NOT ALL PDF COMPRESSION IS CREATED EQUAL

Part I – Scanned Documents

WHITE PAPER

As organizations struggle to keep up with the demands of Enterprise Content Management, many are wisely turning to PDF as the format on which to standardize their document content. The PDF format offers many advantages, and one of them is the versatility of supporting many ways to store an image, some of which can yield much smaller files sizes than others.

> **According to the PDF Association, 2.2 billion PDF files are on the public web, and a staggering 2 billion PDFs are opened every year in Outlook.com.**

These days, various companies offer software or online services to create compressed PDF files from scanned images files or existing PDF files. On the surface, a given solution might seem to do a great job of getting the file size down, sometimes even for free. However, what might look good initially may come with some caveats, and it's important to carefully examine the results to see what you're really getting. Let's take a closer look.

SACRIFICING DOCUMENT CONTENT

Some solutions seek to reduce file size by simply "downsampling" the page image, i.e., resampling it at a lower resolution. For example, they may take a page that is originally scanned at 300 dots per inch (dpi) and resample it to use only 150 dpi. While having fewer pixels to encode can significantly shrink the size of an image file, downsampling can degrade the image quality if you go below 300 dpi. In some use cases, a lower-resolution image may still be acceptable, but this is something to be aware of when choosing a compression solution.

Alternatively, if the original file is a PDF, some vendors will take the approach of stripping metadata from the PDF file to reduce file size. While some metadata may not be necessary, one must be careful here and know exactly what is being stripped out. Some metadata may be important for indexing or for tracking the origins of the file. And stripping metadata won't be very effective in all cases either. It will only be noticeable when the amount of metadata that the PDF file contains is large compared to the overall size of the file.

COMPRESSION WITHOUT COMPROMISE

A sophisticated PDF converter will make use of the most advanced compression technologies currently supported. With converters that use JBIG2, JPEG2000, and MRC compression, you can achieve a high level of compression while maintaining a high-quality image. But the story does not end there; how well those compression technologies are implemented can have a big impact as well. Below we'll examine each of these technologies in more detail.

JBIG2 COMPRESSION

If your scanned documents are purely black and white, JBIG2 compression can provide significant savings in file size. Unlike generalized compression algorithms such a ZIP or older image encoding technologies such as CCITT Group 4, JBIG2 leverages awareness of the actual characters and symbols on the image, allowing for better results. The basic concept is as follows: Let's say, for example, the letter 'a' appears many times on a page in the same font and size. The JBIG2 algorithm will encode just one instance of the arrangement of these pixels, along with a mapping of where this symbol appears all over the page.

The problem is that with a scanned document, few instances of the same character are truly identical pixel-for-pixel, due to artifacts of the scanning process. While you may not notice this on-screen at a typical zoom level for reading, a higher zoom level will reveal very slight shifts in some of the pixels along the edges. If the image were left exactly as-is, JBIG2 may not yield very much savings in file size, since there probably wouldn't be many symbols that match each other precisely.

That's where a matching algorithm comes in. If two symbols are e□ectively the same, we shouldn't care about these slight pixel shifts. A matcher algorithm enables the software to choose one of these images to represent both. Some refer to this as "lossy" encoding, though in fact it does not cause any perceptible visual degradation, and the results can yield compression rates of typically between 60% and 20% of the original file size. Actual compression rates will depend on a few factors such as scan resolution, background pixel noise, page coverage, and how the document was originally encoded. Some products offering JBIG2 encoding will further enhance compression by matching symbols even across pages, or by matching groups of symbols such as whole words.

The trick, of course, is ensuring that two symbols really are the same. If not done well, using lossy encoding can lead to mismatched symbols and replacement of a character on the page with a different character. The poorer the quality of the original scanned image, the greater the risk of a mismatch if the matching algorithm is not suffciently robust. While JBIG2 is a well-defined compression standard supported for years by PDF viewers, choosing PDF compression software with a sophisticated matcher can make all the difference in the size and fidelity of your files.

Nonetheless, in some markets there is still resistance to the use of lossy JBIG2 encoding, regardless of how safe a given matcher may be. So JBIG2 compression vendors will commonly offer a lossless mode as well. To compensate for the issue discussed above concerning matching symbols with slight differences, more sophisticated solutions will still use a lossy matcher here, but for the reference bitmap they will only include the pixels they have in common. Pixels on the edges that are not common to all matching symbols will be stored separately for each symbol instance, and the combination of the two will reconstruct the lossless image. For a high-quality scan, this approach may still help lossless encoding compress a page reasonably well, though not as well as lossy encoding.

JPEG AND JPEG2000 COMPRESSION

JBIG2 compression works only for black and white images. When it comes to color or grayscale images, whose color may vary continuously from pixel to pixel, we must turn to other compression methods. JPEG has been a popular standard for many years, and if an image was encoded as a TIFF, JPEG can reduce the file size quite a bit. However, JPEG encoding can create some unwanted pixel noise, especially at lower target quality levels.

A newer algorithm is JPEG2000, designed to be the successor to JPEG. JPEG2000 can yield compression rates that are 5 to 30 times smaller than JPEG while still maintaining image quality at a respectable level. One thing to know about JPEG2000 compression, however, is that if your images very large (e.g. 600 dpi or large page size), JPEG2000-encoding them may cause PDF pages to lag noticeably while they are being loaded into a viewer. A robust compression solution using JPEG2000 should provide the option to automatically fall back to JPEG when it detects that JPEG2000 encoding might cause this issue.

MRC COMPRESSION

JPEG2000 can achieve great savings for color or grayscale images, but it's possible to do even better.

MRC, which stands for Mixed Raster Content, is a methodology that begins with the premise that the human eye is significantly more sensitive to the quality of an image's foreground (e.g., text) than its background. Given that, MRC separates the image into 3 different layers: a foreground, a background, and a binary mask. The background layer is smoothed using pixel averaging, and the mask determines where the higher-resolution foreground layer should be superimposed, sort of like a stencil. While some smoothing takes place even in the foreground layer, the edges of the mask ensure that the displayed parts of the foreground maintain a crisp edge. The foreground and background are further compressed using JPEG2000 or JPEG, while the mask may be compressed with JBIG2 or CCITT Group 4. All in all, compression rates of as much as 100:1 can be achieved using MRC while still maintaining very high fidelity to the original document, and all image layers are encoded using non-proprietary, industry-standard encoding algorithms.

Since the PDF format supports multiple image layers, it's well suited towards MRC compression. Although MRC is not an encoding standard unto itself like JBIG2, this layer-separation methodology is similarly well known and has been implemented by various vendors. And like JBIG2, the implementation is where a given product supporting MRC will set itself apart from others. Determining which details to pull into the foreground layer or leave in the background layer is a complicated art that not all vendors do as well.

One thing that MRC inherently tends to have trouble with is photographs. In some cases, elements in the photograph might be lifted into the foreground layer, causing some details to appear either smudged or overemphasized. Yet even here, MRC compression that employs a good picture detection algorithm can help avoid

this, by excluding all areas from the mask that are detected to be part of a photograph. That said, no picture detection algorithm is foolproof, so in some cases it may be desirable to either increase the target MRC quality level or turn it off completely in favor of JPEG2000 compression. That will increase file size, but it will safely preserve photo quality.

OTHER FACTORS

While compression ratios and image quality are key criteria in choosing a document compression solution, there are other important factors to consider as well. For example, if you need to make your documents searchable, make sure you choose a solution with a high-quality OCR engine. Or if your inbound PDF documents contain a mix of scanned and born-digital content, such as font-based text with images, check whether a given solution is able to compress your image content while not flattening your text to an image. Not all solutions offer this.

CONCLUSION

Choosing PDF compression software can be challenging and understanding what a solution can really provide is key. With what we've discussed here, you're armed with a lot more knowledge to help identify a robust solution that doesn't necessitate important tradeoffs. Reputation is important in this industry as well, so it's advised to do some research into the solutions you're considering. But no matter what someone else says, whenever an evaluation copy is available, try the software on your own files and see what works best for your use case. Everyone's data is different, so it's imperative to see firsthand what works best for you.

Foxit
39355 California Street, Suite 302
Fremont, CA 94538, USA

Sales: 1-866-680-3668
Support: 1-866-693-6948
Or 1-866-MYFOXIT

Support Center
www.foxit.com