HIGH QUALITY SCANNING VS. SMALL FILE SIZE

This post will not talk about scanning only, but about the whole process of turning a sheet of paper, a magazine or a book into a high quality digital document of the smallest possible size. Nothing will be lost during the conversion as lossless compression will be used and OCR will be run on documents. A low quality scan is hard to read (impossible to run OCR). A high quality scan must preserve as much detail as possible, must have the correct page size (when printed at 100%, the resulting copy should be exactly the same size as the original), must load as quick as possible on low-end devices and shouldn't eat the whole drive space. Software processing of the raw image from scanner plays a very important role. Yet, if the image from scanner has a low resolution, further processing is useless and may have negative results. All software used in this tutorial is free (some apps are open-source). But let's start with the basics.

Hardware

What can be used for scanning:

- a scanner. Even the ones from cheap all-in-one printers work great but are (very) slow.
- a photo camera. As mentioned below a 300 DPI A4 page will have 8.7 Megapixels so if you want great results use a camera of at least 8 MP. In order to remove image distortions as much as possible, you can build your own scanning machine (see diybookscanner.org). The main difference from a scanner is that the raw image will have a totally messed up DPI value, so before any other processing this must be fixed. The final document will never have the exact page size of the real page because DPI can only be estimated and image will be warped.

Avoid/skip any processing done by the scanning software or by the photo camera.

Resolution, DPI and size

Resolution actually refers to image width and height in pixels. It is very important as it determines the amount of 'data' the image contains. Each pixel holds a color. A color image using 24 bit color definition takes 3 bytes for every pixel (the color of a pixel is defined by three values: red, green and blue which can have any value between 0 and 255 - that's one byte).

Size refers to the physical size of the image. It is highly related to *DPI* (dots per inch) value. These parameters are contained in the image metadata so as long as DPI * size remains constant, no pixel is changed in the actual image.

Resolution divided by DPI equals the physical size of the image in inches.

When scanning only DPI matters. You can't change the physical size of the paper. Let's assume you scan an A4 sheet at 300 DPI. A4 measures 8.27 by 11.69 inches. Thus the resulting scanned image will have a resolution of 8.27 x 300 = 2481 px width by 11.69 x 300 = 3507 px height. If this is scanned at full color (3 bytes per pixel) and no compression is applied, the resulting file will have a size of 2481 x 3507 x 3 / 1024 / 1024 = 24.89 MB. Imagine a 500 pages book...

DPI is the only parameter that should be corrected when using a photo camera to scan. There's nothing wrong with pixel resolution and it should never be lowered because the reported size is way too much than physical size.

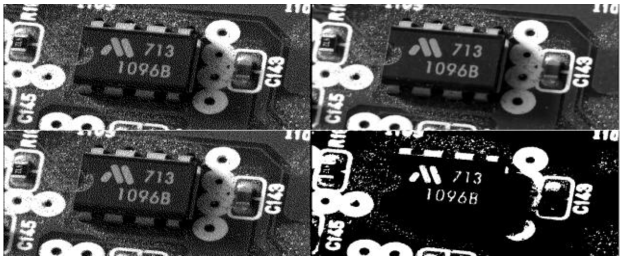
Color depth

Before scanning a page you should correctly determine its color depth. Scanners only have two modes: full color and grayscale. The third offered mode (bitmap, 1-bit, binary) is actually a grayscale scan that is converted by the software to bitmap.

- **Simple page with text:** if the text is black do a **300 DPI grayscale** scan. After software processing the result will be a **600 DPI binary image** (1-bit monochrome, bitmap). If the text is any other color than black do a **300 DPI color** scan. After processing the image will be **600 DPI indexed color** with a limited color palette (the software I'll use doesn't allow less than 4 colors).
- **Black and white page with text and pictures:** you will scan this in grayscale mode. But the pictures matter here because you must figure what their color depth really is in order to make a correct software processing. Take a close look at the printed paper or use a magnifying glass and look for visible dots. If you see dots, then the picture is in 1-bit monochrome. Here comes the tricky part. Zoom the raw image you scanned (try at 300 DPI). Do you still see dots? If yes, then you can process it at **600 DPI monochrome**. If not there are two options:
- The page was printed at a higher DPI than the one you scanned at. You can scan at a higher DPI. I wouldn't recommend this unless you really need to use a higher DPI for the text.
- Process the image as grayscale. This is the selected option also when you can't see any dots on the picture (it was printed in real grayscale mode). The result will be a **300 DPI grayscale** image.

- **Color page:** of course you will scan this in color mode. This mode usually means 24-bit color, a total of 256^3 = 16.7 millions of possible colors. The next step is to determine approximately how many colors are really used on the image. The result will be a **300 DPI image with indexed color** (custom palette).

  300 DPI is a density that usually works for all kind of pages. However you can scan at any higher densities but this will be reflected in the file size and processing/displaying time on computer. The main point is that when converting to a binary image (only two possible color) or to a very small color palette (e.g. maximum 8 colors) always **double** DPI. In all the other situations keep it the same as the scanner.



Monochrome and grayscale images.
Upper row: can you spot the difference? Click the image to see full size.
Left: binary image using only black and white and *Right*: true grayscale image.
Second row: How the above images look after 1-bit monochrome transformation
(left remains unchanged while on the right it could be even worse than this example)

Let's start. Scan or photo a few pages from anything you have around. Here are the required steps to get a quality PDF document.

Source: http://onetransistor.blogspot.in/2014/11/high-quality-scanning-and-low-file-size.html