Digital Conversion Unit
University of Michigan Library
837 Greene St. Room 3212

Digitization (Scanning) Terms and Definitions

(Many shamelessly stolen from other sources without citations)

PPI vs DPI–Technically speaking, PPI (pixels-per-inch) is the way that image resolution is properly described; it affects the size and quality of the image. DPI (dots-per-inch) is better suited to describing the resolution of printers and printed output. PPI and DPI are often used interchangeably.

Optical vs interpolated resolution – Optical resolution is the actual resolution that digitization equipment (scanner, digital camera) is capable of capturing. Interpolation is the computer filling in or guessing to make up the resolution between what can actually be captured and what is being requested. Interpolation is rarely recommended when scanning, but works well for printing images for large posters.

Compression—computer algorithms that make files smaller by replacing redundant data with codes or equations that describe the data more compactly. Compression is either *lossless* or *lossy*.

> **Lossless compression–** no data is lost during the compression process. When the file is decompressed, it has the same number of bits as the original, uncompressed file. Lossless compression does not reduce file size as dramatically as lossy compression, but it is acceptable for use with archival image masters.
>
> **Lossy compression** –these replace redundant data with approximations during the compression process. When the file is decompressed, it has fewer bits than the original, uncompressed file. The amount of lost data depends on the compression type and sometimes user preferences. Lossy compression can achieve incredible file size reductions, but at considerable expense to the quality of an image.

Tagged Image File Format – (abbreviated TIF or TIFF) TIFF is a raster-based image file format. It is used currently as a preservation standard image due to the wide base of support among image viewing software. TIFF by default is an uncompressed form. For bitonal TIFFS, there is a lossless compressed form (Group 4 fax compression or "G4") where the information on the white pixels is thrown out). TIFF also has extensive metadata tags for storing information about each image.

JPEG 2000 (AKA JP2 or .jp2) – a wavelet-based image file compression standard. It has a wide range of compression options available, from lossless to lossy. JP2s also can store metadata in a file header like TIFFs, but JP2s use XML which makes the metadata less standardized but more versatile.

Bitonal – (bilevel, binary, or 1-bit). Bitonal means that each pixel in the image file can only have one of two tonal values, black or white (the tonal value can be stored in one bit of digital data, hence 1-bit or binary). Bitonal images are easier for OCR software to interpret. Because of the limited color range, bitonal images are dramatically smaller than a greyscale or color files.

Continuous tone – (AKA contone) An image in which colors and shades of gray smoothly merge into the neighboring colors or shades, instead of producing distinct, sharply-outlined areas of color or shade. Examples would be traditional black & white or color photography.

Halftone – a printing technique that simulates continuous tone imagery in newspapers and books through the use of dots, varying either in size or in spacing. Halftones are problematic during digitization and can generate distortions in the image when displayed on a monitor.

Dithering – Dithering is another way to simulate continuous tone imagery through dots. Dithering creates dots of equal size and scatters different colored pixels in varying concentrations to simulate shading and blending. Dithering works well for image files with limited color palettes, and can be used to represent black-and-white photographs in bitonal image files and in print.

Grayscale – a black-and-white form of continuous tone imagery. Unlike bitonal images, where one two tonal values can be described, grayscale images are (typically) composed of 256 shades of gray (2^8 or 8-bit), varying from black at the weakest intensity to white at the strongest. High-end scanners are capable of capturing 12-bit (2^{12}) and 16-bit (2^{16}) grayscale. Grayscale images are also called monochromatic, as they only capture one channel of color.

Color Depth – (also known as Bit Depth) The number of possible shades or tonal gradations that a color can have from black to white. A bitonal image is 1-bit (2^1, or 2 colors). Grayscale images are typically 8-bit (2^8 or 256 values). Color images are typically 24-bit (3 colors, 8-bits per color, 16 million values).

Color – ('true color') The representation of color images on a monitor is done with the RGB (red-green-blue) color model. Whereas grayscale uses one color channel, color images use 3 channels (one each for red, green, and blue). Typically, each color channel has 8 bits or 256 values from darkest to lightest, resulting in 24-bit color. On Macintosh computers, 24-bit color is referred to as "millions of colors" because 256 x 256 x 256 = 16,777,216 possible color combinations. As with grayscale, high-end scanners can also capture 36-bit (12 bits per channel) and 48-bit color (16 bits per channel).

RGB – Red, Green, and Blue. This is an additive (transmissive) color model where three colors are added together in varying degrees to get the correct color. White is the addition of all three color channels at their fullest intensity, black is the lack of light across all three color channels. RGB is the most commonly used colorspace for image files for viewing on computer monitors.

CMYK – short for Cyan, Magenta, Yellow, and Key (black). This is a subtractive (reflective) color model used in printing. The three color inks plus black are combined together on a base of white (paper) to form color images. The more ink is added, the darker the image becomes.

Brightness – is an attribute of visual perception in which a source appears to be radiating or reflecting light.

Contrast – in an image is determined by the difference between light and dark tones in a scene. If there is not enough contrast a picture may appear too gray or dull. Bitonal images are considered high contrast, because they are only black and white. High contrast images of text files are easier to read and result in much more accurate OCR results.

Threshold – the cut off value for distinguishing one color from another. When converting grayscale to bitonal images, the threshold level determines whether a gray tone becomes white or black.

Noise – variation in the pixels within a 'solid' color due to equipment or software interpretation of the color variation.

Artifacts – visible defects in an image that are not present in the original item, and were introduced either by software, hardware, or both. Artifacts may include:

- blocky text or obvious zones of color (caused by from heavy file compression)
- bizarre patterns in a halftone illustration when viewed on a monitor (from the lack of proper image editing)
- speckles and noise (from dust on a scanner)
- color bands across the entire image (caused by a bad scanning element)

MD5 checksum – A checksum is generated by software that reads the bits in the file and generates a unique 32 character alphanumeric string. These strings can be used to determine if the file has been altered in any way simply by running it through the checksum algorithm again. If the strings are exactly the same, the file is unchanged. If they are not, the file has been altered in some way. Checksums are vital when transferring files across the Internet or from one storage medium to another.

Wavelet compression – A method of describing raster image files through the amount and distribution of color in an image. Areas of mostly uniform color and few details are seen as flat areas with few data points. Areas with lots of detail and color show up as huge waves or spikes of data. Wavelet compression uses an algorithm to describe the flat areas and waves, with fewer data points along the flat areas and more on the waves. The algorithm also makes it possible to scale the image in size, because the data points are described in relative position, rather than absolute position, to each other.

Scanning Equipment-

Flatbed Scanner – a scanner that is set up like a photocopier: the item is placed face down on a glass plate and a scanning array captures the image from below. Flatbeds work very well with flat, single page items; they can be used to scan bound materials, but this subjects the binding to stress and may cause damage.

Auto feed or sheet feed scanner – an attachment to a flatbed scanner that allows the scanner to work through a stack of loose pages unaided. Some autofeeders redirect the page to the flatbed via rollers, and these may curl the page, causing damage to fragile or brittle paper. Other models of autofeeders have separate scanning arrays which allow the feeder to scan the page without curling the paper.

Planetary/Overhead scanner – a scanner that is set up like a copystand camera: the scanning array is fixed above the scanning bed rather than under (as with flatbed scanners). Overhead scanners are generally less damaging to bound materials.

Digital back or camera back – A traditional film camera which has been fitted with a digital image array in place of the film. Digital backs can be used with 35mm film cameras or copystand cameras.

Fixity – quality of digital object showing that it has not changed over time.

OCR – (Optical Character Recognition) Text in image files are just photographs. They can neither be searched by keywords nor altered. In order to make them searchable, you process the image files through an OCR program, which identifies the letters as appropriate computer code (e.g, A=&0041) which is saved as a text file.

"Raw" or unedited OCR results typically misidentify letters and punctuation, especially in old and difficult typescripts with an accuracy rate of 60-80%. Most OCR programs can identify Roman printed typescript of most European languages. Image files with handwriting, decorated type, German Gothik, non-Western ideographic languages, or with lines of text running in different directions are much harder to identify properly.

Keyed text – text entered into a file by hand rather than generated by OCR.

Born Digital – objects which never had a physical counterpart. Its content was created on a computer from the beginning (i.e., web pages, computer documents, etc.)

Access vs preservation digitization – Access digitization is done with the intent of creating digital files that will provide short-term, easy access online for patrons. Lower resolution/bit depth and lossy compression are favored. Preservation digitization is done with the intent of having the file serve as a long-term "digital master". Preservation digitization often includes high-end scanners, extensive metadata collection, use of uncompressed file formats, and migration/disaster recovery plans.

Metadata – Latin term meaning "information about information." In the digital realm, metadata is data that describes key information about the digital files (image files, text files, digital audio/video) and when appropriate, the original objects they represent. There are different kinds of 'metadata', including:

- bibliographic (author/artist, publisher, publication/release date)
- technical (related to software things like scanning equipment, software programs, settings used to create/modify the file);
- preservation (fixity, checksum information; conservation treatment performed);
- provenance (history of ownership);
- structural (how the original item is put together hierarchically – page numbers, titles, chapter headings, etc.)