the Netherlands (KB). (2007d). *Online deposit of electronic publications*. Retrieved August 20, 2007, from <http://www.kb.nl/dnp/e-depot/loket/index-en.html>
- National Library of the Netherlands (KB). (2007e) *Digitisation programmes & projects*. Retrieved August 20, 2007, from <http://www.kb.nl/hrd/digi/digdoc-en.html>

Folk, M. & Barkstrom, B. R. (2002). *Attributes of File Formats for Long-Term Preservation of Scientific and Engineering Data in Digital Libraries*. Retrieved August 20, 2007, from http://www.ncsa.uiuc.edu/NARA/Sci_Formats_and_Archiving.doc

Christensen, S.S. (2004). *Archival Data Format Requirements*. Retrieved August 20, 2007, from http://netarchive.dk/publikationer/Archival_format_requirements-2004.pdf

Brown, A. (2003). *Digital Preservation Guidance Note 1: Selecting File Formats for Long-Term Preservation*. Retrieved August 20, 2007, from http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf

Arms, C. & Fleischhauer, C. (2005). Digital formats: Factors for sustainability, functionality and quality. In *Proceedings Society for Imaging Science and Technology (IS&T) Archiving 2005* (pp. 222-227).

Library of Congress. (2007). *Sustainability of Digital Formats Planning for Library of Congress Collections*. Retrieved August 20, 2007, from <http://www.digitalpreservation.gov/formats/sustain/sustain.shtml>

Anderson, R., Frost, H., Hoebelheinrich, N. & Johnson, K. (2005) The AIHT at Stanford University, *D-Lib Magazine* (11), 12. Retrieved 20 August, 2007, from <http://www.dlib.org/dlib/december05/johnson/12johnson.html>

Rog, J. (2007). *PDF Guidelines: Recommendations for the creation of PDF files for long-term preservation and access*, Retrieved from http://www.kb.nl/hrd/dd/dd_links_en_publicaties/PDF_Guidelines.pdf

Author Biography

Caroline van Wijk (1973) has a BA degree in Art and an MA in Political Science. She finished a Java software engineer training in 2000. Directly after, she had been working at a number of web development companies for well over four years before she joined the KB in 2004. At the KB, she had worked on the pilot project Tiff-archive as the technical project staff member until December 2005. Since 2006, she leads the migration implementation project and takes part in the European project Planets as a digital preservation researcher and work package leader.

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.

Appendix I

Table 1: All criteria, weighting factors, characteristics and values that can be applied

Criterion	Characteristic (weighing factor)	values	
Openness			
	Standardisation (9)		
		2	De jure standard
		1,5	De facto standard, specifications made available by independent organisation
		1	De facto standard, specifications made available by manufacturer only
		0,5	De facto standard, closed specifications
		0	No standard
	Restrictions on the interpretation of the file format (9)		
		2	No restrictions
		1	Partially restricted
		0	Heavily restricted
	Reader with freely available source (7)		
		2	Freely available open source reader
		1	Freely available reader, but not open source
		0	No freely available reader
Adoption			
	World wide usage (4)		
		2	Widely used
		1	Used on a small scale
		0	Rarely used
	Usage in the cultural heritage sector as archival format (7)		
		2	Widely used
		1	Used on a small scale
		0	Rarely used
Complexity			
	Human readability (3)		
		2	Structure and content readable
		1	Structure readable
		0	Not readable
	Compression (6)		
		2	No compression
		1	lossless compression
		0	lossy compressed
	Variety of features (3)		
		2	Small variety of features
		1	Some variety of features
		0	Wide variety of features
Technical Protection Mechanism (DRM)			
	Password protection (3)		
		2	Not possible
		1	Optional
		0	Mandatory
	Copy protection (3)		

Criterion	Characteristic (weighing factor)	values	
		2	Not possible
		1	Optional
		0	Mandatory
	Digital signature (3)		
		2	Not possible
		1	Optional
		0	Mandatory
	Printing protection (3)		
		2	Not possible
		1	Optional
		0	Mandatory
	Content extraction protection (3)		
		2	Not possible
		1	Optional
		0	Mandatory
Self-documentation			
	Metadata (1)		
		2	Possibility to encapsulate user-defined metadata
		1	Possibility to encapsulate a limited set of metadata
		0	No metadata encapsulation
	Technical description of format embedded (1)		
		2	Fully self-describing
		1	Partially self-describing
		0	No description
Robustness			
	Format should be robust against single point of failure (2)		
		2	Not vulnerable
		1	Vulnerable
		0	Highly vulnerable
	Support for file corruption detection (2)		
		2	Available
		0	Not available
	File format stability (2)		
		2	Rare release of new versions
		1	Limited release of new versions
		0	Frequent release of new versions
	Backward compatibility (2)		
		2	Large support
		1	Medium support
		0	No support
	Forward compatibility (2)		
		2	Large support
		1	Medium support
		0	No support
Dependencies			
	Not dependent on specific hardware (8)		
		2	No dependency
		1	Low dependency
		0	High dependency
	Not dependent on specific operating systems (8)		

Criterion	Characteristic (weighing factor)	values	
		2	No dependency
		1	Low dependency
		0	High dependency
	Not dependent on one specific reader (8)		
		2	No dependency
		1	Low dependency
		0	High dependency
	Not dependent on other external resources (7)		
		2	No dependency
		1	Low dependency
		0	High dependency

Appendix II

Table 2: Example application of the file format assessment method to MS Word 97-2003 and PDF/A-1

Criteria	Characteristics	Weight	PDF/A-1		MS Word 97-2003	
			Score	Total	Score	Total
Openness		3				
	Standardisation	9	2	6	0,5	1,5
	Restrictions on the interpretation of the file format	9	2	6	0	0
	Reader with freely available source	7	2	4,666666667 ⁶⁹	0	0
Adoption		2				
	World wide usage	4	2	4	2	4
	Usage in the cultural heritage sector as archival format	7	2	7	0	0
Complexity		3				
	Human readability	3	1	1	0	0
	Compression	6	1	2	0	0
	Variety of features	3	1	1	0	0
Technical Protection Mechanism (DRM)		5				
	Password protection	3	2	1,2	1	0,6
	Copy protection	3	2	1,2	1	0,6
	Digital signature	3	2	1,2	1	0,6
	Printing protection	3	2	1,2	2	1,2
	Content extraction protection	3	2	1,2	2	1,2
Self-documentation		2				
	Metadata	1	2	1	2	1
	Technical description of format embedded	1	0	0	0	0
Robustness		7				
	Format should be robust against single point of failure	2	0	0	0	0
	Support for file corruption detection	2	0	0	0	0
	File format stability	2	2	0,8	1	0,4
	Backward compatibility	2	2	0,8	2	0,8
	Forward compatibility	2	1	0,4	0	0
Dependencies		4				
	Not dependent on specific hardware	8	2	4	0	0
	Not dependent on specific operating systems	8	2	4	0	0
	Not dependent on one specific reader	8	2	4	0	0
	Not dependent on other external resources	8	2	4	1	2
Total score				56,66666667		13,9
	Normalised to percentage of 100⁷⁰			89,01 %		21,83 %

⁶⁹ $4,6667 = 2 \text{ (score)} * 7 \text{ (weight for the characteristic)} / 3 \text{ (normalisation factor because there are 3 sub-characteristics for the openness criterion)}$

⁷⁰ The maximum score a format can receive is 63,667. By multiplying the total score by 100 and dividing it by 63,667 it is normalised to a scale from 0-100.

Appendix 4: Storage Tests

As said in the introduction tow limitations were places upon this test:

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

The test images on which the below data are based are 94, 300 ppi, 24 bits RGB low contrast scans of popular ballads.⁷¹ The originals vary in format between slightly larger than A4 to smaller than A5.

File Format and Compression	File Size of Test Batch	Average File Size ⁷²	Storage Savings Compare to Uncompressed TIFF ⁷³	Storage Interpolated for 500.000 Files ⁷⁴
Uncompressed TIFF	623 MB	6.6 MB		3.1 TB
TIFF LZW lossless	428 MB	4.6 MB	31%	2.2TB
JPEG 10 ⁷⁵	66 MB	0.7 MB	89%	343 GB
JPEG 8	35 MB	0.4 MB	94%	195 GB
JPEG 6	26 MB	0.3 MB	96%	146 GB
JPEG 1	10 MB	0,1 MB	98%	49 GB
PNG lossless	355 MB	4 MB	43%	2 TB
JPEG2000 lossless ⁷⁶	298 MB	3,2 MB	52%	1.5 TB
JPEG2000 compression ratio10	54 MB	0.6 MB	91%	280 GB
JPEG2000 compression ratio 25	25 MB	0.3 MB	96%	146 GB
JPEG2000 compression ratio 50	13 MB	0,1 MB	98%	68 GB

In addition to this set of popular ballads, a test was conducted on 104 scans from photo prints, scanned in RGB. The results were almost identical.

⁷¹ Scanned within the scope of the Geheugen van Nederland (Memory of the Netherlands) project <http://www.geheugenvannederland.nl/straatliederen>.

⁷² The number of files – 94 – divided by the file size of all files together.

⁷³ Percentage of total storage of 94 uncompressed TIFFs (RGB 653 GB and grey 218 GB) compared to total storage of 94 compressed files.

⁷⁴ Average file size multiplied by 500.000.

⁷⁵ JPEG Adobe Photoshop scale quality 10.

⁷⁶ Lead JPEG2000 plugin for Photoshop is used, whereby the amount of compression is set by means of the compression ratio. Compression ratio 10 is minimum compression and is qualitatively comparable to JPEG10. Compression ratio 25 is average compression and is qualitatively comparable to JPEG6. Compression ratio 50 is strong compression and is qualitatively comparable to JPEG1.

Additional testing was performed with the Photoshop native plugin. Lossless compression proved to be slightly less successful than that of the LEAD plugin: 53% for the Lead plugin, versus 52% for the Photoshop plugin.

Additional testing with other converters, like the Lurawave tool (<http://www.luratech.com/products/lurawave/jp2/clt/>), is necessary.

Bibliography

General

- LOC Sustainability of Digital Formats - Planning for Library of Congress Collections - Still Images Quality and Functionality Factors
http://www.digitalpreservation.gov/formats/content/still_quality.shtml.
- *Sustainability of Digital Formats Planning for Library of Congress Collections, Format Description for Still Images*
http://www.digitalpreservation.gov/formats/fdd/still_fdd.shtml.
- Florida Digital Archive Format Information
<http://www.fcla.edu/digitalArchive/formatInfo.htm>.
- Roberto Bourgonjen, Marc Holtman, Ellen Fleurbaay, *Digitalisering ontrafeld. Technische aspecten van digitale reproductie van archiefstukken* (Digitisation Unraveled. Technical Aspects of Digital Reproduction of Archive Pieces).
http://stadsarchief.amsterdam.nl/stadsarchief/over_ons/projecten_en_jaarverslagen/digitalisering_ontrafeld_web.pdf.

JPEG2000

- Judith Rog, *Notitie over JPEG 2000 voor de KB* (Note regarding JPEG 2000 for the RL), version 0.2 (August 2007).
- JPEG2000 homepage: <http://www.jpeg.org/jpeg2000/>.
- Wikipedia JPEG 2000 lemma: http://en.wikipedia.org/wiki/JPEG_2000.
- Robert Buckley, *JPEG2000 for Image Archiving, with Discussion of Other Popular Image Formats*. Tutorial IS&T Archiving 2007 Conference.
- Robert Buckley, *JPEG 2000 – a Practical Digital Preservation Standard?*, a DPC Technology Watch Series Report 08-01, February 2008:
<http://www.dpconline.org/graphics/reports/index.html#jpeg2000>
- Florida Digital Archive description of JPEG 2000 part 1:
http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgrounds/jp2_bg.pdf.
- *Sustainability of Digital Formats Planning for Library of Congress Collections, JPEG 2000 Part 1, Core Coding System*
<http://www.digitalpreservation.gov/formats/fdd/fdd000138.shtml>.

PNG

- Portable Network Graphics (PNG) Specification (Second Edition)
Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification. ISO/IEC 15948:2003 (E), W3C Recommendation 10 November 2003 <http://www.w3.org/TR/PNG/>.
- Wikipedia PNG lemma: http://nl.wikipedia.org/wiki/Portable_Network_Graphics.
- *Sustainability of Digital Formats Planning for Library of Congress Collections, PNG, Portable Network Graphics*
<http://www.digitalpreservation.gov/formats/fdd/fdd000153.shtml>.
- Greg Roelofs, *A Basic Introduction to PNG Features*
<http://www.libpng.org/pub/png/pngintro.html>.

JPEG

- JPEG standard: <http://www.w3.org/Graphics/JPEG/itu-t81.pdf>.

- JPEG homepage: <http://www.jpeg.org/jpeg/index.html>.
- JFIF standard: <http://www.jpeg.org/public/jfif.pdf>.
- Florida Digital Archive description of JFIF 1.02:
http://www.fcla.edu/digitalArchive/pdfs/action_plan_bgrounds/jfif.pdf.
- *Sustainability of Digital Formats Planning for Library of Congress Collections, JFIF JPEG File Interchange Format*
<http://www.digitalpreservation.gov/formats/fdd/fdd000018.shtml>.
- Wikipedia JPEG lemma: <http://en.wikipedia.org/wiki/JPEG>.
- JPEG-LS homepage: <http://www.jpeg.org/jpeg/jpegls.html>.
- Wikipedia JPEG-LS lemma: <http://www.jpeg.org/jpeg/jpegls.html>.

TIFF LZW

- TIFF 6.0 standard: <http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf>.
- TIFF/EP extension ISO 12234-2 http://en.wikipedia.org/wiki/ISO_12234-2.
- TIFF/IT (ISO 12369) extension for prepress purposes
<http://www.digitalpreservation.gov/formats/fdd/fdd000072.shtml>.
- DNG Adobe TIFF UNC extension for storing RAW images
<http://www.digitalpreservation.gov/formats/fdd/fdd000188.shtml>.
- EXIF technical metadata of cameras and camera settings
(<http://www.digitalpreservation.gov/formats/fdd/fdd000145.shtml>).
- LOC overview of TIFF extension
http://www.digitalpreservation.gov/formats/content/tiff_tags.shtml.
- Unisys patent LZW http://www.unisys.com/about_unisys/lzw/.
- *Sustainability of Digital Formats Planning for Library of Congress Collections, TIFF, Revision 6.0* <http://www.digitalpreservation.gov/formats/fdd/fdd000022.shtml>.
- *Sustainability of Digital Formats Planning for Library of Congress Collections, TIFF, Bitmap with LZW Compression*
<http://www.digitalpreservation.gov/formats/fdd/fdd000074.shtml>.