Image Capture

IMPACT Best Practice Guide

IMPACT project
Neil Fitzgerald, British Library
Gunter Muhlberger, University Innsbruck Library
Niall Anderson, British Library

Table of Contents

A Best Practice Guide to Image Capture

Executive Summary

One of the most important stages in any digitisation project is to design and implement a consistent approach to capturing digital images. How an institution does this will to a large extent depend on what kind of material it wants to digitise, the aim for which it digitises, available funds and how scaleable the process needs to be. Historically, most text-based digitisation has been done by professional imaging specialists with more or less standard photographic tools and apparatus. Photography is still often the default method for digitally capturing high value material and unusual format items, although development work in this area continues.

As more institutions become interested in mass digitisation, however, there has been a drive towards greater automation of the capture process to reduce costs and increase the volume of items processed, with the overall workflow becoming more integrated. To obtain the best outcomes, all the steps within the workflow and how they impact on each other need to be considered.

The digitisation of newspapers has also improved over time, because of their size and the complexity of their physical make-up, newspapers have always been difficult to digitise directly from the source.

Libraries and archives have thus taken to scanning newspaper images from microfilm. IMPACT presents a case study and recommendations on digitisation from microfilm later in this chapter.

General reading on image capture

Information about standards and methods of capturing digital images is widely available. These sources all provide background to capture approach:

JISC Digital Media; 2008-2011; the Joint Information Systems Committee: http://www.jiscdigitalmedia.ac.uk/ Retrieved 10.03.2011

Digital Imaging Primer; Gueguen, G; 2007; Office of Digital Collections and Research at the University of Maryland: http://www.lib.umd.edu/dcr/publications/DigitalImagingPrimer.pdf Retrieved 10.03.2011

Technical Standards for Digital Conversion of Text and Graphic Materials; 2007; Library of Congress: http://memory.loc.gov/ammem/about/techStandards.pdf Retrieved 12.02.2010

Preservation Imaging Handbook; Metamorfoze, the Dutch National Programme for the Preservation of Paper Heritage: http://www.metamorfoze.nl/methodiek/preservationimaging.html Retrieved 10.03.2011

History and development of scanning for digitisation

The creation of professional scanning devices remains a developing field, with different models and types of scanner specialising in the capture of particular material. In the formative days of mass scanning, a book was placed in a cradle, scanned, and its pages turned one by one by a manual operator.

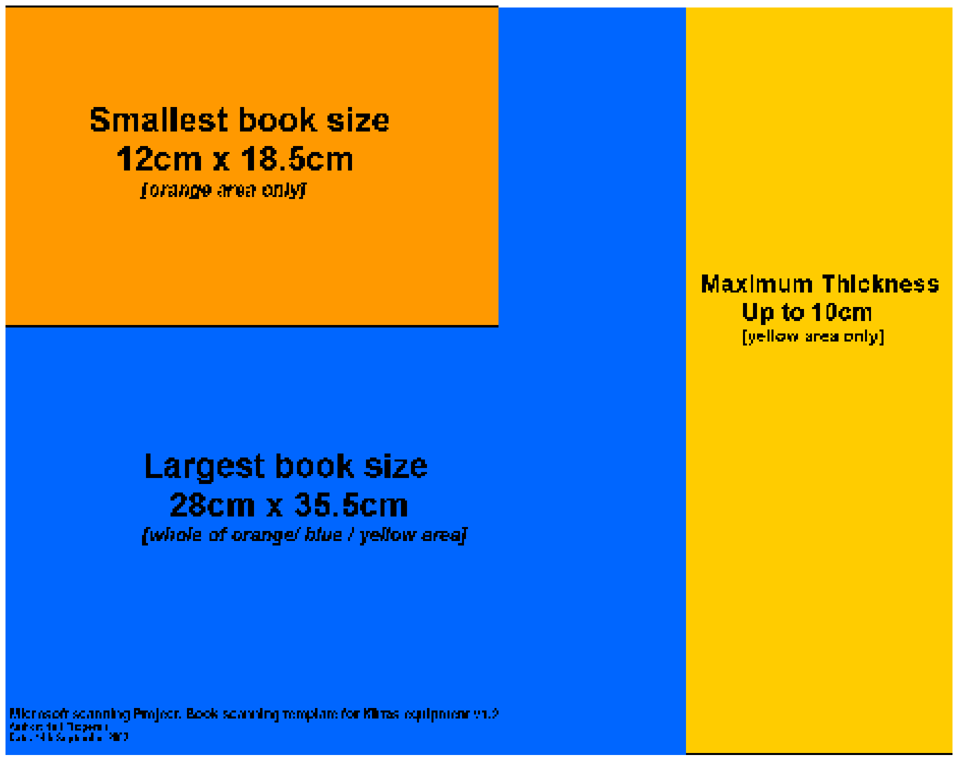
Manual scanners of this sort have growingly been superseded by automatic scanners (also known as scan robots), which offer faster and more cost-effective digitisation for a range of standardised material, books in particular. Appearing on the market in the early years of this century, this new type of scanner automatically turns pages using air jets or a bionic finger. Books are held in a cradle that ensures they won't open wider than 60-120 degrees, minimising the possible damage to the binding. Like manually operated scanners, robot scanners capture two pages at once, thereby significantly increasing throughput of material.

There are several automatic book scanners on the market, each using a different capture approach and turn pages by different methods. Some of these were exhibited at the Bayerische Staatsbibliothek in Munich in July 2008. JISC published a report on the exhibition, concluding that none of the scanners was suitable for all source material, handling requirements, or fast output for mass digitisation[1]. This is in part because not all material can be scanned accurately and efficiently by a specific robot. To digitise all the material in a substantial collection, several scanners of different types would be needed, which has cost and efficiency implications.

At optimal level on ideal material, each of the scan robots surveyed can supply a very high throughput, from 1,500 to 2,500 pages per hour. In the context of a mass digitisation project, however, throughput depends on the overall project workflow and the reliability of the machines themselves. Scan robots can be difficult to operate and maintain, and the accompanying capture software is often complicated. Nonetheless, modern text books can realistically be scanned at a rate of more than 2,000 pages per hour. For older material (pulp-based paper from the 19th-century and earlier), robot scanners perform more slowly. With structurally eccentric or damaged material, manual interventions are still needed.

[1] *Public Exhibition of Automated Book Scanners at BSB-Munich 18-20 June 2008*; Ball, Dr. J.H.: http://digitisation.jiscinvolve.org/files/2008/10/automated-book-scanners-munich-2008-final.pdf Retrieved 13.03.2011

In practice, it is possible to identify material that needs manual intervention prior to digitisation. This involves comparing the specifications of the scanning machine with those of the items to be digitised. In the British Library's 19th Century Book Digitisation Project, for example, books were assessed according to size and processed in batches of similar dimensions to reduce manual intervention. The general book scanning template used by the project can be seen below:



Book template front
[2]

Scan robots range in price from €80,000 to €210,000 per machine, about the same price as other large-format book scanning devices. An example of the usage of a robot scanner is given on the website of Stanford University.[3] Robot scanners can also be seen at work on YouTube.[4]

CCD vs. CMOS

As well as their physical properties and production capacity, capture devices also differ in terms of the type of image sensor they use. In digital still imaging, there are two main types of image sensor (a CCD chip and a CMOS chip), and they both work by capturing available light from a scene and converting it into electrical signals.

A Charged Couple Detector (CCD) chip is an analogue sensor. When light strikes the chip it is held as a small electrical charge in each photo sensor. The charges are converted to voltage one pixel at a time as

[2]Please note that the dimensions cited in this graphic are all 5mm smaller than the Kirtas APT 2400's stated technical specifications. This is because of the difference in size between a book's binding and the pages it contains, and also because the books themselves often had pages of minutely differing size

[3]*Robotic Book Scanning at Stanford University*: http://www-sul.stanford.edu/depts/dlp/bookscanning/ Retrieved 13.03.2011

[4] *SR300: Up to 2400 pages per hour - book scanner ScanRobot*; Treventus, Inc.; 2009 http://uk.youtube.com/watch?v=pb6E4Hrgi9Y&feature=related Retrieved 13.03.2011

they are read from the chip. Additional circuitry in the camera converts the voltage into digital information drawn from what the camera is viewing.

A Complementary Metal–Oxide–Semiconductor (CMOS) chip differs in that the conversion of light to voltage is active and instantaneous. Extra circuitry next to each photo sensor converts the light energy to a voltage. Additional circuitry on the chip may be included to convert the voltage to digital data.

CCD and CMOS imagers were both developed over the late 1960s and the early 1970s. Both offered excellent image capture potential if set up correctly, though CCD was the clear market favourite, largely because the set-up was less intensive from chip level to operational level. CMOS, on the other hand, offered lower power consumption, better integration with hardware, and (related to this) the potential to reuse hardware capabilities – especially digital memory – in a single operation. The downside of CMOS imagers was that they took longer to create, and produced noticeably worse images than CCD imagers unless set up correctly.

While CCD imagers have maintained their position as the most likely sensor to be used for capturing material in high quality, the mainstreaming of silicon-chip production has made CMOS sensors far cheaper to produce, allowing CMOS developers to concentrate on producing image standards as high – or higher – than the standard CCD chip. The cost difference to produce or buy both types of chip has narrowed, though CMOS chips still require buttressing from other software to make them work. CCD and CMOS scanners both exist and function well in their ideal environments, but the following tables summarise their various strengths and weaknesses, and the history of their development to overcome their weaknesses:

Feature and Performance Comparison

Feature	**CCD**	**CMOS**
Signal out of pixel	Electron packet	Voltage
Signal out of chip	Voltage (analog)	Bits (digital)
Signal out of camera	Bits (digital)	Bits (digital)
Fill factor	High	Moderate
Amplifier mismatch	N/A	Moderate
System Noise	Low	Moderate
System Complexity	High	Low
Sensor Complexity	Low	High
Camera components	Sensor + multiple support chips + lens	Sensor + lens possible, but additional support chips common
Relative R&D cost	Lower	Higher
Relative system cost	Depends on Application	Depends on Application

Performance	**CCD**	**CMOS**
Responsivity	Moderate	Slightly better
Dynamic Range	High	Moderate
Uniformity	High	Low to Moderate
Uniform Shuttering	Fast, common	Poor
Uniformity	High	Low to Moderate
Speed	Moderate to High	Higher
Windowing	Limited	Extensive
Anti-blooming	High to none	High

Biasing and Clocking	Multiple, higher voltage	Single, low-voltage

CMOS Development's Winding Path

Initial Prediction for CMOS	**Twist**	**Outcome**
Equivalence to CCD in imaging performance	Required much greater process adaptation and deeper submicron lithography than initially thought	High performance available in CMOS, but with higher development cost than CCD
On-chip circuit integration	Longer development cycles, increased cost, tradeoffs with noise, flexibility during operation	Greater integration in CMOS, but companion chips still required for both CMOS and CCD
Reduced power consumption	Steady improvement in CCDs	Advantage for CMOS, but margin diminished
Reduced imaging subsystem size	Optics, companion chips and packaging are often the dominant factors in imaging subsystem size	CCDs and CMOS comparable
Economies of scale from using mainstream logic and memory foundries	Extensive process development and optimization required	CMOS imagers use legacy production lines with highly adapted processes akin to CCD fabrication

From: *CCD vs. CMOS*; Teledyne Dalsa Ltd.; 2011: http://www.dalsa.com/corp/markets/ccd_vs_cmos.aspx Retrieved 14.03.2011

What image file format is suitable?

This section details image file formats suitable for use in text-related mass digitisation projects, listing the pros and cons of each and providing guidance on which file formats to use according to circumstance. It is intended as a compact and general guide, with links to helpful resources containing more detailed information.

Choosing the appropriate file format

When selecting a file format, several issues have to be considered carefully;

- The intended use of the image files, e.g. for reprint, OCR, web presentation;
- Whether both master and derivative files need to be kept;
- Whether the image needs to be digitally enhanced in order to make it readable on screen, or translatable into OCR text;
- The use of open vs. proprietary file formats;
- Whether the images need to be preserved beyond the life of the project;
- The digital storage required during image processing and afterwards;
- Whether images can be processed in batches;
- Whether to use lossy or losslessly compressed images; or whether to deal with uncompressed master files only;

- The depth of colour representation required by the end resource, and whether a particular file format supports it;
- Whether it is necessary to store additional information alongside the main image data, e.g. metadata, colour management profiles, descriptive information about the image.

The following external resources all provide general information about and references to file formats in connection with digitisation projects:

Choosing a File Format; 2008; JISC Digital Media: http://www.jiscdigitalmedia.ac.uk/stillimages/advice/choosing-a-file-format-for-digital-still-images Retrieved 12.03.2010

Planning an Imaging Project; 1998; Digital Library Federation: http://www.diglib.org/standards/imageoutl1.htm Retrieved 12.03.2011

Digital Imaging Best Practices, Version 2.0; 2008; Bibliographical Center for Research (BCR): http://www.bcr.org/dps/cdp/best/digital-imaging-bp.pdf Retrieved 12.03.2011

Technical Guidelines for Digital Cultural Content Creation Programmes; 2008; MINERVA Europe: http://www.minervaeurope.org/publications/MINERVA-Technical-Guidelines-Version-1.2.pdf Retrieved 12.03.2011

Guidelines for Digitisation Projects; 2002; International Federation of Library Associations and Institutions: http://archive.ifla.org/VII/s19/pubs/digit-guide.pdf Retrieved 12.03.2011

Digital Imaging Primer; 2007; Gueguen, G; Office of Digital Collections and Research at the University of Maryland: http://www.lib.umd.edu/dcr/publications/DigitalImagingPrimer.pdf Retrieved 12.03.2011

Best Practice Guidelines for Digital Collections; 2007; Schreibman, S. et al; Office of Digital Collections and Research at the University of Maryland: http://www.lib.umd.edu/dcr/publications/best_practice.pdf Retrieved 12.03.2011

Image Compression

Image compression is done to minimise digital file size without degrading the quality of the image to an unacceptable level. The reduction in file size allows more images to be stored in a digital memory space, be it a server, a memory stick, or a hard drive. It also reduces the time required for images to be sent over the internet (or downloaded from web pages) and reduces storage costs.

Any use of compression should be carefully considered in a digitisation project. Compression technology may introduce errors in the image, and even small digital artefacts may compromise the entire image. As a result, bitmap images have tended to be standard in text-based mass digitisation projects. A bitmap image is one in which the file size (in bytes) is directly related to the image’s pixel dimensions and colour information, as in the TIFF file format. This has the obvious effect of making bitmap images larger than their compressed equivalents, which can mean that running a text-based mass digitisation project becomes hugely storage intensive and therefore expensive.

For storage, delivery or web presentation, therefore, it may be necessary to compress images. For delivery in particular, the method that preserves the best balance between content and size is a lossless compression format. In lossless compression, the file size is temporarily compressed for delivery to the end user. Once delivered and decompressed it should contain the same information as the original image. It should be borne in mind that the decompression process may slow down the opening of an image at the user end.

By contrast, lossy compression discards some information from the original digital file. In doing so, it achieves a greater size reduction. Where storage needs take priority over image quality (or where an image needs only to be legible), lossy compression may be sufficient. Because most lossy compression involves averaging out small differences between colours and contrast in a digital image, it is often visually undetectable and can be used as a delivery file format for text-based resources.

The most commonly used image file formats and their capabilities are discussed in the next section.

Helpful information and references to image compression:

File Formats and Compression; 2008; JISC Digital Media: http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/file-formats-and-compression Retrieved 12.03.2011

File Formats Benchmark; 2008; The Data Compression Newsblog: http://www.imagecompression.info [http://www.imagecompression.info/] Retrieved 12.03.2011Image quality and long-term preservation

The image quality delivered by a digitisation project to its end users depends on the purpose of the project. But even where a lossy or compressed image is finally delivered to the user, those images will tend to have been derived from a lossless master image. This section looks at the implications of long-term preservation of those master images, and the means necessary to do so.

The optimum condition for long-term storage is to archive all data exactly as created at the moment of capture, but this may not be practicable if the capture device uses a proprietary standard (either commercial or institutional). In cases where the source material itself makes good image capture difficult, it may also be necessary to create and store an optimised version of these files (usually known as “service masters”).

Against this background, the following minimum standards for file formats are recommended:

- For sources containing text with colour illustrations or backgrounds: TIFF or JPEG2000, resolution 300 ppi, bit depth 24, no compression, lossless
- For sources containing text with black-and-white illustrations (pictures or shades of grey): TIFF or JPEG2000, resolution 300 ppi, bit depth 8, no compression, lossless
- For sources containing text with or without line drawings: TIFF or JPEG2000, resolution 400 ppi, bit depth 1, no compression. If necessary for storage purposes, lossless compression is recommended.

Please note in all cases that a bit depth of 8 and a resolution of 400ppi is the minimum recommended where Optical Character Recognition is to be applied. For very small text, a resolution of 600ppi is recommended.

In addition to TIFF, lossless image formats such as PNG and JPEG2000 should be considered. As a rule of thumb, the resolution should allow for the smallest relevant details to be clearly visible when the file display is reduced to 25 percent its original size.

The most commonly used image file formats and their capabilities are discussed in the next section.

Helpful information and references to image quality and long-term preservation:

The Digital Still Image; 2008; JISC Digital Media: http://www.jiscdigitalmedia.ac.uk/stillimages/advice/the-digital-still-image Retrieved 12.03.2011

Recommended Data Formats for Preservation Purposes; 2008; Florida Digital Archive: http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf Retrieved 12.03.2011

Specifications for Files; 2008; Penn State University: https://secureapps.libraries.psu.edu/digital/tools/files.html Retrieved 12.03.2011

Reproduction Qualities of Digital Masters; 1998; Digital Library Federation: http://www.diglib.org/standards/imageoutl4.htm Retrieved 12.03.2011

File Formats for Digital Masters; 1998; Digital Library Federation: http://www.diglib.org/standards/imageoutl5.htm Retrieved 12.03.2011

Preserving Access to Digital Information; 2009; National Library of Australia: http://www.nla.gov.au/padi Retrieved 12.03.20111

Digital Preservation; 2009; Library of Congress: http://www.digitalpreservation.gov/ Retrieved 12.03.2011

Digital Preservation Handbook; 2008; Digital Preservation Coalition: http://www.dpconline.org/advice/introduction-background.html Retrieved 12.03.2011

Digital Best Practices; Washington State Library: http://digitalwa.statelib.wa.gov/newsite/imagequality.htm [http://digitalwa.statelib.wa.gov/newsite/imagequality.htm%20Accessed%2010.02.2010] Retrieved 12.03.2011

Preservation of Digital Projects; Washington State Library: http://digitalwa.statelib.wa.gov/newsite/collection/preservation.htm Retrieved 12.03.2011

Key Terms

Bit depth: The bit depth of an image refers to the number of bits used to describe the colour of each pixel. Greater bit depth allows more colours to be used in the colour palette for the image. 1 bit per pixel will allow black and white, 8 bits per pixel will allow 256 colours, 8 bits per colour component in a RGB image (24 bit) will allow 16777216 colours

Optimisation: As it pertains to digital images, optimisation is any technical process undertaken to prepare an image for final delivery to users. Within IMPACT, optimisation usually refers to processes used to enhance the OCR accuracy of a digital image

PPI: Pixels (or points) per inch. A measurement used to describe both the spatial resolution of a digital image and the physical size of an image printed from it. An image that has a higher number of pixels per inch will show more detail than one which has fewer pixels per inch

Resolution: Spatial resolution, normally expressed as the number of pixels per linear unit e.g., 300 ppi (pixels per inch)

RGB: An additive colour model in which red, green and blue light is combined to create colours, combining full intensities of all three makes white. Digital cameras, scanners and monitors use RGB to record and display colours

Recommended file formats

In general, open standard file formats should be preferred to proprietary ones. Proprietary formats (whether commercial or institutional) may not have technical support into the future. However, some hardware or software used in digital capture may necessitate proprietary formats. In this case, a migration strategy to an open standard should be considered from the outset of the work. Mass digitisation has an obvious role in maintaining the long-term accessibility of collection items, as well as the long-term preservation.

The number of suitable formats for mass digitisation has been somewhat decreased by the practice of mass digitisation. Because scanners, robots and digital cameras invariably produce bitmap images, file formats that do not support such images can be excluded from consideration.

The most commonly used image file formats and their capabilities are discussed in the next section.

Helpful information and references to file formats in general:

Sustainability of Digital Formats; 2007; Library of Congress: http://www.digitalpreservation.gov/formats/ Retrieved 12.03.2011

Comparison of graphics file formats; 2009; Wikipedia: http://en.wikipedia.org/wiki/Comparison_of_graphics_file_formats Retrieved 12.03.2011

Image file formats compared

TIFF (Tagged Image File Format)

TIFF is established as the standard format for archival and master images. It is flexible, open-standard, widespread, cross-platform compatible, considered as a professional image standard and supported by most image processing software.

Properties:

- Usable as master and derivative format
- Support of colour management
- Support of CMYK colour model
- Allows to store metadata via tags
- Large file size when uncompressed
- Usable with lossy or lossless compression

Not supported by all image processing software packages, due to high complexity of format

Helpful information and resources related to the TIFF file format:

TIFF 6.0 Specification; 1992; Adobe Systems Incorporated: http://partners.adobe.com/public/developer/tiff/index.html Retrieved 12.03.2011

Tiff 6.0 specification; 1992; International Telecommunications Union: http://www.itu.int/itudoc/itu-t/com16/tiff-fx/docs/tiff6.pdf Retrieved 12.03.2011

TIFF compression format descriptions; 2009; The Library of Congress. LZW Compression: http://www.digitalpreservation.gov/formats/fdd/fdd000135.shtml Group 4 Compression: http://www.digitalpreservation.gov/formats/fdd/fdd000136.shtml Retrieved 12.03.2011

Tiff Information; Aware Systems: http://www.awaresystems.be/imaging/tiff.html Retrieved 12.03.2011

JPEG File Interchange Format (JFIF)

JPEG is a widely used format based on a lossy compression algorithm allowing the production of very small files that are suitable for online access to full screen image files. It is supported by most image processing software.

Properties:

- Usable as derivative format, commonly used for access copies, especially for web delivered images
- Web browser support
- Metadata support
- Lossy compression
- Truecolor support (up to 16,777,216 colours)

Helpful information and references related to the JPEG file format:

JPEG home page: http://www.jpeg.org/ Retrieved 12.03.2011

*JPEG format description*s: Sustainability of Digital Formats; 2008; Library of Congress: http://www.digitalpreservation.gov/formats/fdd/fdd000018.shtml Retrieved 12.03.2011

JPEG JFIF; 2003; The World Wide Web Consortium (W3C): http://www.w3.org/Graphics/JPEG/ Retrieved 12.03.2011

JPEG2000 and JP2

JPEG2000 is a wavelet-based image compression format developed with the intention of superseding the JPEG standard. The compression performance over JPEG is improved and the data is less easily corrupted. JPEG2000 is licensed format, but the core coding system (JP2) can be obtained free of charge. It is being used increasingly as an archival image format and may be considered as master, archive and access file in the future depending on the popularity and software support.

Properties:

- Usable as master and derivative format within same file
- Allows for metadata storage
- Usable with lossy or lossless compression
- Limited browser support, but has sophisticated delivery options
- Truecolor support

Helpful information and references related to the JPEG2000 file format:

JPEG 2000 home page: http://www.jpeg.org/jpeg2000/ Retrieved 12.03.2011

JPEG2000 format descriptions; Sustainability of Digital Formats; 2009; Library of Congress: http://www.digitalpreservation.gov/formats/fdd/fdd000143.shtml

JP2k Bibliography; JP2k Bibliography; The Wellcome Trust; 2011: http://jp2k-uk.wikidot.com/jp2k-bibliography Retrieved 13.03.2011 [http://jp2k-uk.wikidot.com/jp2k-bibliography%20Retrieved%2013.03.2011]

From TIFF to JPEG 2000? Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings; Kulovitz, Kugler, et al; D-Lib Magazine November/December 2009; http://www.dlib.org/dlib/november09/kulovits/11kulovits.html Retrieved 13.03.2011

PNG (Portable Network Graphics)

PNG is an open bitmapped image format developed with the intention to overcome possible patent problems associated with the GIF format (see below). It is widely used for web images and supports lossless data compression.

Properties:

- Open, extensible format
- Usable as master and derivative format
- Lossless compression
- No loss of information when edited or migrated
- Truecolor support
- Limited nonstandard metadata support
- Web browser support

Helpful information and references related to the PNG file format:

PNG Homepage: http://www.libpng.org/pub/png/ Retrieved 12.03.2011

PNG Format Description: Sustainability of Digital Formats; 2009; Library of Congress: http://www.digitalpreservation.gov/formats/fdd/fdd000153.shtml Retrieved 12.03.2011

Portable Network Graphics; 2003; The World Wide Web Consortium (W3C): http://www.w3.org/TR/PNG/ Retrieved 12.03.2011

GIF (Graphics Interchange Format)

Although the GIF format is out of date from the technical point of view, it is still widely used in web graphics. It supports lossless compression up to 256 colours and is compatible with most image processing software.

Properties:

- Usable as derivative format
- Lossless Compression up to 256 colours
- Web browser support
- Limited metadata support
- Truecolor support
- Supports animation

Helpful information and references related to the GIF file format:

GIF Format Description: Sustainability of Digital Formats; 2009; Library of Congress: http://www.digitalpreservation.gov/formats/fdd/fdd000133.shtml Retrieved 12.03.2011