Image File Formats, Digital Archival and TI/A

Peter Fornaro & Lukas Rosenthaler

A Short Introduction
into
Image File Formats

1 Introduction

In general, long-term archival of digital data is a difficult task. On one hand the media, where the digital data is recorded on may be instable and decay with time. On the other hand, the rapid evolution cycle of digital technologies which is measured in years or even months leads to the obsolescence of recording technologies at a fast pace. Old[1] data carriers may not be read anymore because the necessary machinery (tape reader, disk interface etc.) is no longer commercially available. Also, the the information about the file formats – that is the information about the meaning of the bits – may be lost because new formats have become standard. Thus, digital archiving is basically the task of guaranteeing the meaningful reading and decoding of bits in the far future. This task can be divided into parts:

Bitstream preservation

It has to be guaranteed that the bits which are basically *analogue symbols on a* analogue medium[2] can be correctly detected. Since most often the permanence of the bits is higher than the lifetime of a given recording technology, bitstream preservation is basically limited by the obsolescence of a given recording technologies. Thus, copying the bits onto a new data carrier using the latest technology just before a recording technology becomes obsolete will preserve the bitstream. This task called *bitstream migration has to be repeated every 3 - 5 years. Since a bitstream can be copied without information loss and the copies will be* identical to the “original”, this process can be repeated an indefinite number of times (contrary to analogue copies where each generation is affected by more degradation until all information is lost).

File format preservation

File format preservation is much more complex and requires some in-depth knowledge about file formats. In a broad sense, digital data can be defined as anything recorded using a symbol based code on a medium. Such a code uses a finite set S of Symbols

$$S = \{s1, s2,23, ..., sn\}, n \geq 2 \qquad (1)$$

Thus, written text using latin characters, cunei form scripts, Egyptian hieroglyphs are perfect examples of digital data.If $n = 2$, that is if the code uses only two symbols, it is called a binary code. Binary codes are the most simple codes which can easily implemented by computing machinery[3]. In order to *understand a text (that is to extract the by the writer intended information) we have to know the syntax and language the text is written in. For example the symbol “a” may have many different meanings depending on the language and context it is being used. The same holds*

[1] which means more than 5 years old!

[2] e.g. analogue changes in the direction of a magnetic field of a magnetic support.

[3] e.g. $S = \{0, 1\}$, $S = \{TRUE, FALSE\}$, $S = \{+3V, -3V\}$

true for binary information. The meaning of the bits is defined by the file formats. If the file format is not known, the information of a digital file cannot be extracted even if the bits can be read without problems.

In the following, we will concentrate on the problem of selection the proper file format for the preservation of digital images.

2 Generic Requirements for File Formats

Long term preservation of digital data poses some obvious requirements to file formats:

Documentation

The file format must be fully documented and the documentation must be openly available. Therefore all proprietary file formats are principally not suitable for long term archival.

Prevalence

The wider a file format is in use the better it is suited for long term archival. A common file format which is widely used and many people understand is more probable to be in use for a long time and thus will be more easily interpreted in the future.

Simplicity

The simpler a file format is, the easier it is to write a short documentation and the easier it is to write a decoder in the future.

The first requirement is a *hard* requirement, while the two others have some leeway.

2.1 Metadata

Metadata, that is the data about the data, is extremely important, especially in the case of digital images where there are no standard methods for indexing or automatic content description (in contrary to text files such as PDF etc. where such methods have been established). In case of digital images, we have to distinguish between different types of metadata:

Technical metadata

Technical metadata is essential for correctly rendering the bits to an image visible to the human eye. The data consists of information about the dimension of the image in pixels, color model, number of bits per sample etc. These metadata often are an integral part of an image file format.

Descriptive metadata

Descriptive metadata may not be essential from a technical point of view, but contains important descriptive information about an image. This also includes technical data such as shutter speed, GPS coordinates, ISO-values etc., but may also include image description, copyright information, ownership etc. . This kind of metadata may be essential for *using an image in a day-to-day context.*

Most file formats provide some methods to include descriptive metadata into the digital image file. Unfortunately these methods have not been standardized[4].

Further it has to be noted, that image data can be recorded in many different ways. While most digital image sensors are based on an additive color model where for each pixel a *red*, *green* and *blue* (RGB) value is provided, the printing industry often relies on colors based on subtractive color model using *cyan*, *magenta*, *yellow* and *black* as (CMYK) primary colors. However, the additive color model (RGB) is much more widespread and should therefore be used for long term digital image preservation.

2.1.1 EXIF

EXIF[5] (EXchangeable Imagefile Format) is standard proposed by the CIPA[6] and JIETA[7]. The mechanism of recording metadata is based on the TIFF standard where key-value pairs of “tags” and associated values are used. EXIF is targeted to record camera parameters (e.g. ISO speed, aperture etc.) including GPS coordinates if available.

2.1.2 IPTC

IPTC[8] (International Press Telecommunications Council) is targeted towards the use for news media. It contains

Administrative metadata

> Identification of the creator, creation date and location, contact information for licensors of the image, and other technical details.

Descriptive Metadata

> Information about the visual content. This may include headline, title, captions and keywords and can be done using free text or codes from a controlled vocabulary.

Rights metadata

> Copyright information and underlying rights in the visual content including model and property rights, and rights usage terms.

IPTC has its own specification how to record the data independent of a given image file format. Thus IPTC data is usually embedded into an image file just as a binary object.

2.1.3 ICC

The International Color Consortium (ICC)[9] defines a standard for embedding color information within an image file. Color reproduction is a very complex topic and the ICC has

[4] It has to be noted that there are efforts to provide standards such as XMP which has been introduced by Adobe and has been adopted widely

[5] See http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf.

[6] Camera & Imaging Products Association

[7] Japan Electronics and Information Technology Industries Association

[8] 8See https://iptc.org.

[9] See http://color.org

defined a metadata scheme which allows – within the limitations of the capturing device (camera, scanner) and the rendering device (screen, printer etc.) for a faithful reproduction of the colors. By nature these metadata can be very complex and special color management software has to be used. However there are a few color profiles which are widely used and can be rendered by any modern device.

2.1.4 XMP

The eXtensible Metadata Platform (XMP)[10] is an XMLfootnoteeXtended Markup Language, an open standard defined by the World Wide Web Consortium, see http://www. w3.org/XML. and RDF (Resource Description Framework)[11] based standard for embedding metadata in media files. It has been proposed by Adobe and established itself as a quasi standard. Its extensibility and versatility make it very attractive for memory institutions.

2.2 Data Compression

Claude Shannon's fundamental work about information theory "A mathematical theory of communication"[12] contains two important statements, given here in a simplified form:

1. . Any code where the probability of occurrence is not the same for all symbols contains redundancy. In such a case it is possible to find a new code which minimizes redundancy.
2. If a communication path introduces errors into the transmitted symbols, a new code can be found which allows to correct for these errors.

The first statement addresses the possibility of lossless compression whereas the second statement deals with the possibilities of error correction codes. Shannon's theory shows that there is a trade-off between efficiency (lossless compression) on one hand and error correction (Redundancy) on the other hand. Many codes such as the written language contain a lot of redundancy and are therefore quite fault tolerant. For digital computer systems however, a high efficiency is required and therefore often compression techniques are used.

In general, data compression would be very welcome for digital images since these file are often quite heavy. However, lossless compression as provided for example by the widely used ZIP program does not yield good results for images. On the other hand, images usually *do* contain a lot of redundancy. E.g., the probability that a neighbouring pixel of a pixel in the blue sky is also blue is very high. Therefore compression schemes have been devised which try to use this kind of redundancy. These compression schemes

[10] See http://www.adobe.com/products/xmp.html.

[11] Technology for the semantic web, see http://www.w3.org/RDF.

[12] Claude E. Shannon, A mathematical theory of communication, Bell System Technical Journal (1948)

will modify the original data (and thus the original uncompressed bitstream cannot be reconstructed), but they are constructed in a way the visual content should remain the same (and thus the compression should not become visible to the human eye). But at the end, some information is being destroyed by these compression schemes and may, if used incorrectly, introduce visible artifacts in the resulting image.

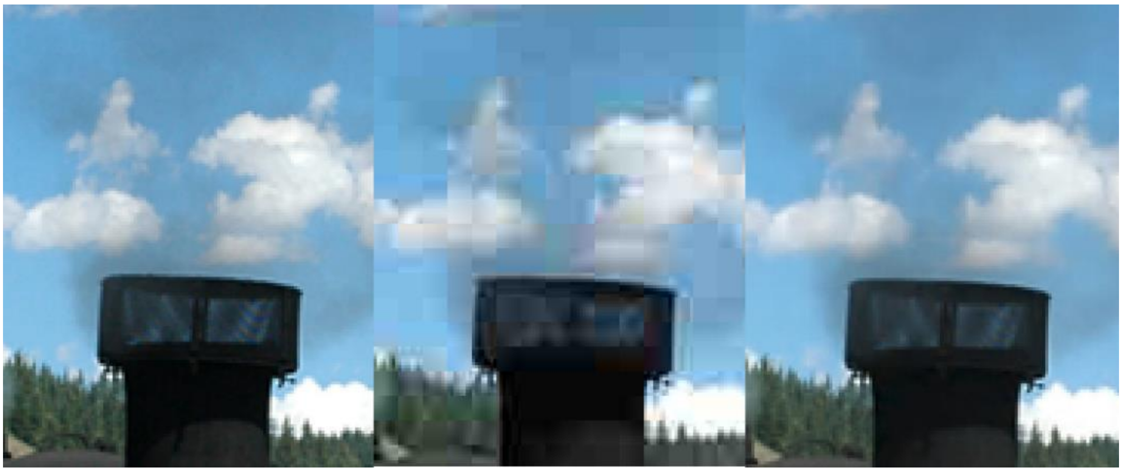

Figure 1: On the left a detail of a TIFF image, in the middle the same detail compressed with the JPEG algorithm 1:100. The middle image clearly shows “blocky” artifacts. On the right the same detail compressed with the JPEG2000 algorithm, also with a compression factor 1:100. The image clearly shows less artifacts, but seems a little blurred.

Thus the ability to represent compressed data is an important factor of file formats for images.

3 Common File Formats

In the following section, some of the most common file formats are described and analyzed. The given list of file formats is not exhaustive at all. Many file formats can be identified by a (hopefully) unique signature of the first few bytes of the file. This signature is often an important aid to identify a file format

3.1 JPEG (extensions: .jpg .jpeg)

The JPEG[13] is probably the most common file format for images. Its success is due to the fact that based on a lossy compression scheme that usually introduces very little artefacts for reasonable compression factors (1:10 to 1:25). Metadata is embedded using “chunks”

[13] Joint Photographers Experts Group file interchange format

of binary data where each chunk is limited to a maximum of 65533 bytes. There are some convention how metadata has to be written, but it’s not strictly standardised.

3.2 PNG (extensions: .png)

The Portable Network Graphics (PNG) has been created as an improved, non-patented replacement for Graphics Interchange Format (GIF) and is one of the most used lossless image compression format on the Internet. Also possible, the ways to include metadata are quite limited and peculiar. While XMP is possible, EXIF and IPTC metadata are not supported unless converted to XMP.

3.3 JPEG2000 (extensions: .jp2 .jpx)

JPEG2000 has been created as a much more powerful replacement of the standard JPEG file format. Its compression scheme is extremely powerful and allows both lossy and lossless compression. It is based on a mathematical functions of the wavelet transform which is close to the neuronal processing of images in the human brain and thus reduces the visual effect of compression artifacts. The algorithms are used are very complex and pose high demands on computing power for compression (decompression is much less demanding). Therefore camera manufacturers are very reluctant in implementing it as standard format for digital cameras even if it would result in higher quality for compressed images. However, the JPEG2000 has been selected by the moving image industry as standard for presenting digital films to the public. The Digital Cinema Pack (DCP) which is used to present the films in the theaters is based on a series of JPEG2000 images. The inclusion of metadata is possible with the JPEG2000 format, but many aspects thereof have not yet been standardized.

3.4 TIFF (extensions: .tif .tiff)

The Tagged Image File Format (TIFF), originally defined by Aldus and now under the auspices of Adobe, is a well documented, open file format for uncompressed images. More recent version of the TIFF standard also allow to include image data that has been compressed without loss (especially united for binary images) or using the JPEG algorithm. As the name suggests the Tiff uses tags with associated values for recording both technical and descriptive metadata. TIFF is very versatile and allows for many different image representations (binary, gray value, palette color and full color). In addition, metadata schemes such as EXIF, IPTC, ICC, XMP etc. may be included by adding the binary data as a special tag (e.g. TIFFTAG XMP). Because of its flexibility and versatility, not all TIFF readers are able to read all TIFF images even if they conform completely to the standard. Within TIFF, the so called *baseline TIFF* defines a least common denominator which all TIFF reads must be able to interpret. However, the baseline TIFF is restricted to the most rudimentary set of metadata.

3.4.1 TIFF/IT

The TIFF/IT has been defined as a subset of the TIFF standard targeted at markets like the exchange of ads for newspapers or magazines and the exchange of pages for magazine printers. It has never been in widespread use and has now largely been replaced by the PDF format.

3.4.2 TIFF/EP

The Tag Image File Format/Electronic Photography (TIFF/EP) is a digital image file format standard (ISO 12234-2, titled ”Electronic still-picture imaging Removable memory Part 2: TIFF/EP image data format”) based on a subset of the TIFF and EXIF standards with the addition of some extensions. The goal was to create a standard format for camera manufacturers to store *raw data images* from camera sensors based on color filter arrays. The standard has not found wide adoption as the Exif/DCF has become the quasi standard of the camera industry. However, the Digital Negative format proposed by Adobe and adopted by some camera manufacturers is loosely based on TIFF/EP and to some degree compatible.

4 TI/A – a recommendation for long term archiving

The versatility of the TIFF format has made it very attractive for memory institution for long term archival of their digital images. However, since the TIFF format offers such a great flexibility, it is not guaranteed that in the future a standard TIFF reader will be able to read some TIFF images. However, the limitations of the baseline TIFF are too severe for many applications in digital archiving. It is important that, besides crucial technical metadata such as ICC color profiles (in case of color images) also important descriptive metadata is stored within the image file. Having descriptive metadata available (such as content description, iconography, copyright and ownership information etc.) is crucial for every archive. Having this information in the same file as the image data guarantees that this information will always be associated with the image.

The TI/A recommendation defines a subset of standard TIFF tags which are either required, optional of forbidden for the purposes of long term archival. Within this context, the goal must be that

1. The image can be opened with standard software even in the far future. Since the TI/A documentation is open and simple, even in case there is no standard software around, a reader can be programmed easily in the future which will render the image correctly and extract the essential descriptive metadata.
2. The image data does not contain features that are not documented and therefore cannot be understood and rendered correctly in the future.

Conforming to the TI/A recommendation will guarantee, that the essential digital information of an image file always can be read and interpreted correctly. Since TI/A is a subset of the TIFF standard, all current TIFF readers are able to correctly and completely render TI/A just out-of-the-box.