Yongli Zhou

Are Your Digital Documents Web Friendly?: Making Scanned Documents Web Accessible

The Internet has greatly changed how library users search and use library resources. Many of them prefer resources available in electronic format over traditional print materials. While many documents are now born digital, many more are only accessible in print and need to be digitized. This paper focuses on how the Colorado State University Libraries creates and optimizes text-based and digitized PDF documents for easy access, downloading, and printing.

To digitize print materials, we normally scan originals, save them in archival digital formats, and then make them Web-accessible. There are two types of print documents, graphic-based and text-based. If we apply the same techniques to digitize these two different types of materials, the documents produced will not be Web-friendly.

Graphic-based materials include archival resources such as historical photographs, drawings, manuscripts, maps, slides, and posters. We normally scan them in color at a very high resolution to capture and present a reproduction that is as faithful to the original as possible. Then we save the scanned images in TIFF (Tagged Image File Format) for archival purposes and convert the TIFFs to JPEG (Joint Photographic Experts Group) 2000 or JPEG for Web access. However, the same practice is not suitable for modern text-based documents, such as reports, journal articles, meeting minutes, and theses and dissertations. Many old text-based documents (e.g., aged newspapers and books), should be treated as graphic-based material. These documents often have faded text, unusual fonts, stains, and colored background. If they are scanned using the same practice as modern text documents, the document created can be unreadable and contain incorrect information. This topic is covered in the section "Full-Text Searchable PDFs and Troubleshooting OCR Errors."

Currently, PDF is the file format used for most digitized text documents. While PDFs that are created from high-resolution color images may be of excellent quality, they can have many drawbacks. For example, a multipage PDF may have a large file size, which increases download time and the memory required while viewing. Sometimes the download takes so long it fails because a time-out error occurs. Printers may have insufficient memory to print large documents. In addition, the Optical Character Recognition (OCR) process is not accurate for high-resolution images in either color or grayscale. As we know, users want the ability to easily download, view, print, and search online textual documents. All of the drawbacks created by high-quality scanning defeat one of the most important purposes of digitizing text-based documents: making them accessible to more users.

This paper addresses how Colorado State University Libraries (CSUL) manages these problems and others as staff create Web-friendly digitized textual documents. Topics include scanning, long-time archiving, full-text searchable PDFs and troubleshooting OCR problems, and optimizing PDF files for Web delivery.

Preservation Master Files and Access Files

For digitization projects, we normally refer to images in uncompressed TIFF format as master files and compressed files for fast Web delivery as access files. For text-based files, access files normally are PDFs that are converted from scanned images.

"BCR's CDP Digital Imaging Best Practices Version 2.0" says that the master image should be the highest quality you can afford, it should not be edited or processed for any specific output, and it should be uncompressed.[1] This statement applies to archival images, such as photographs, manuscripts, and other image-based materials. If we adopt the same approach for modern text documents, the result may be problematic. PDFs that are created from such master files may have the following drawbacks:

- Because of their large file size, they require a long download time or cannot be downloaded because of a timeout error.
- They may crash a user's computer because they use more memory while viewing.
- They sometimes cannot be printed because of insufficient printer memory.
- Poor print and on-screen viewing qualities can be caused by background noise and bleed-through of text. Background noise can be caused by stains, highlighter marks made by users, and yellowed paper from aged documents.
- The OCR process sometimes does not work for high-resolution images.
- Content creators need to spend more time scanning images at a high resolution and converting them to PDF documents.

Web-friendly files should be small, accessible by most users, full-text searchable, and have good

Yongli Zhou is Digital Repositories Librarian, Colorado State University Libraries, Colorado State University, Fort Collins, Colorado

on-screen viewing and print qualities. In the following sections, we will discuss how to make scanned documents Web-friendly.

Scanning

There are three main factors that affect the quality and file size of a digitized document: file format, color mode, and resolution of the source images. These factors should be kept in mind when scanning text documents.

File Format and Compression

Most digitized documents are scanned and saved as TIFF files. However, there are many different formats of TIFF. Which one is appropriate for your project?

- *TIFF*: Uncompressed format. This is a standard format for scanned images. However, an uncompressed TIFF file has the largest file size and requires more space to store.
- *TIFF G3*: TIFF with G3 compression is the universal standard for faxs and multipage line-art documents. It is used for black-and-white documents only.
- *TIFF G4*: TIFF with G4 compression has been approved as a lossless archival file format for bitonal images. TIFF images saved in this compression have the smallest file size. It is a standard file format used by many commercial scanning vendors. It should only be used for pages with text or line art. Many scanning programs do not provide this file format by default.
- *TIFF Huffmann*: A method for compressing bi-level data based on the CCITT Group 3 1D facsimile compression schema.
- *TIFF LZW*: This format uses a lossless compression that does not discard details from images. It may be used for bitonal, grayscale, and color images. It may compress an image up to 50 percent. Some vendors hesitate to use this format because it was proprietary; however, the patent expired on June 20, 2003. This format has been widely adopted by much software and is safe to use. CSUL saves all scanned text documents in this format.
- *TIFF Zip*: This is a lossless compression. Like LZW, ZIP compression is most effective for images that contain large areas of single color. [2]
- *TIFF JPEG*: This is a JPEG file stored inside a TIFF tag. It is a lossy compression, so CSUL does not use this file format.

Other image formats:

- *JPEG*: This format is a lossy compression and can only be used for nonarchival purposes. A JPEG image can be converted to PDF or embedded in a PDF. However, a PDF created from JPEG images has a much larger file size compared to a PDF created from TIFF images.
- *JPEG 2000*: This format's file extension is .jp2. This format offers superior compression performance and other advantages. JPEG 2000 normally is used for archival photographs, not for text-based documents.

In short, scanned images should be saved as TIFF files, either with compression or without. We recommend saving text-only pages and pages containing text and/or line art as TIFF G4 or TIFF LZW. We also recommend saving pages with photographs and illustrations as TIFF LZW. We also recommend saving pages with photographs and illustrations as TIFF uncompressed or TIFF LZW.

How Image Format and Color Mode Affect PDF File Size

Color mode and file format are two factors that determine PDF file size. Color images typically generate the largest PDFs and black-and-white images generate the smallest PDFs. Interestingly, an image of smaller file size does not necessarily generate a smaller PDF. Table 1 shows how file format and color mode affect PDF file size.

The source file is a page containing black-and-white text and line art drawings. Its physical dimensions are 8.047" by 10.893". All images were scanned at 300 dpi.

CSUL uses Adobe Acrobat Professional to create PDFs from scanned images. The current version we use is Adobe Acrobat 9 Professional, but most of its features listed in this paper are available for other Acrobat versions. When Acrobat converts TIFF images to a PDF, it compresses images. Therefore a final PDF has a smaller file size than the total size of the original images. Acrobat compresses TIFF uncompressed, LZW, and Zip the same amount and produces PDFs of the same file size. Because our in-house scanning software does not support TIFF G4, we did not include TIFF G4 test data here. By comparing similar pages, we concluded that TIFF G4 works the same as TIFF uncompressed, LZW, and Zip. For example, if we scan a text-based page as black-and-white and save it separately in TIFF uncompressed, LZW, Zip, or G4, then convert each page into a PDF, the final PDF will have the same file size without a noticeable quality difference. TIFF JPEG generates the smallest PDF, but it is a lossy format, so it is not recommended. Both JPEG and JPEG 2000 have smaller file sizes but generate larger PDFs than those converted from TIFF images.

Recommendations

1. Use TIFF uncompressed or LZW in 24 bits color for pages with color graphs or for historical documents.
2. Use TIFF uncompressed or LZW

in grayscale 8 bits for pages with black-and-white photographs or grayscale illustrations.
3. Use TIFF uncompressed, LZW, or G4 in black-and-white for pages containing text or line art.

To achieve the best result, each page should be scanned accordingly. For example, we had a document with a color cover, 790 pages containing text and line art, and 7 blank pages. We scanned the original document in color at 300 dpi. The PDF created from these images was 384 MB, so large that it exceeded the maximum file size that our repository software allows for uploading. To optimize the document, we deleted all blank pages, converted the 790 pages with text and line art from color to black-and-white, and retained the color cover. The updated file has a file size of 42.8 MB. The example can be accessed at http://hdl.handle.net/10217/3667. Sometimes we scan a page containing text and photographs or illustrations twice, in color or grayscale and in black-and-white. When we create a PDF, we combine two images of the same page to reproduce the original appearance and to reduce file size. How to optimize PDFs using multiple scans will be discussed in a later section.

How Image Resolution Affects PDF File Size

Before we start scanning, we check with our project manager regarding project standards. For some funded projects, documents are required to be scanned at no less than 600 dpi in color. Our experiments show that documents scanned at 300 or 400 dpi are sufficient for creating PDFs of good quality. Resolutions lower than 300 dpi are not recommended because they can degrade image quality and produce more OCR errors. Resolutions higher than 400 dpi also are not recommended because they generate large files with little improved on-screen viewing and print quality. We compared PDF files that were converted from images of resolutions at 300, 400, and 600 dpi. Viewed at 100 percent, the difference in image quality both on screen and in print was negligible. If a page has text with very small font, it can be scanned at a higher resolution to improve OCR accuracy and viewing and print quality.

Table 2 shows that high-resolution images produce large files and require more time to be converted into PDFs. The time required to combine images is not significantly different compared to scanning time and OCR time, so it was omitted. Our example is a modern text document with text and a black-and-white chart.

Most of our digitization projects do not require scanning at 600 dpi; 300 dpi is the minimum requirement. We use 400 dpi for most documents and choose a proper color mode for each page. For example, we scan our theses and dissertations in black-and-white at 400 dpi for bitonal pages. We scan pages containing photographs or illustrations in 8-bit grayscale or 24-bit color at 400 dpi.

Table 1. File format and color mode versus PDF file size

File Format	Scan Specifications	TIFF Size (KB)	PDF Size (KB)
TIFF	Color 24 bits	23,141	900
TIFF LZW	Color 24 bits	5,773	900
TIFF ZIP	Color 24 bits	4,892	900
TIFF JPEG	Color 24 bits	4,854	873
JPEG 2000	Color 24 bits	5,361	5,366
JPEG	Color 24 bits	4,849	5,066
TIFF	Grayscale 8 bits	7,729	825
TIFF LZW	Grayscale 8 bits	2,250	825
TIFF ZIP	Grayscale 8 bits	1,832	825
TIFF JPEG	Grayscale 8 bits	2,902	804
JPEG 2000	Grayscale 8 bits	2,266	2,270
JPEG	Grayscale 8 bits	2,886	3,158
TIFF	Black-and-white	994	116
TIFF LZW	Black-and-white	242	116
TIFF ZIP	Black-and-white	196	116

Note: Black-and-white scans cannot be saved in JPEG, JPEG 2000, or TIFF JPEG formats.

Other Factors that Affect PDF File Size

In addition to the three main factors we have discussed, unnecessary edges, bleed-through of text and graphs, background noise, and blank pages also increase PDF file sizes. Figure 1 shows how a clean scan can largely reduce a PDF file size and

simultaneously improve its viewing and print quality.

Recommendations

1. *Unnecessary edges*: Crop out.
2. *Bleed-through text or graphs*: Place a piece of white or black card stock on the back of a page. If a page is single sided, use white card stock. If a page is double sided, use black card stock and increase contrast ratio when scanning. Often color or grayscale images have bleed-through problems. Scanning a page containing text or line art as black-and-white will eliminate bleed-through text and graphs.
3. *Background noise*: Scanning a page containing text or line art as black-and-white can eliminate background noise. Many aged documents have yellowed papers. If we scan them as color or grayscale, the result will be images with yellow or gray background, which may increase PDF file sizes greatly. We also recommend increasing the contrast for better OCR results when scanning documents with background colors.
4. *Blank pages*: Do not include if they are not required. Blank pages scanned in grayscale or color can quickly increase file size.

PDF and Long-Term Archiving PDF/A

PDF vs. PDF/A

PDF, short for Portable Document Format, was developed by Adobe as a unique format to be viewed through Adobe Acrobat viewers. As the name implies, it is portable, which means the file created on one computer can be viewed with an Acrobat viewer on other computers, handheld devices, and on other platforms.[3]

A PDF/A document is basically a traditional PDF document that fulfills precisely defined specifications. The PDF/A standard aims to enable the creation of PDF documents whose visual appearance will remain the same over the course of time. These files should be software-independent and unrestricted by the systems used to create, store, and reproduce them.[4] The goal of PDF/A is for long-term archiving. A PDF/A document has the same file extension as a regular PDF file and must be at least compatible with Acrobat Reader 4.

There are many ways to create a PDF/A document. You can convert existing images and PDF files to PDF/A files, export a document to PDF/A format, scan to PDF/A, to name a few. There are many software programs you can use to create PDF/A, such as Adobe Acrobat Professional 8 and later versions, Compart AG, PDFlib, and PDF Tools AG.

Many PDF files cannot be saved as PDF/A files. If an error occurs when saving a PDF to PDF/A, you may use Adobe Acrobat Preflight (Advanced > Preflight) to identify problems. See figure 2.

Errors can be created by non-embedded fonts, embedded images with unsupported file compression, bookmarks, embedded video and audio, etc. By default, the Reduce File Size procedure in Acrobat Professional compresses color images using JPEG 2000 compression. After running the Reduce File Size procedure, a PDF may not be saved as a PDF/A because of a "JPEG 2000 compression used" error. According to the PDF/A Competence Center, this problem will be eliminated in the second part of the PDF/A standard—PDF/A-2 is planned for 2008/2009. There are many other features in new PDFs; for example, transparency and layers will be allowed in PDF/A-2.[5] However, at the time this paper was written PDF/A-2 had not been announced.[6]

Table 2. Color Mode and Image Resolution vs. PDF File Size

Color mode	Resolution (DPI)	Scanning time (sec.)	OCR time (sec.)	TIFF LZW (KB)	PDF size (KB)
color	600	100	N/A*	16,498	2,391
color	400	25	35	7,603	1,491
color	300	18	16	5,763	952
grayscale	600	36	33	6,097	2,220
grayscale	400	18	18	2,888	1370
grayscale	300	14	12	2,240	875
B/W	600	12	18	559	325
B/W	400	10	10	333	235
B/W	300	8	9	232	140

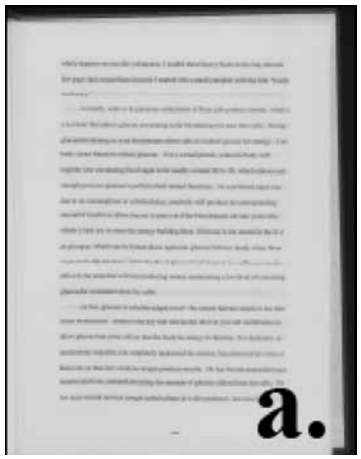
*N/A due to an OCR error

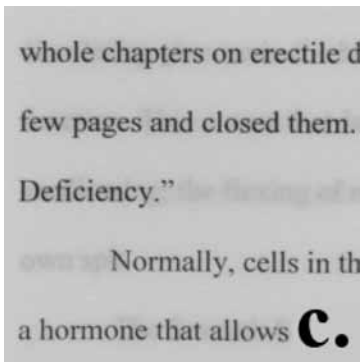
Full-Text Searchable PDFs and Trouble-shooting OCR Errors

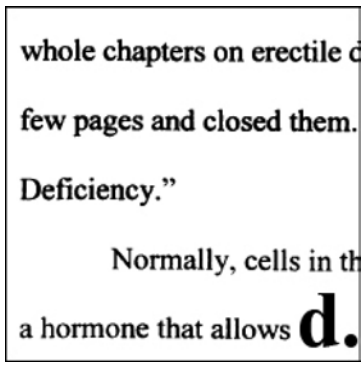
A PDF created from a scanned piece of paper is inherently inaccessible because the content of the document is an image, not searchable text. Assistive technology cannot read or extract the words, users cannot select or edit the text, and one cannot manipulate the PDF document for accessibility. Once OCR is properly applied to the scanned files, however, the image becomes searchable text with selectable graphics, and one may apply other accessibility features to the document.[7]

Acrobat Professional provides three OCR options:

1. *Searchable Image (Exact)*: Ensures that text is searchable and selectable. This option keeps the original image and places an invisible text layer over it. Recommended for cases requiring maximum fidelity to the original image.[8] This is the only option used by CSUL.
2. *Searchable Image*: Ensures that text is searchable and selectable. This option keeps the original image, de-skews it as needed, and places an invisible text layer over it. The selection for downsample images in this same dialog box determines whether the image is downsampled and to what extent.[9] The downsampling combines several pixels in an image to make a single larger pixel; thus some information is deleted from the image. However, downsampling does not affect the quality of text or line art. When a proper setting is used, the size of a PDF can be significantly reduced with little or no loss of detail and precision.
3. *ClearScan*: Synthesizes a new Type 3 font that closely approximates the original, and preserves the page background using a low-resolution copy.[10] The final PDF is the same as a born-digital PDF. Because Acrobat cannot guarantee the accuracy of







Dimensions: 9.127″ X 11.455″
Color Mode: grayscale
Resolution: 600 dpi
TIFF LZW: 12.7 MB
PDF: 1,051 KB

Dimensions: 8″ X 10.4″
Color Mode: black-and-white
Resolution: 400 dpi
TIFF LZW: 153 KB
PDF: 61 KB

Figure 1. PDFs Converted from different images: (a) the original PDF converted from a grayscale image and with unnecessary edges; (b) updated PDF converted from a black-and-white image and with edges cropped out; (c) screen viewed at 100 percent of the PDF in grayscale; and (d) screen viewed at 100 percent of the PDF in black-and-white.

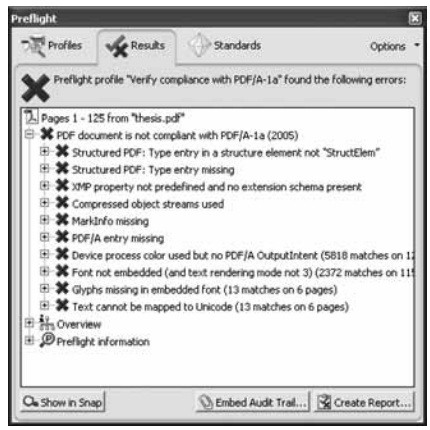


Figure 2. Example of Adobe Acrobat 9 Preflight

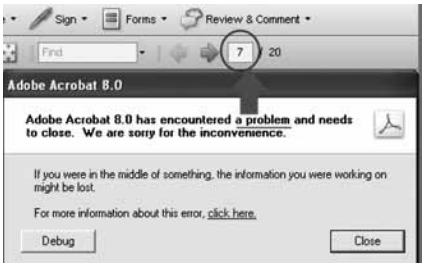


Figure 3. Adobe Acrobat 8 Professional crash window

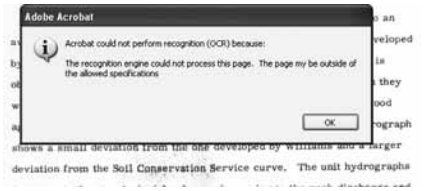


Figure 4. "Could not perform recognition (OCR)" error

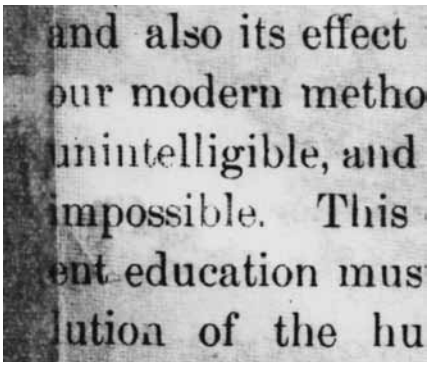
OCRed text at 100 percent, this option is not acceptable for us.

For a tutorial on to how to make a full-text searchable PDF, please see appendix A.

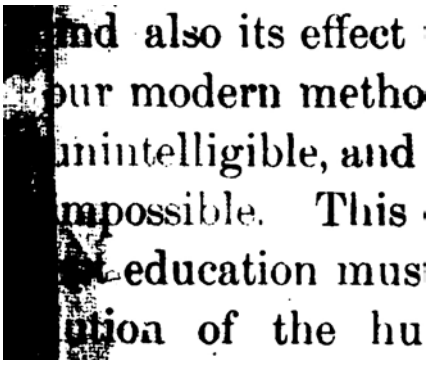
Troubleshoot OCR Error 1: Acrobat Crashes

Occasionally Acrobat crashes during the OCR process. The error message does not indicate what causes the crash and where the problem occurs. Fortunately, the page number of the error can be found on the top shortcuts menu. In figure 3, we can see the error occurs on page 7.

We discovered that errors are often caused by figures or diagrams. For a problem like this, the solution is to skip the error-causing page when running the OCR process. Our initial research was performed on Acrobat 8 Professional. Our recent study shows that this problem has been significantly improved in Acrobat 9 Professional.



Aged Newspaper Scanned in Color



Aged Newspaper Scanned in Black-and-White

Figure 5. An aged newspaper scanned in color and black-and-white

Troubleshoot OCR Error 2: Could Not Perform Recognition (OCR)

Sometimes Acrobat generates an "Outside of the Allowed Specifications" error when processing OCR. This error is normally caused by color images scanned at 600 dpi or more.

In the example in figure 4, the page only contains text but was scanned in color at 600 dpi. When we scanned this page as black-and-white at 400 dpi, we did not encounter this problem. We could also use a lower-resolution color scan to avoid this error. Our experiments also show that images scanned in black-and-white work best for the OCR process.

In this article we mainly discuss running the OCR process on modern textual documents. Black-and-white scans do not work well for historical textual documents or aged newspapers. These documents may have faded text and background noise. When they are scanned as black-and-white, broken letters may occur, and some text might become unreadable. For this reason they should be scanned in color or grayscale. In figure 5, images scanned in color might not produce accurate OCR results, but at least users can read all text, while the black-and-white scan contains unreadable words.

Troubleshoot OCR Error 3: Cannot OCR Image Based Text

The search of a digitized PDF is actually performed on its invisible text layer. The automated OCR process inevitably produces some incorrectly recognized words. For example, Acrobat cannot recognize the Colorado State University Logo correctly (see figure 6).

Unfortunately, Acrobat does not provide a function to edit a PDF file's invisible text layer. To manually edit or add OCR'd text, Adobe Acrobat Capture 3.0 (see figure 7) must be purchased. However, our tests show that Capture 3.0 has many drawbacks. This software is complicated and produces it's own errors. Sometimes it consolidates words; other times it breaks them up. In addition, it is time-consuming to add or modify invisible text layers using Acrobat Capture 3.0.

At CSUL, we manually add searchable text for title and abstract pages only if they cannot be OCR'd by Acrobat correctly. The example in

Original Logo

~dO
UniversitY

Text OCRed by Acrobat

Figure 6. Incorrectly recognized text sample

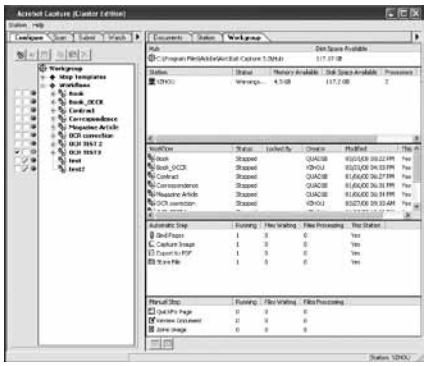


Figure 7. Adobe Acrobat capture interface

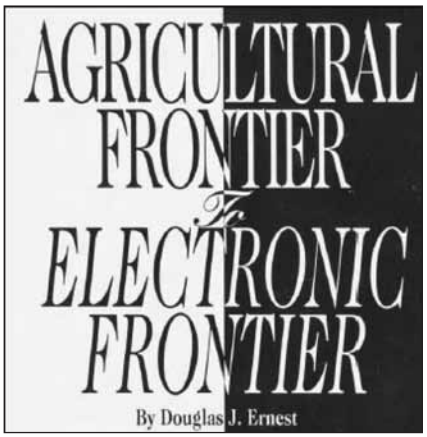


Figure 8. Image-based text sample

figure 8 is a book title page for which we used Acrobat Capture 3.0 to manually add searchable text. The entire book may be accessed at http://hdl.handle.net/10217/1553.

Optimizing PDFs for Web Delivery

A digitized PDF file with 400 color pages may be as large as 200 to 400 MB. Most of the time, optimizing processes may reduce files this large without a noticeable difference in quality. In some cases, quality may be improved. We will discuss three optimization methods we use.

Method 1: Using an Appropriate Color Mode and Resolution

As we have discussed in previous sections, we can greatly reduce a PDF's size by using an appropriate color mode and resolution. Figure 9 shows two different versions of a digitized document. The source document has a color cover and 111 bitonal pages. The original PDF, shown in figure 9 on the left, was created by another university department. It was not scanned according to standards and procedures adopted by CSUL. It was scanned in color at 300 dpi and has a file size of 66,265 KB. We exported the original PDF as TIFF images, batch-converted color TIFF images to black-and-white TIFF images, and then created a new PDF using black-and-white TIFF images. The updated PDF has a file size of 8,842 KB. The image on the right is much cleaner and has better print quality. The file on the left has unwanted marks and a very light yellow background. The undesirable marks and background contribute to its large file size and create ink waste when printed.

Method 2: Running Acrobat's Built-In Optimization Processes

Acrobat provides three built-in processes to reduce file size. By default, Acrobat use JPEG compression for color and grayscale images and CCITT Group 4 compression for bitonal images.

Optimize Scanned PDF

Open a scanned PDF and select Documents > Optimize Scanned PDF. A number of settings, such as image quality and background removal, can be specified in the Optimize Scanned PDF dialog box. Our experiments show this process can noticably degrade images and sometimes even increase file size. Therefore we do not use this option.

Reduce File Size

Open a scanned PDF and select Documents > Reduce File Size. The Reduce File Size command resamples and recompresses images, removes embedded Base-14 fonts, and subset-embeds fonts that were left embedded. It also compresses document structure and cleans up elements such as invalid bookmarks. If the file size is already as small as possible, this command has no effect.[11] After process, some files cannot be saved as PDF/A, as we discussed in a previous section. We also noticed that different versions of Acrobat can create files of different file sizes even if the same settings were used.

PDF Optimizer

Open a scanned PDF and select Advanced > PDF Optimizer. Many settings can be specified in the PDF Optimizer dialog box. For example, we can downsample images from

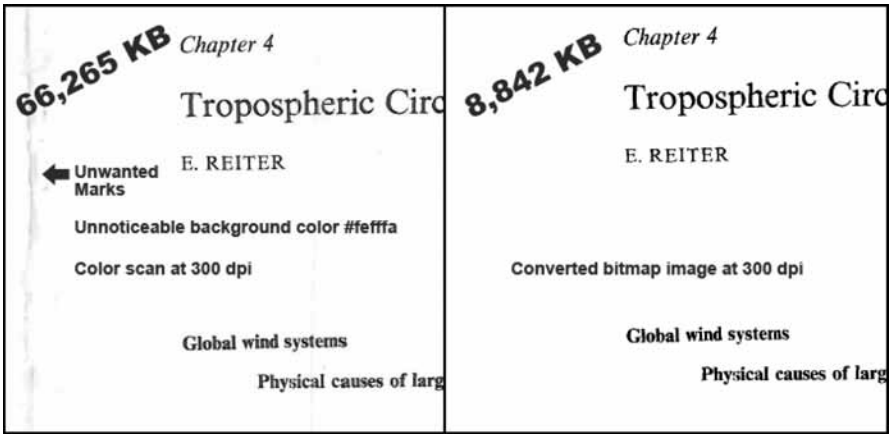


Figure 9. Reduce file size example

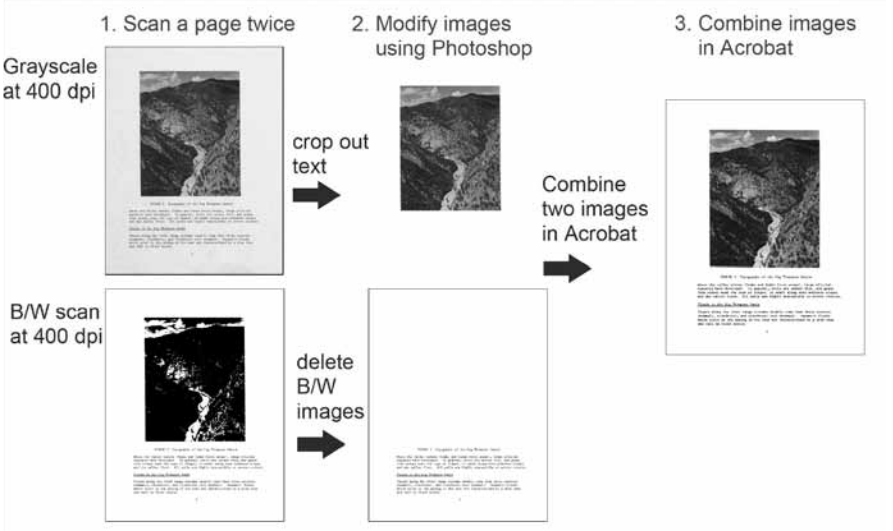


Figure 10. Reduce file size example: combine images

a higher resolution to a lower resolution and choose a different file compression. Different collections have different original sources, therefore different settings should be applied. We normally do several tests for each collection and choose the one that works best for it. We also make our PDFs compatible with Acrobat 6 to allow users with older versions of software to view our documents. A detailed tutorial of how to use the PDF Optimizer can be found at http://www.acrobatusers.com/tutorials/understanding-acrobats-optimizer.

Method 3: Combining Different Scans

Many documents have color covers and color or grayscale illustrations, but the majority of pages are text-only. It is not necessary to scan all pages of such documents in color or grayscale. A PDF may contain pages that were scanned with different color modes and resolutions. A PDF may also have pages of mixed resolutions. One page may contain both bitonal images and color or grayscale images, but they must be of the same resolution.

The following strategies were adopted by CSUL:

1. Combine bitmap, grayscale, and color images. We use grayscale images for pages that contain grayscale graphs, such as black-and-white photos, color images for pages that contain color images, and bitmap images for text-only or text and line art pages.
2. If a page contains high-definition color or grayscale images, scan that page in a higher resolution and scan other pages at 400 dpi.
3. If a page contains a very small font and the OCR process does not work well, scan it at a higher resolution and the rest of document at 400 dpi.
4. If a page has both text, color, or grayscale graphs, we scan it twice. Then we modify images using Adobe Photoshop and combine two images in Acrobat.

In figure 10, the grayscale image has a gray background and a true reproduction of the original photograph. The black-and-white scan has a white background and clean text, but details of the photograph are lost. The PDF converted from the grayscale image is 491 KB and has nine OCR errors. The PDF converted from the black-and-white image is 61KB and has no OCR errors. The PDF converted from a combination of the grayscale and black-and-white images is 283 KB and has no OCR errors.

The following are the steps used to create a PDF in figure 10 using Acrobat:

1. Scan a page twice—grayscale

and black-and-white.
2. Crop out text on the grayscale scan using Photoshop.
3. Delete the illustration on the black-and-white image using Photoshop.
4. Create a PDF using the black-and-white image.
5. Run the OCR process and save the file.
6. Insert the color graph. Select Tools > Advanced Editing > TouchUp Object Tool. Right-click on the page and select Place Image. Locate the color graph in the Open dialog, then click Open and move the color graph to its correct location.
7. Save the file and run the Reduce File Size or PDF Optimizer procedure.
8. Save the file again.

This method produces the smallest file size with the best quality, but it is very time-consuming. At CSUL we used this method for some important documents, such as one of our institutional repository's showcase items, *Agricultural Frontier to Electronic Frontier*. The book has 220 pages, including a color cover, 76 pages with text and photographs, and 143 text-only pages. We used a color image for the cover page and 143 black-and-white images for the 143 text-only pages. We scanned the other 76 pages as grayscale and black-and-white. Then we used the procedure described above to combine text pages and photographs. The final PDF has clear text and correctly reproduced photographs. The example can be found at http://hdl.handle.net/10217/1553.

Conclusion

Our case study, as reported in this article, demonstrates the importance of investing the time and effort to apply the appropriate standards and techniques for scanning and optimizing digitized documents. If proper techniques are used, the final result will be Web-friendly resources that are easy to download, view, search, and print. Users will be left with a positive impression of the library and feel encouraged to use its materials and services again in the future.

References

1. BCR's CDP Digital Imaging Best Practices Working Group, "BCR's CDP Digital Imaging Best Practices Version 2.0," June 2008, http://www.bcr.org/dps/cdp/best/digital-imaging-bp.pdf (accessed Mar. 3, 2010).

2. Adobe, "About File Formats and Compression," 2010, http://livedocs.adobe.com/en_US/Photoshop/10.0/help.html?content=WSfd1234e1c4b69f30ea53e41001031ab64-7757.html (accessed Mar. 3, 2010).

3. Ted Padova *Adobe Acrobat 7 PDF Bible*, 1st ed. (Indianapolis: Wiley, 2005).

4. Olaf Drümmer, Alexandra Oettler, and Dietrich von Seggern, *PDF/A in a Nutshell—Long Term Archiving with PDF*, (Berlin: Association for Digital Document Standards, 2007).

5. PDF/A Competence Center, "PDF/A: An ISO Standard—Future Development of PDF/A," http://www.pdfa.org/doku.php?id=pdfa:en (accessed July 20, 2010).

6. PDF/A Competence Center, "PDF/A—A new Standard for Long-Term Archiving," http://www.pdfa.org/doku.php?id=pdfa:en:pdfa_whitepaper (accessed July 20, 2010).

7. Adobe, "Creating Accessible PDF Documents with Adobe Acrobat 7.0: A Guide for Publishing PDF Documents for Use by People with Disabilities," 2005, http://www.adobe.com/enterprise/accessibility/pdfs/acro7_pg_ue.pdf (accessed Mar. 8, 2010).

8. Adobe, "Recognize Text in Scanned Documents," 2010, http://help.adobe.com/en_US/Acrobat/9.0/Standard/WS2A3DD1FA-CFA5-4cf6-B993-159299574AB8.w.html (accessed Mar. 8, 2010).

9. Ibid.

10. Ibid.

11. Adobe,"ReduceFileSizebySaving," 2010, http://help.adobe.com/en_US/Acrobat/9.0/Standard/WS65C0A053-BC7C-49a2-88F1-B1BCD2524B68.w.html (accessed Mar. 3, 2010).

Appendix A. Step-by-Step Creating a Full-Text Searchable PDF

In this tutorial, we will show you how to create a full-text searchable PDF using Adobe Acrobat 9 Professional.

Creating a PDF from a Scanner

Adobe Acrobat Professional can create a PDF directly from a scanner. Acrobat 9 provides five options: Black and White Document, Grayscale Document, Color Document, Color Image, and Custom Scan. The custom scan option allows you to scan, run the OCR procedure, add metadata, combine multiple pages into one PDF, and also make it PDF/A compliant. To create a PDF from a scanner, go to File > Create PDF > From Scanner > Custom Scan. See figure 1.

At CSUL, we do not directly create PDFs from scanners because our tests show that it can produce fuzzy text and it is not time efficient. Both scanning and running the OCR process can be very time consuming. If an error occurs during these processes, we would have to start over again. We normally scan images on scanning stations by student employees

or outsource them to vendors. Then library staff will perform quality control and create PDFs on seperate machines. In this way, we can work on multiple documents at the same time and ensure that we provide high-quality PDFs.

Creating a PDF from Scanned Images

1. From the task bar select Combine > Merge Files into a single PDF > From Multiple Files. See figure 2.
2. In the Combine Files dialog, make sure the Single PDF radio button is selected. From the Add Files dropdown menu select Add Files. See figure 3.
3. In the Add Files dialog, locate images and select multiple images by holding shift key, and then click Add Files button.
4. By default, Acrobat sorts files by file names. Use Move Up and Move Down buttons to change image orders and use the Remove button to delete images. Choose a target file size. The smallest icon will produce a file with a smaller file size but a lower image quality PDF, and the largest icon will produce a high image quality PDF but with a very large file size. We normally use the default file size setting, which is the middle icon.
5. Save the file.

At this point, the PDF is not full-text searchable.

Making a Full-Text Searchable PDF

A PDF document created from a scanned piece of paper is inherently inaccessible because the content of the document is an image, not searchable text. Assistive technology cannot read or extract the words, users cannot select or edit the text, and one cannot manipulate the PDF document for accessibility. Once optical character recognition (OCR) is properly applied to the scanned files, however, the image becomes searchable text with selectable graphics, and one may apply other accessibility features to the document.

Adobe Acrobat Professional provides three OCR options, Searchable Image (Exact), Searchable Image, and Clean Scan. Because Searchable Image (Exact) is the only option that keeps the original look, we only use this option.

To run an OCR procedure using Acrobat 9 Professional:

1. Open a digitized PDF.
2. Select Document > OCR text recognition > Recognize text using OCR.
3. In the Recognize Text dialog, specify pages to be OCRed.
4. In the Recognize Text dialog, click the Edit button in the Settings section to choose OCR language and PDF Output Style. We recommend the Searchable Image (Exact) option. Click OK. The setting will be remembered by the program and will be used until a new setting is chosen.

Sometimes a PDF's file size increases greatly after an OCR process. If this happens, use the PDF optimizer to reduce its file size.

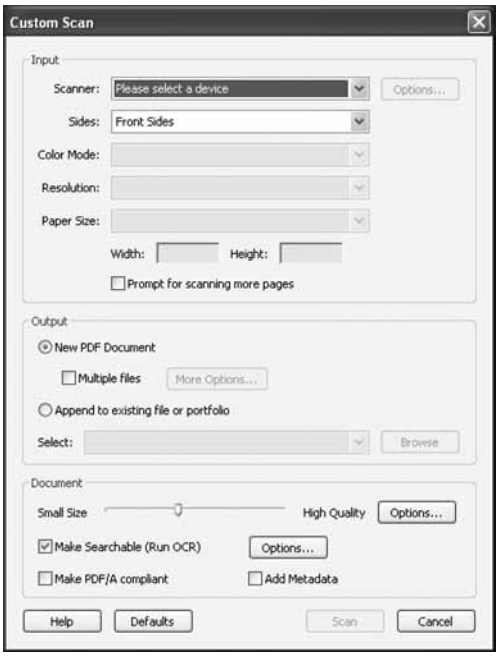


Figure 1. Acrobat 9 Professional's Create PDF from Scanner Dialog

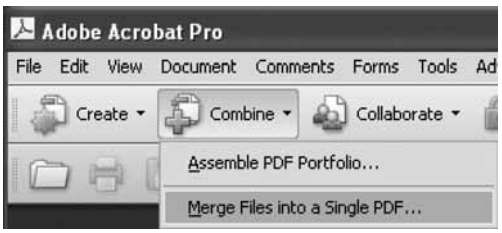


Figure 2. Merge files into a single PDF

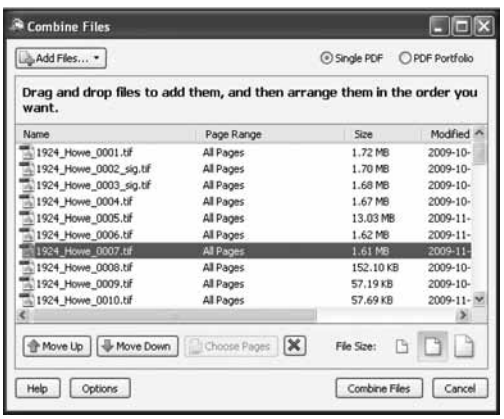


Figure 3. Combine Files dialog