Scanning Business Documents To PDF Best Practices

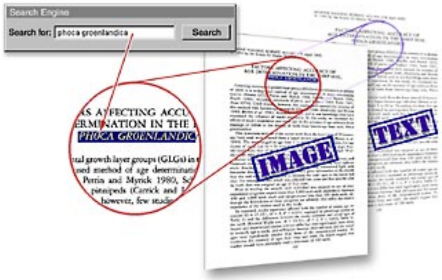
AIIM (Association for Information and Image Management)

This article offers some key best-practices for scanning business documents.

AIIM's Standards Program has developed more formal materials for this purpose, but ISO standards cost money, and most users just want some straightforward (and fast) advice on how to set up their scanning software to avoid 10 MB monster files from a couple of pages. That's what this article is all about.

Background

The very first PDF file I saw was a "PDF/Image + Hidden Text" file, nowadays referred to as PDF/Searchable Image.



I was seriously impressed by the fact that search hits would cause a highlight to appear on the scanned page.

That was cool!

Right at the outset, Adobe recognized that PDF could unify content from any source together in a single document. Whether scanned or produced from an electronic source (such as Word), PDF technology makes it easy to deliver any document in a universal viewer.

Even better, because a PDF can contain and knit together a wide variety of objects, it would be possible to place scanned images and OCR-generated text in alignment with each other.

Thus, the nifty highlighting effect.

How to scan a document

We hear a lot of complaints about PDF files produced from scanned pages. Either the document is too large or the quality is poor. In the vast majority of cases, these complaints arise due to incorrect choices made when scanning or processing.

"Document" is a broad term, but we're going to assume for these purposes that documents are pieces of paper, perhaps notepad or larger in size (although business cards may also be "documents") – the typical stuff of office and institutional records-keeping. If you are (legally, of course) scanning books or magazines for website or archival purposes, other general principles will apply.

We're also going to assume that a PDF file is the logical end-product of your scanning efforts. PDF is the world's de facto electronic document format, after all. You can scan to any format you want, but these days, that scan is likely to become a PDF, and not remain a "naked" TIFF or JPEG.

How to Choose Your Hardware

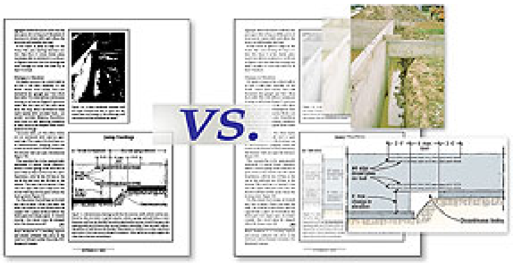
Quite frankly, this article isn't about hardware. I'm going to assume you understand that price, performance and quality vary in scanners just like they do in furniture, cars, and baseball teams. Beyond that, you'll want a scanner that's capable of handling the documents you intend to scan (naturally). For the vast majority of applications, the hardware issues boil down to:

- Do you need to scan in color at all? If not can you also live without grayscale?
- What's the largest size (width and height) document you'll need to scan?
- How fast do you need to scan?
- How high a resolution might you need?
- What software should you use?

These questions are best answered in consultation with your systems integrator – even if that person has a day job at Best Buy. You can get perfectly decent scans from a $100 scanner – and if you only scan 4 pages per month, there's no need to pay for more. There are no general rules about hardware, and this article is about GENERAL principles.

Color or Black and White?

The single most common mistake is to scan in color when it's not necessary.



Color scans range from MUCH larger to COLOSSALLY larger than a *functionally equivalent*black-and-white scan. A 10 page document scanned at 300 dpi black and white might be 0.8 MB; that same document in 24-bit color could be 10, or 30… or even 60 MB!

I say "functionally equivalent", because the vast majority of documents, even if they contain some color, don't need to be scanned in color. Colored letterhead, for example, is rarely a reason to switch over into "color scanning mode". Plenty of forms, invoices, brochures, diagrams and other documents include color for various reasons, but it's rare for those reasons to apply when scanning the documents for records-keeping purposes.

Simply put, the vast majority of business documents may be scanned in black-and-white, even though many of those same documents can and will contain some color. If a regular black-and-white photocopy would be acceptable, then so is a black-and-white scan.

Obviously, if the material includes highlighter markup, color charts and diagrams and color photographic images, then color scanning is indicated. Don't throw away color information if it's needed – just be careful about deciding that you need color, because it's adds expense and hassle to almost every interaction with the document thereafter.

What about Gray?

When reviewing the options in your scanner's software, there may be a "gray" or "grayscale" option. While it's important when capturing black-and-white photographs or content that explicitly includes shades of gray, this can be important – otherwise, it adds most of the complications of color, even if at a somewhat reduced cost in file-size.

File Type and Compression

Your choice of file type is driven by your decision on how to scan – black and white or color.

For black-and-white scans intended for conversion to PDF, the most common way to go is "CCITT-G4 TIFF", more colloquially known as "G4 tiff". These may be either single or multipage; either way they convert to PDF files very nicely.

Color scanning and color image processing is more challenging because different compression systems are preferable for different types of images. Photographic images, for example, get JPEG or JPEG-2000 "lossy" compression, a way to reduce the image-size with (hopefully) minimal impact on quality. Screen-shots and vector artwork may be best handled as PNG files.

Leave uncompressed files to the archivists – no-one else ever wants to deal with business document images without some sort of compression!

Some image conversion systems use a variety of advanced software to improve results or reduce file-size (or both). Adobe Systems includes several optimization options for scanned documents. In addition, Adobe also provides an option they call ClearScan, which recognizes the text and images in each scan and converts the content appropriately.

Luratech and Cvision offer high-volume PDF Compressor products which can dramatically reduce file size for both scanned and conventional PDFs. Vendors such as PrimeRecognition, ABBYY and Nuance offer packaged OCR solutions that include the option of human cleanup of OCR errors, a vital factor in accessibility and search optimization.

Resolution

Once you’ve decided about color and image type, the next question is image resolution, often expressed in “dots per inch” or “dpi”.

Here, it’s a simple question of choosing the right resolution for your content. Color documents can be scanned at lower resolutions, but the file size, on a per-page basis, will still tend to exceed black and white scans.

Black-and-white documents

- Lowest acceptable resolution = 200 dpi (equal to “high” fax quality)
- Conventional resolution = 300 dpi
- Highest conventional resolution = 600 dpi

There’s not a lot of reason to worry about compression with black-and-white documents, and there’s no need for lossy compression, as there is with color scans. With G4 TIFF images, you’ve got a decent lossless compression, ready for conversion to PDF. JBIG2, while technically lossy, is commonly understood as “perceptually lossless”, and maybe safely used as well. Reserve 600 dpi for documents with VERY small type and lots of small details.

Color documents (and color compression)

- Lowest acceptable resolution = 150 dpi
- Conventional resolution = 200 dpi
- Highest conventional resolution = 300 dpi
- Archive resolution = 600 dpi (and lossless, ie, uncompressed TIFF)

In color content, compression works by throwing away some of the information in the hopes that you won’t notice. Usually it works. If you’ve chosen the right resolution, the only remaining question is compression.

In color documents, resolution interacts with compression in several ways; but in general, the lower the resolution, the more sensitive you’ll be to compression. At resolutions of 150 dpi and under, strong JPEG compression can convert text into a muddy blur of pixels.

Select the resolution you think you want to use, then find the smallest, lightest, most italicized type in your documents, and be sure that it still OCRs adequately. Choose higher resolutions (300 and above) for documents that include very small text (ie, below 8 point).

Conclusion

Scanners are cheap, and scanning is easy. The results are not guaranteed, but a little forethought will ensure that your end-product PDF files are as small as reasonably possible while retaining the information that caused you to scan them in the first place.

Source: AIIM (Association for Information and Image Management)
http://www.aiim.org/
May 6, 2012