BEST PRACTICES FOR SCANNING DOCUMENTS

By Frank Harrell

Recommended Scanning Settings.

- Scan at a minimum of 300 DPI, or 600 DPI if expecting to OCR the document
- Scan in full color
- Save pages as JPG files with 75% compression and store them permanently
- Import JPG files into a PDF creator for OCR and distribution purposes

I bought my first scanner in 1992 and have been scanning documents ever since then. Over the last three years, in conjunction with several historical projects, I've scanned roughly 40,000 pages. The following article discusses the techniques I have learned and worked out which have given me the best results for both scanning and storage of historical documents. This article may seem to ramble a bit, and you may wonder when is he going to get to the scanning part, but everything here is very relevant to our subject.

The process of scanning is essentially to take the equivalent of a digital photograph of an object, then saving that photograph in a computer file.

There are two basic types of graphical computer files. Vector and Bitmapped (also called Raster) files.

In a Vector file, lines, dots, curves, shapes and colors are stored using a series of mathematical formulas representing the shapes, values and locations of those features within the image. Vector files work extremely well with line art created on a computer because, consisting of mathematical formulas, the patterns can be scaled to any size, large or small, without losing any detail. However, Vector files do not work well with photographs or scanned items, partly because the files generated quickly become extremely large.

In a Bitmapped graphic file, the content consists of a series of dots, each represented by numbers defining location, hue (color) and brightness. If you look very closely at your computer or TV screen, you can see the image is made up of tiny colored dots. Actually on your computer or TV, each point is made up of three dots, (one red, one green and one blue) but the three together are considered a single dot, which is called a pixel.

Human beings (and all other animals who see as we do) see in a bitmap style. Your eye's retina is made up of tiny light sensitive receptors which respond to various colors within the visible spectrum. We cannot see Vector images in pure form, only computers can do that. In order for us to see a Vector image, our computer must convert that image to a bitmap so it can be shown on the monitor.

Your scanner, digital camera, smartphone, and even the satellites orbiting our planet looking at the weather, all accumulate their images using a bitmapped process.

Looking back at your computer screen, the color and brightness the monitor displays on each dot, is defined by a number stored in the graphic file. There are several ways to define that number, of which one of the more popular being a hexadecimal number. In our everyday life we tend to use a base-10 (decimal) numbering system. However, math isn't restricted to a 10 digit number system. Hexadecimal (HEX) has 16 available digits. To avoid having to create six unique characters on a computer just for this, we take the short cut of using some letters in addition to numbers, so the hexadecimal digits run, 0 1 2 3 4 5 6 7 8 9 a b c d e f. Inside the computer file we use two Hex numbers to represent each of the three primary colors displayed on the monitor (red, green and blue).

The numbers come out looking like this. ff1493 which is a deep pink, or d2b48c which is a tan color. The first two digits control the color red. The second two digits represent green and the last two digits are for blue. If all six digits are ffffff, then all three colors are on full generating a white pixel. If the first four digits are ffff, and the last 2 are 00, then the red and green dots are turned totally on while the blue dot is off (black), then we get a bright pink. This system gives us approximately 65,000,000 possible combinations of colors and brightness levels.

Inside our graphic file, each of those number or letter characters is represented using a string of eight ones and zeroes. (It's actually much more complicated than that, but we don't need to get into those details.) Now take a look at how many dots are on your computer screen in total. Each dot, times six (Hex digits) times eight (binary ones and zeroes). That's a LOT of ones and zeroes.

File types:

There are roughly 100 different kinds of bitmapped file types. The most common found on the Internet are BMP, PNG, GIF, and JPG. (Don't worry; we'll get to PDF files in a bit.) Those names are derived from the file name extension used for each type. By default Windows and Mac computers hide the extension from the user so you may not be aware they are even there. In Windows the file extension determines what program the file will be opened up with. For this article I'm only going to concentrate on BMP and JPG graphics files.

The BMP (Device Independent Bitmapped Image File Format) was developed way back in the early days of graphic computers and is the most simple of all the graphic files. Each pixel is represented by a Hex number. If you think about it, much of the screen area taken up by a document scan is nothing but the white background. In fact, on average, 81% of a page is just white. So having to store all of those redundant identical points as separate numbers is very wasteful. (Yeah, I know, using the phrase 'redundant identical' is redundant and identical.)

Introducing the JPEG file (Joint Photographic Experts Group) - or JPG for short.

JPG files were one of the first, and currently most popular, compressed format graphic file developed specifically for storing photographs. One of the main reasons for becoming the most popular is that, unlike many of the other formats, there were never any royalties associated with the JPG format. In other words, you weren't required to pay the developer for the privilege of using it.

A compressed file is one where, by using mathematical algorithms, the computer is able to reduce the actual size of the file as stored on the computer. Think along the lines of representing all the same color pixels in your picture with one reference number, and then defining all the spots where that color is to be used in shorthand. That would be more efficient than repeating the color for every point. It's really a lot more complicated than that, but I find using this scenario usually gets the point across.

The JPG format uses a sliding compression scale ranging from 1% to 100%. In most programs and systems, the 100% level means there is no compression, while the 1% level is maximum compression. However, I have seen a few programs that have their scale the other way round.

The drawback to JPG compression is the more you compress the file the more detail you lose in your image. So the object is to choose the optimum compression level, to obtain the smallest file size without losing too much detail for any given purpose. Once the detail is lost, there is no way to get it back without reshooting the photo, or rescanning the document. Enhancing doesn't work like they show on TV's CSI. Therefore, I always tell people when saving a JPG image, never overwrite your original, and always save it as a new file.

Introducing the PDF (Portable Document Format) file:

PDF was developed by Adobe Systems in 1993 and to begin with was strictly controlled by them. Adobe was smart in that they made the reader freely available to anyone, while any software for creating a PDF document had to be purchased from them at a very high cost. In the beginning, the cheapest package was over $800. In 2008, they made the standard freely available and now there are a slew of other company versions available including some free ones.

While actually more complicated than this, a PDF file is essentially a wrapper or package containing numbers, strings, arrays, collections and objects such as a JPG images.

The PDF format has a number of advantages and some disadvantages as a distribution platform for images and scans. One advantage is, in the case of distribution of a scanned book, all the pages are together in proper order and can be printed easily. Another advantage is PDF readers are freely available to anyone who wants to download one. If you have a decent PDF generator program, you can OCR the book. OCR stands for Optical Character Recognition, a process where the software attempts to recognize the writing in an image and converts it to computer text which can be searched on within the document or across the web.

One disadvantage of PDF is file size. If you have a PDF containing a long book, then the user must download the entire book in order to see any of it, even if they only want one page. Another disadvantage comes from improper creation of the PDF document, usually caused by not understanding how PDF files are constructed.

Depending on the type of object being scanned, and what software you are using to produce the PDF file, the software may or may not perform OCR automatically. If it does not, then what you have essentially done is to embed your image into the PDF document as a JPG image. Depending on the

source file or scan process, this can produce some very large PDF files. In fact, in some cases, the PDF file may be larger than the source document. It is also possible, depending on settings and software, that the PDF creator will compress your file, possibly loosing quality which is not recoverable without rescanning the document.

In Adobe Acrobat, when you perform OCR on a document, what happens is the software compresses what you see on the screen into a reduced size JPG, then underlays the image with machine readable text which you can search on or copy and paste. So what you are actually looking at is a JPG image. But… you have now lost the higher quality image you may have started with, which you cannot retrieve without rescanning the document.

My recommended process

For my scanning I use a graphics viewing program called IrfanView as my primary tool. It is freeware for personal use and is quite powerful. (http://www.IrfanView.com/) A simple Ctrl+A will open the scanning dialogue within IrfanView. There I can chose the type of scan I intend to do for this job. You have a choice of “Single image” or “Multiple images (batch mode)”. In batch mode I define an output file name which will have a number appended to the end representing the page number. There are also a number of other settings dealing with file naming, saved image type and format as well as compression level and destination of the scanned files. When you click the OK button, your scanner’s application is opened and you go from there. Each page is automatically named and saved in the chosen directory, as a separate file, and then the program is ready for the next page.

Some points on file names.

Do not use spaces in your file names if there is any chance that your documents might end up on the Internet sometime in the future. Spaces are not allowed on the Internet. If you have ever noticed some web pages with a %20 in the name, that is where the original file had a space in name. Yes, it does work; however, it can cause lots of problems when trying to put links to the file in a webpage or email. Instead of spaces use underscores _ hyphens – or MixedCaseFileNameing. They are easier to read than nonmixedcasefilenaming.

If your project will be extensive, it is extremely important to choose a standard format for your file names. For instance, one of my projects involves scanning hundreds of programs from various rodeos that have been held across the country over the last 35 years. Over time, there have been about 40 different rodeo associations, some of which produce a rodeo annually; some only produced a single rodeo. I chose to name my files using a four digit year, followed by the acronym of the association, followed by a word or words indicating the type of document, followed by page numbers beginning with 01 for the front cover.

Another of my projects is my collection of Bell System Practice manuals. In this case my files are named by the practice number, followed by the version, then the published date.

There is nothing wrong in indicating what the document contains within the file name, but the first series of characters in the name must have a logical purpose with some type of order in it. Whatever naming scheme you chose, it must be logical, and you have to stick with it. Otherwise, you will end up with a mess that is impossible to sort out.

Back to scanning.

DPI stands for Dots per Inch. Most modern books are printed at 300 DPI. I have found setting my scanner at 300 DPI for most projects works pretty well if the finished document is only intended for viewing. However, if I plan on performing OCR on the document, I get better results if I scan them at 600 DPI. It takes longer and creates much larger files but the OCR result is far more reliable. If the original DPI is too low, or the source material was poorly printed, I have found OCR results can have as much as 90% failure rate. Best practice is to scan all documents in at least 300 DPI.

I save all my scanned pages in JPG format, compressed to about 75%. After I finish scanning, I import these files into Acrobat, then OCR the document. I then use Acrobat to save my PDF as a "reduced size PDF" (one of the options under the file menu). When I am finished with the book, the original JPG files are stored permanently for future use. In some cases I make them available on one of my websites in case the user needs a higher quality version. Also, as technology improves and better OCR software is developed, I can go back to my original JPG files and run them through the better software for more accurate OCR.

Color Depth:

Scan everything in full color, even if the item is only black and white text. The difference between file sizes of gray scale and full color is minimal. In the beginning, I tried using two bit or B&W scanning but realized early on that two bit scans were only useful for very specialized purposes.

# # #