The Art of Scanning

Paul Royster
University of Nebraska-Lincoln, proyster@unl.edu

The Art of Scanning

Paul Royster

University of Nebraska–Lincoln

January 5, 2011

Yes, it is presumptuous to call scanning an "art," when it is really more of a craft, but "The Craft of Scanning" doesn't sound as sexy, so we will consider it for the time being as one of the fine arts, like music, or painting, or dance.

This short treatise derives from work done in the process of scanning published and original materials to create PDF files for online publication or deposit in our institutional repository.

This approach assumes you have a scanner and software to drive it, and also three software programs from Adobe (sold together as their Creative Suite): Adobe Acrobat, Photoshop, and InDesign.

Contents

1. The Scanner

Scanning requires a scanner. No surprise there, and an adequate scanner is within the reach of most aspirants to the art. Scanners now often come as part of a multi-purpose copy-scan-fax machine, and these are adequate for most uses, and priced under $100. There are many models of scanners on the market; almost all will do photographs, color, line artwork, and text, with resolution up to 600 dpi, which will meet most needs. Two additional features that are welcome are "descreening" and "optical character recognition."

Descreening (or re-screening) is used for photographs that have been printed as halftones. Halftone printing of photographs uses ink on paper to represent the continuous range of grays (or colors) in a photographic print. Since ink is only one color, the darkness is modified by varying the size of the dot that is printed: a 100% dot produces a solid ink color (usually black); a 10% dot produces the effect of a 10% gray (or color) by reducing the dot to 10% of its full size. These are now produced by computer programs, but before computers they were produced by laying a white gauze net or "screen" over the photograph, thus dividing it into many small squares. The fineness of the screen is given in "lines per inch"—thus a (standard) 133 line screen had 133 lines per inch (each way).

If you look at a printed photograph (as opposed to a photographic print, which is produced by a chemical process), you will see it is composed of tiny dots in regular rows and columns. Note that newspapers will use fewer and larger dots than fine-printed photography or art books. When a scanner scans these dots it tends to hit each dot at the same relative point, and this will create a regular checkerboard pattern, called moire (pronounced "more-ray"), that is an unsightly distortion. Descreening is a computer subroutine in the scanner that compensates for this and avoids the appearance of moire. Descreening works best when it is set for the same value as the original screening. For most printed materials, the standard is 133 line screen; for fine art and photography books, it will be higher, usually 150 or 200 lines per inch. If your scanner does not have a "descreen" function (or if it was not used on the file you have to work with), there are filters in Photoshop that will also eliminate the moire patterns; see "Adjusting Scanned Images," below.

Optical character recognition, often abbreviated OCR, is software that will recognize and read scanned letters and convert them to live text. Thus instead of a picture of an "A" the resultant file includes an actual "A" that can be manipulated as text in a word processor. Most scanners do include this function, though sometimes it needs to be installed separately from the CD or software that is included. OCR can also be run "after the fact" in Adobe Acrobat, but this is sometimes not as accurate as OCR done as part of the original scan. Note that when you scan a page of text using OCR, what you will get will be a string of words and letters and numbers, not a 2-dimensional copy of the original scanned page.

There are far too many models of scanning to attempt to describe them all. Probably the most popular line is manufactured and sold by Hewlitt-Packard. These are reasonably good scanners, but the HP software that operates them will often drive you to distraction with its many frustrations. Good scanners, bad software.

One other feature that is nice to have on occasion is a **transparency adapter**—which allow you to scan slides, transparencies, negative film, and other see-through media that needs to be lit from the back to show the image.

I mainly use a Microtek Scanmaker i700 desktop scanner that retailed for under $300. I have no reservations about recommending that model, with the proviso that (as reviewers have pointed out) it is a veritable "swamp turtle" when it comes to speed.

Figure 1. Top: Example of moire patterns. Printed photograph scanned at 200 dpi without descreening. The diagonal and checkerboard patterns are especially noticeable in the gray background. Bottom: Same printed photo, scanned at 200 dpi, with descreening. Note the absence of diagonals and checkerboarding, and also the smoother transition from lights to darks. (Photo of Father Pierre-Jean De Smet, by Matthew Brady [detail]).

2. Scanner Software

The software that operates the scanner varies for almost every scanner model on the market, so this section will be relatively non-specific. The software should allow you to set (and not continually need to re-set):

1. original media
2. scan area
3. type of scan
4. resolution
5. threshold
6. output file type
7. output file destination and naming

Original media — Is the object to be scanned opaque (such as a printed book or sheet of paper) or transparent (like a slide)? Is it positive (normal) or negative (reversed)?

Scan area — Most scanners will allow you to do a pre-scan and then select the area you wish to have scanned. (Some "bad" scanner software will do this automatically and then perversely "select" for you.) There is no need to scan unnecesary areas of the scanner bed, but sometimes, as when scanning multiple pages from a book, you may wish to set an area larger than needed for a single page to avoid re-setting the scan area for each page.

Type of scan — The 3 choices are grayscale, color, and line art (also called bitmap or black & white). **Grayscale** scans are used for black & white photographs; they produce rows and columns of pixels in 256 levels of gray, from 0 = white to 255 = black. **Color scans** do the same, but in 3 or 4 layers. **RGB color** (for monitors, screens, and online presentation) does 3 layers: one each for red, green, & blue. **CMYK color** (for color process printing) does 4 layers: one each for cyan, magenta, yellow, and black. Logically, RGB color files will be 3 times as large as grayscale files (for the same image size and resolution), and CMYK color files will be 4 times as large.

    Line art (also called **bitmap** or **black & white**) scans recognize only 2 colors—black or white. These are appropriate for text and for artwork that has only black lines on white background. There is no gray. Line art files are much smaller than grayscale or color files, since they only have 2 levels, not 256 (or 2^8) for each dot. However, line art scans cannot be compressed, while color and grayscale scans can be.

Resolution — This is how many dots-per-inch vertically as well as horizontally will be recorded and reproduced in the file. Recommended resolutions for various scan types are:

- grayscale   300 dpi
- color (RGB & CMYK)   300 dpi
- line art/text   300 to 600 dpi
- grayscale or color artwork
    with text   400 dpi

Resolution is directly related to file size, so overscanning can lead to unmanageably large files that are hard to manipulate and take too long to download. Rough typewriter text is adequately reproduced at 300 dpi; smaller type does better at 600 dpi. Sometimes charts and graphs require grayscale or color scanning but also have (usually small) type on them; for these 400 dpi is a good compromise.

Threshold — This applies only to line art scans, and it is the point of darkness at which white changes over to black. Normally, 50% is the cut-off, but sometimes the original requires the threshold be adjusted between 30% and 75% to get an appropriate image. If your letters are broken and the thin parts are becoming invisible, lower the threshold. If the letters are too thick and every dirt spot is reproducing or the gutter shadow is creeping into the text, then raise the threshold. Works printed on coated paper tend to have finer and lighter type and may need to have the threshold adjusted.

Output file type — What kind of file shall the scanner save the scan as? The choices are text (for OCR scans), TIFF, and JPG. TIFF is required for line art. JPG is appropriate for grayscale and color scans. JPG is a compression formula that reduces file size with some cost in image quality, depending on the JPG setting: "maximum" quality hardly reduces the file size; "high" quality reduces the file size by half with almost no perceptible decline in image resolution; "medium" and "low" quality are not recommended.

    Some scanners may offer PDF as an output type. You can try it and see what you get, but oftentimes this produces an OCR'ed file with the type all converted to Times Roman (regardless of the original typeface) and any word or letter not recognized is reproduced as a picture, so the resulting page looks somewhat like a ransom note, and is a mixture of text and images.
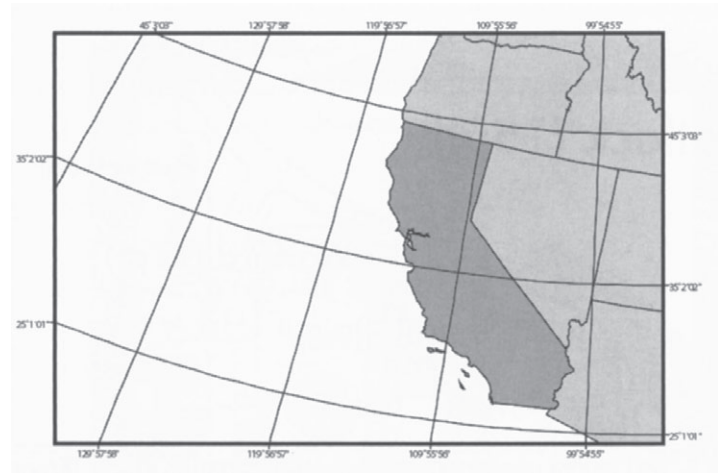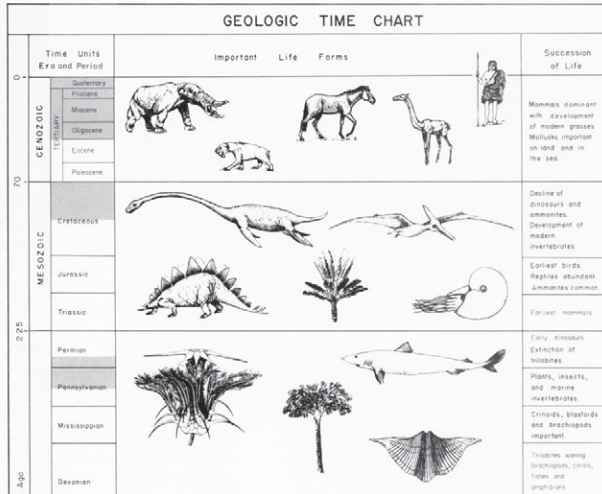
Figure 2. Examples of grayscale scans; note there are different shades of gray.
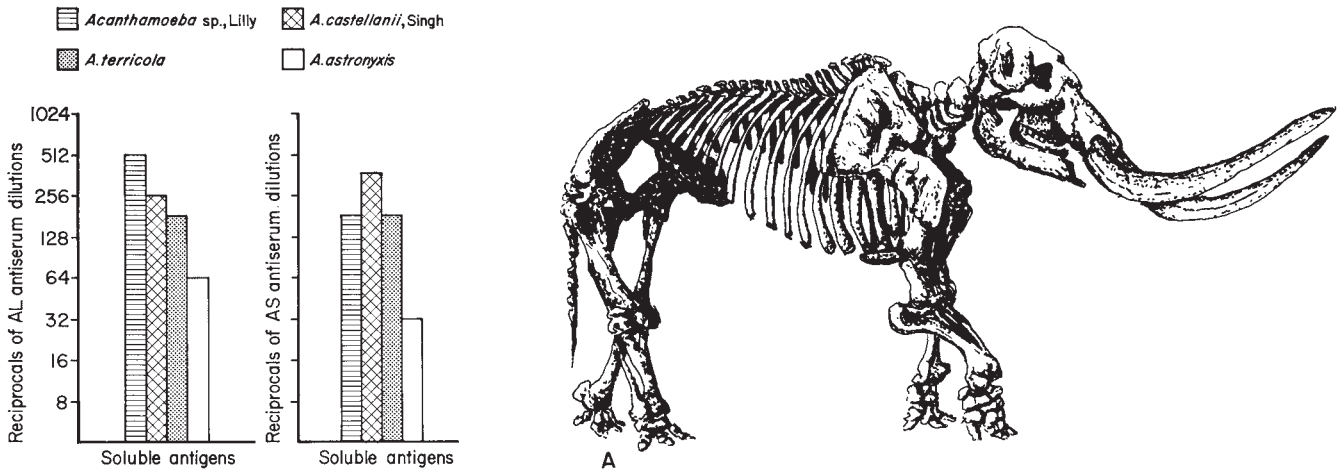

**Figure 3.** Examples of line art (bitmap) scans; note there are only 2 colors—black or white.

Figure 4. Text (scanned as line art/bitmap): Top: 300 dpi; Bottom: 600 dpi (enlarged 300%).

Some OCR programs offer MS Word or RTF word processing format as an output type, and this is tempting in that it may capture italics, bold, and other text characteristics. However, some programs attempt to reproduce the page layout by placing everything in separate text boxes, which are awkward and unwieldy to handle. I have found it best to go with unformatted Unicode text, producing a single text stream, and to manually restore italics, boldface, etc. as necessary.

Output file destination and naming — Where shall the scanner store the file? Some scanner software automatically places the file in the "My pictures" folder or in some other out-of-the-way or hard-to-find location and names it "Myscan00001, " etc. Good scanner software will allow you to choose the location and will let you set up a naming structure so that you can easily find and manipulate the files further as needed.

Some OCR scanning programs allow scanning multiple pages into one file, and this is a nice feature. Be careful, though, not to overload the system—to have it crash on the 38th page of a 40-page document will force you to start all over again.
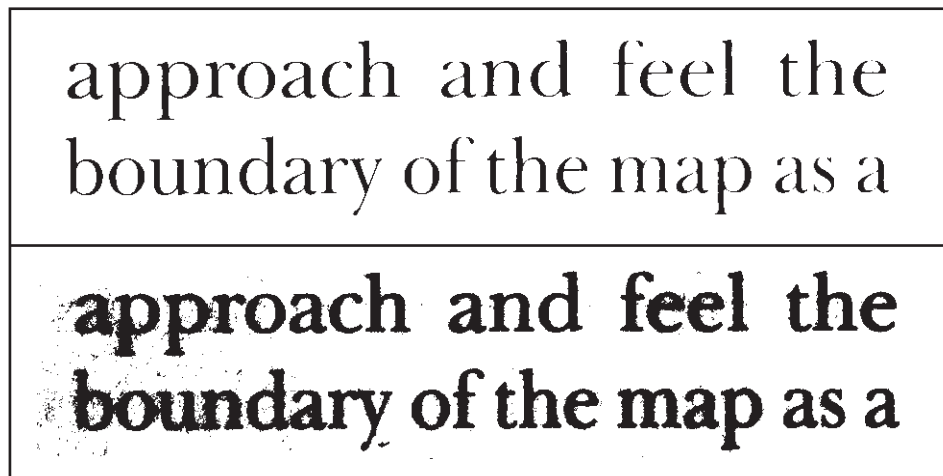
Figure 5. Text (scanned as line art/bitmap) with improper "threshold" settings:: Top: too high; Bottom: too low (enlarged 300%).

3. What is a "Good Scan"?

A good scan is one where:

- The type is black on a white background; i.e. scanned as line art, not as grayscale, which gives dark gray type on a light gray background.

- The text is searchable and can be copied and pasted.

- The artwork is grayscale or RGB color when appropriate, without moire effects.

- The pages are in the correct order.

- The pages are straight and right-side-up.

- The pages are all the same size.

- The final file size is reasonable — usually 100 kilobytes per page or less, so that a 20-page document is about 2.0 Mbytes. Sometimes, if there is a lot of art, this is not possible. But, in general, file sizes over 100 Mb should be avoided if possible.

The Lewis-Syford House, located at 700 North 16th Street on the University of Nebraska-Lincoln campus, was built about 1878, for the Reverend Elisha Moses Lewis, early Presbyterian pastor and missionary. The house is Lincoln's only example (and one of Nebraska's last remaining examples) of the French Second Empire style of architecture, popular in America beginning about 1865, when Lincoln's earliest homes were being built.[1] It is also valuable as a surviving home of a middle income Nebraska family of the 1880s.

The Lewis-Syford House, located at 700 North 16th Street on the University of Nebraska-Lincoln campus, was built about 1878, for the Reverend Elisha Moses Lewis, early Presbyterian pastor and missionary. The house is Lincoln's only example (and one of Nebraska's last remaining examples) of the French Second Empire style of architecture, popular in America beginning about 1865, when Lincoln's earliest homes were being built.[1] It is also valuable as a surviving home of a middle income Nebraska family of the 1880s.

Figure 6. Top: text scanned as grayscale at 600 dpi (shown at 100%). Bottom: as line art.

Figure 7. Printed (halftone) photograph scanned as line art/bitmap (top) and as grayscale (bottom) (scanned at 300 dpi; shown at 100%).

4. Simple Scanning: Text + Line Art

The most straightforward type of scanning to create a PDF file involves documents with only text and/or line art. These can be scanned together, one scan per page is all that is required. We scan these as line art, crop them in Adobe Photoshop, assemble them in Adobe Acrobat, run OCR, and standardize the page sizes.

This method is appropriate for a book chapter, journal article, or entire book that has no photos, color, or grayscale artwork—just text and charts or graphs with only black lines or areas.

With a little practice, and depending on the speed of the scanner, my work-study students have achieved output rates approaching 100 pages per hour. More realistically, I think an estimate of half of that speed (say 50 pages/hour) would be appropriate for most budgetary or planning purposes.

Phase 1: Scanning

- Place the book on the scanner bed and use the top edge and fore-edge to align the book page as straight as possible.

- Do an "overview" or "preview" scan.

- Select a scan area that is the full height of the book and the full width of the scanner bed (even though this may give you part of the previous or following page; this is so you do not have to reset the scan area for each page).

- Select the settings for "line art" (or "bitmap" or "black & white").

- Set the resolution for 600 dpi; if there are more than 40 or 50 pages, you may wish to reduce this to 400 dpi, to keep the final file size manageable. 600 × 600 dots makes 360,000 dots per square inch; 400 × 400 dots makes 160,000 dots per square inch, or less than half the size.

- Scan each page and save the files as TIFF. These files will be fairly large but will reduce significantly when converted to PDF.

- You should review the first few pages (by opening the files in Adobe Photoshop) before scanning them all to make sure that the threshold cutoff is correct.
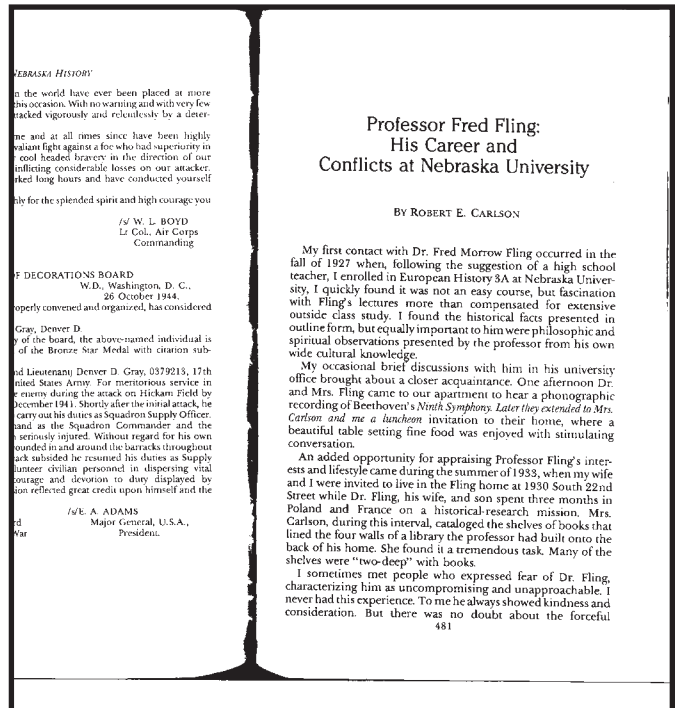
Figure 8. Proper scan area for book scanning. Using the entire width allows you to do both left-hand and right-hand pages without resetting the scan area.

Phase 2: Adobe Photoshop

- Open the TIFF files with Adobe Photoshop

- Crop the image to the type block (see illustration). You can use the selection rectangle and the keyboard shortcut "Alt + OP" to crop (this is faster than using the pull-down menus for Image > Crop, etc., or the "crop" tool). It is not necessary to crop to the exact edge of the type; just get reasonably close and be sure to crop out all shadow, edges, and other unwanted things.

  If a page is short—such as the last page of a chapter that has only a few lines—crop to where the full page would end, not to where the type stops.

  If a page contains part of another (unwanted) article, crop to the normal page size, and then delete (i.e. erase) the unwanted section, leaving the page its full size.
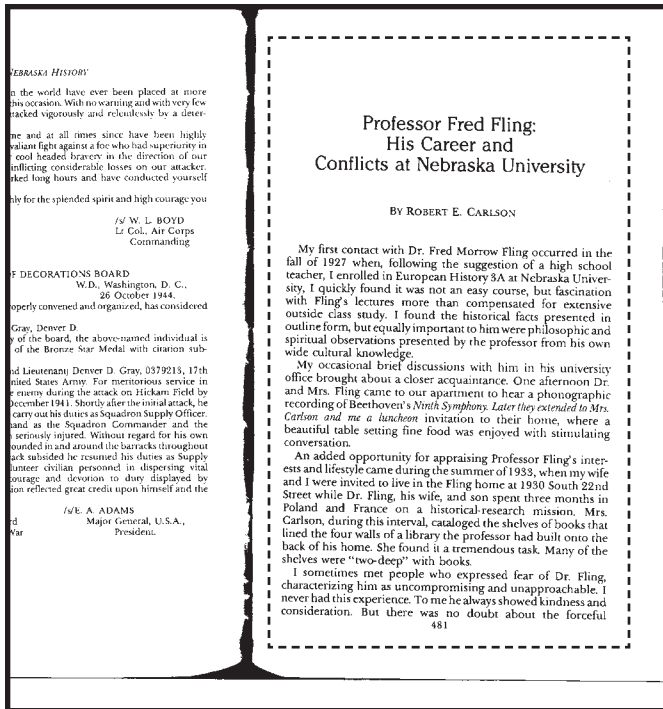
Figure 9. Cropping to the type block. Include areas where type would normally appear—even if it does not appear on that particular page.
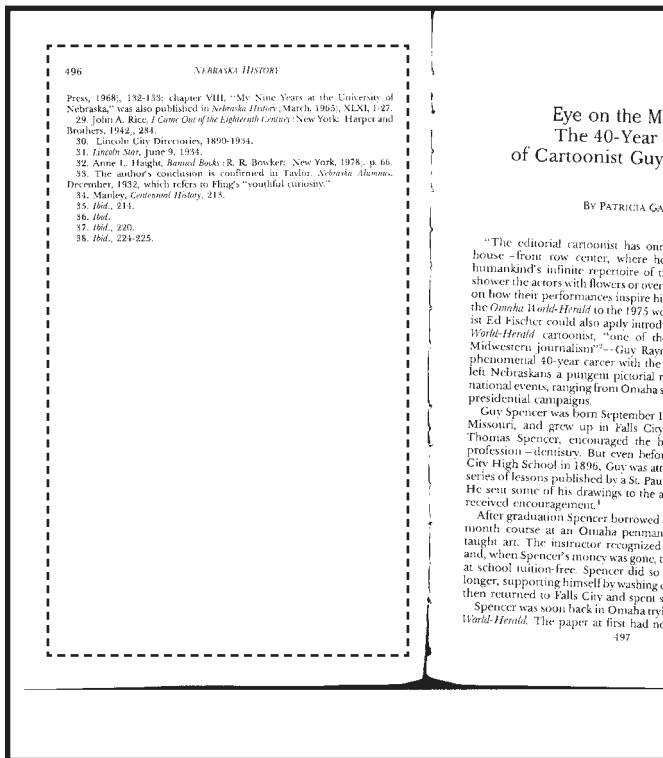

Figure 10. Cropping a short page. Include the full height of the page so it will show properly in the ensuing PDF file.

- Using the "Save As" (keyboard shortcut = "Ctrl+Shift S") function save the file as a Photoshop PDF file. You may notice that the file size will go from several Mb down to less than 100 kb. Interestingly, the size of the PDF file is a function of how much black there is on the page.

Phase 3: Adobe Acrobat

- Use Acrobat to "Combine" the PDF files. This can be done by opening the first one and then using "Document > Insert >Pages" to insert the following pages at the end of the document, one at a time. Acrobat also has a "Combine files" feature that will allow you to drag all the PDF files into a new PDF file at once; this is useful for longer articles. This function is under the "File" menu.*

- Page through to verify that all the pages are there and in the correct order.
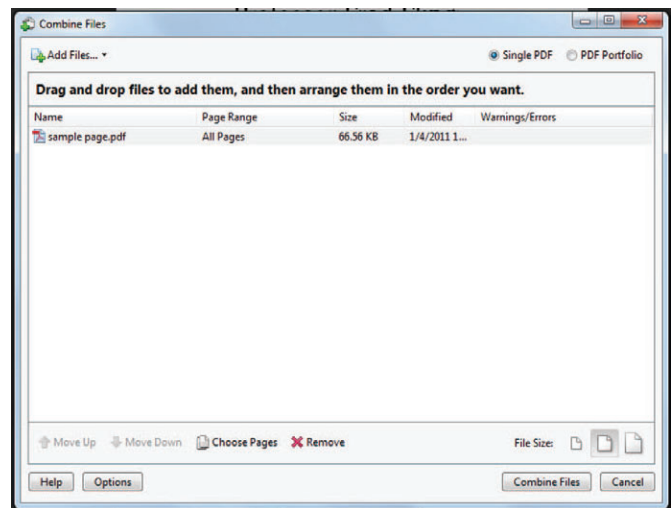
Figure 11. Adobe Acrobat's Combine function is found under the File > Combine > Merge Files into a Single PDF menu.

---

* The newer releases of Adobe Acrobat will convert TIFF files to PDF as part of the "Combine" function. That approach is not recommended here for two reasons: 1) the TIFF files have not been cropped, so they will show the gutter, edges, etc., and 2) PDF files created by the Acrobat "Combine" function are (for reasons unknown to this author) larger—i.e., more bytes—than the equivalent files created by the Photoshop "Save as ... Photoshop PDF" function.

- Run OCR (optical character recognition) on the entire document. This command is normally under the "Document"menu. If the language is not English, "Edit" the settings to select the proper language. Acrobat's OCR program will also straighten any pages that were not completely square before. Acrobat produces a hidden text string that sits behind the image of the scanned page. The typeface remains whatever it was before, and if something is not recognized or mis-recognized, only the hidden text is incorrect, and this is not observable unless the text is copied and pasted into another venue.

- Use the "Crop pages" command to set all the pages to the same size. Choose (or set) a size that is at least 1 inch larger than the cropped pages you created in Photoshop (Acrobat will show you this as the current page size). A 1-inch excess will create a ½-inch margin all around (skimpy); a 1.5-inch excess will produce a ¾-inch margin (better); and a 2-inch excess will give a 1-inch margin (luxurious).

- Odds are that the original document was one of the following sizes: 6 × 9 inches, 6.13 × 9.25 inches, 7 × 10 inches, 8 × 10 inches, or 8.5 × 11 inches ("Letter" size); although if it was a British document, it's harder to say.

- Acrobat places the old page right smack dab in the middle of the new page size (so if you had cropped that short page to be only 4 inches tall, those 4 inches will show in the middle of the new page, not towards the top as they ought to).

- Be sure to select the box that applies this new size to "All pages" not just the "Current page".

- Save the file. You are done. You can delete all those single-page TIFF and PDF files.

Figure 12. Adobe Acrobat's Crop screen is found under the Document > Crop pages menu..

Figure 13. Result of improper cropping of a short type page. Acrobat's Crop (Page Resize) function puts it in the middle, not at the top.
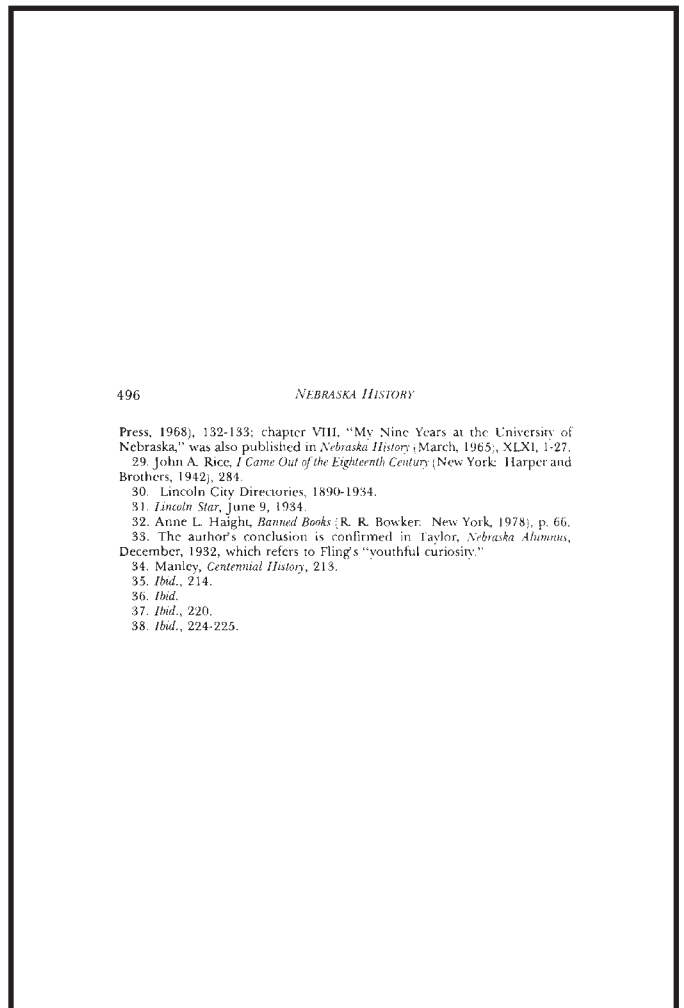
5. Scanning Documents with Grayscale or Color Images

If all documents were just text and line art, life (and the professional literature) would be simpler but less interesting. Making PDF files from documents with grayscale or color illustrations involves some additional steps and the use of an additional software, Adobe InDesign. In this method we scan the pages as line art and delete the illustrations, then scan the illustrations separately as grayscale or color, re-combine the two parts in a page-layout program, and generate PDF pages. Adobe InDesign comes as part of the Creative Suite package (along with Acrobat and Photoshop); Quark Xpress is a similar page-layout program that could be used instead.

Following is a step-by-step guide to this method:

Phase 1: Scanning

- Follow the instructions in #4 Phase 1, above to scan all the pages as line art and save as TIFF files.

- Go back and rescan the grayscale or color images as grayscale or color at 300 dpi. Use the scanner's "descreen" setting to prevent moire patterns. Save these files as JPG format. Don't worry about tight cropping of these images at the scanning stage, just be sure you get the whole image.

Phase 2: Adobe Photoshop

- Open the TIFF files with Adobe Photoshop and crop to the type block.

- If a page had no grayscale or color artwork, then "Save as" a Photoshop PDF.

- If a page has grayscale or color artwork, delete (erase) this artwork, leaving a blank white space where it used to be. Save these pages in their original TIFF format.

- If an illustrated page is noticeably crooked, you should straighten it; recombining the type with the illustration is easier if they are both straight. To straighten one of these pages, first convert it from bitmap/line art to grayscale (Image > Mode > Grayscale; or Alt + I M G), then rotate it clockwise or counterclockwise until a line of type all sits on the same horizontal guideline (which you can pull down from the top ruler bar). Rotation

Figure 63. Howietown Fish Hatchery in 1974. Photo by the author.

which was still unknown that day! Later I gave a two hour slide show talk on fish parasites. I was urged three times to return for a sabbatical of a year or less, but it never materialized.

Stirling University of Scotland has the most beautiful campus I've seen. That opinion was probably influenced by a nice trout pond *on campus* where the pond water is cool enough year round for trout. Interestingly, I was able to understand all the Scots I talked with except one! Surprisingly, some of them had no noticeable Scottish accent.

Saturday morning we visited what may be one of the oldest fish (trout) hatcheries anywhere. The hatching house, resembling a small castle, was built in 1860 and must have been designed by a man with "vision" because it was still very functional even in 1974.

Next I flew to London where the hotel accommodations near Heathrow Airport were convenient and rather inexpensive. I had a "real" breakfast the next morning—bacon and eggs! The return to Little Rock, AR, was uneventful. Upon my return to the Arkansas lab I learned that my first Fish and Wildlife Service assistant, Dr. Robert Putz, was now a G.S. 15 (near the top) and was in charge of all USFWS research, a dramatic rise!

Figure 14. Formerly illustrated page with illustration deleted in Photoshop. Save this page in its original TIFF format. It will later be placed in an InDesign file.

goes by degrees: 1 degree is a lot; less than .1 degree is probably not worth doing. Re-convert the image back to bitmap/line art (Image > Mode > Bitmap; or Alt + I M B) before saving as TIFF.

- Open the illustration JPG files and straighten and crop them as needed.

- Sometimes you will discover that a picture was not straight or not square to start with. When in doubt, straighten it so the top is horizontal.
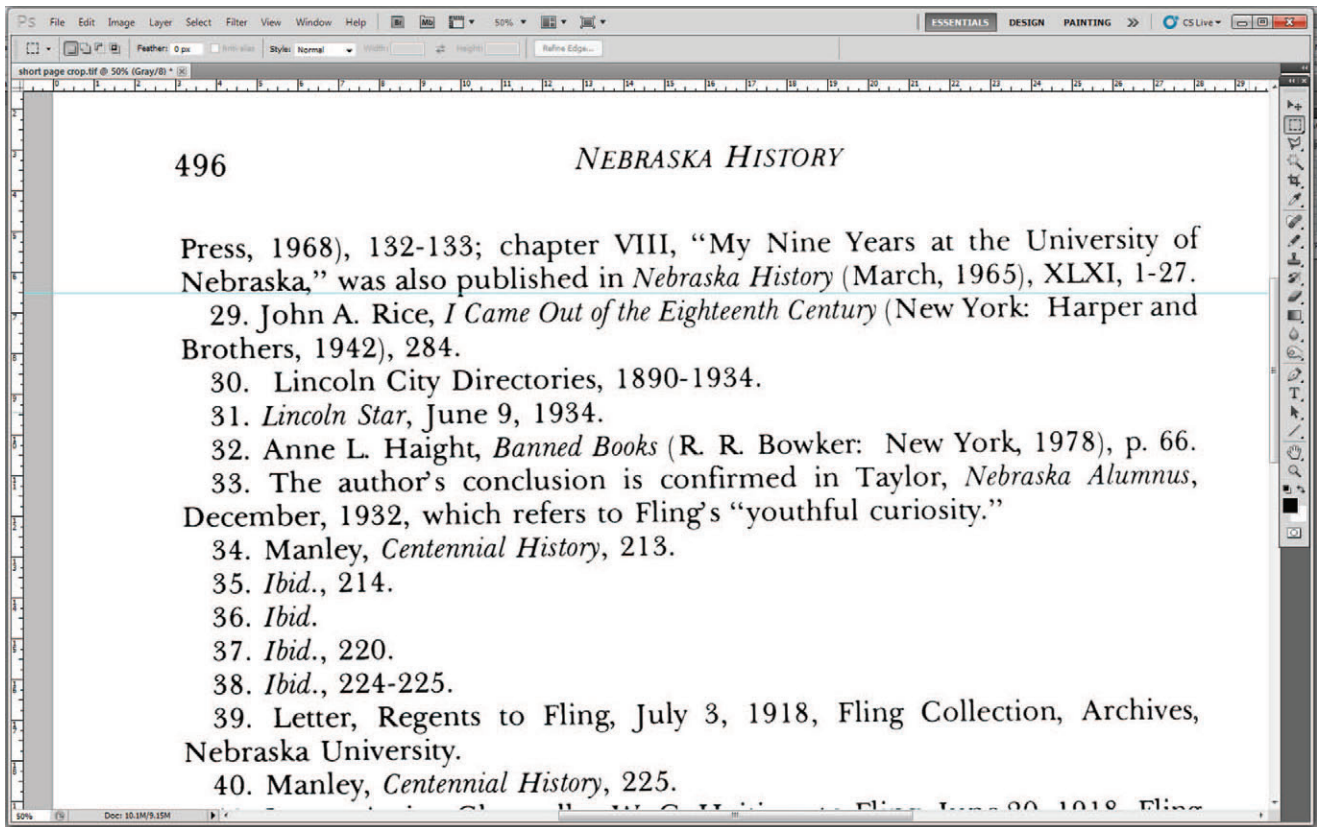
Figure 15. Straightening text pages in Photoshop. The light blue guideline is placed by clicking on the ruler at the top and dragging down to the bottom of a line of letters on the left. The file has been converted from bitmap/line art to grayscale so that Rotation can occur; this is why the "Doc" size (lower left of screen) is so large. This page needs to rotate approximately .5% clockwise to make the type horizontal. Convert back to bitmap/line art before saving.
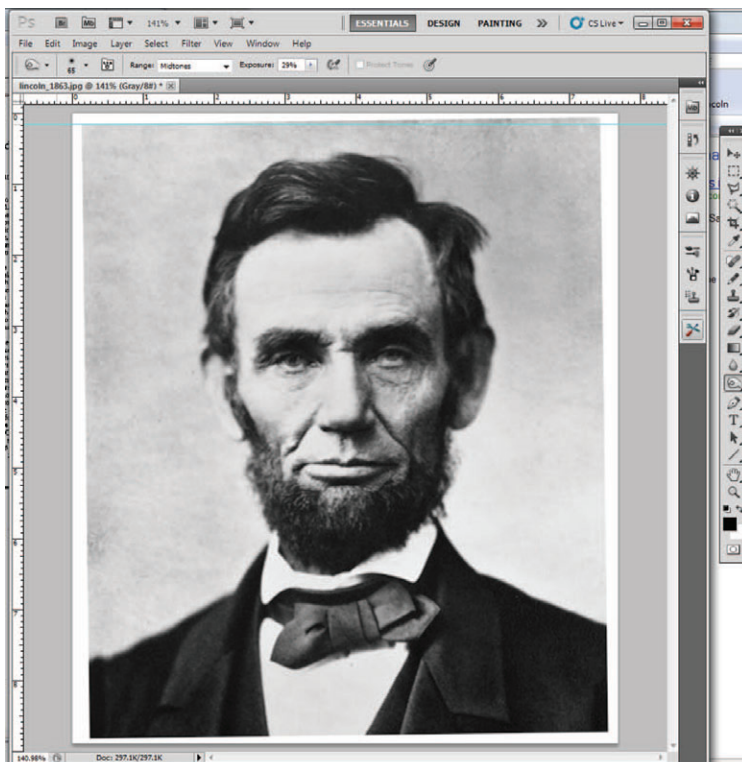

Figure 16. Straightening an image in Photoshop. The light blue guideline is placed by clicking on the ruler at the top and dragging down to the top of the image on the left. This image needs to rotate about .25% clockwise to be straight.

- If moire patterns remain, use the Gaussian blur filter (Filter > Blur > Gaussian blur; range = 1 or 2 pixels) to fix that.

- If adjustments are needed for darkness, contrast, etc., see the section below on "Adjusting Scanned Images" for suggestions and guidance.

- Save the cropped, straightened, and adjusted images in their original JPG file format.

Phase 3: Adobe InDesign

- Start a new InDesign file (New > Document > ...). The page size should be larger than the type block you have cropped to and smaller than the final page size you intend to wind up with. InDesign by default measures space in picas (6 picas = 1 inch); if you wish, this can be changed to inches by the "Edit > Preferences > Units and Measures" command.

- Don't worry about the margins, since you won't be setting any type. The default margins of 3 picas (½ inch) on all sides are fine.

- The number of pages should be the number of pages with illustrations that you wish to combine with the type. If you mis-estimate, more pages can always be added with the "Layout > Pages > Insert pages" command.

- For each page with illustrations, File > Place (Ctrl + D) the TIFF file containing the type on the InDesign page. Then place (again Ctrl + D) the JPG illustration file(s) and move it to the appropriate position

- "Export" the combined pages to an Adobe PDF file. These pages will show the type as black-on-white line art and the illustrations as either grayscale or color. Alternatively, you may also generate the PDF file via the "Print" function or via the "Adobe PDF Presets ..." function (which I prefer).
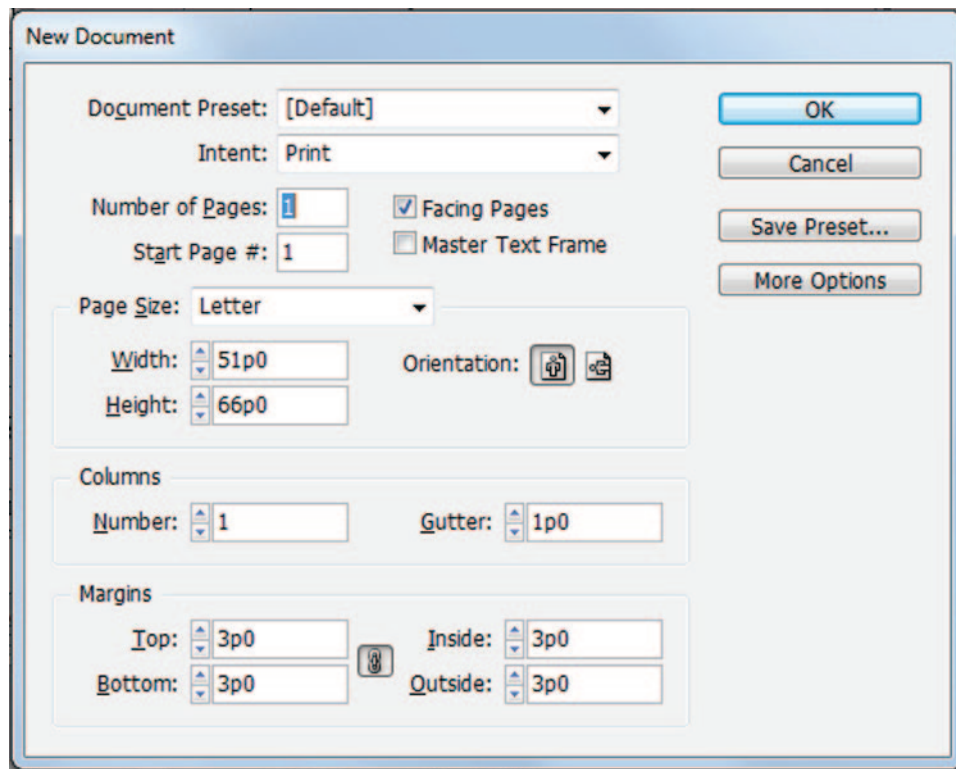
Figure 17. New Document dialogue box for Adobe InDesign. Enter the "Number of Pages," un-select "Facing Pages" and select a page size that is larger than the cropped type pages but smaller than the final intended page size. Don't worry about the margins.
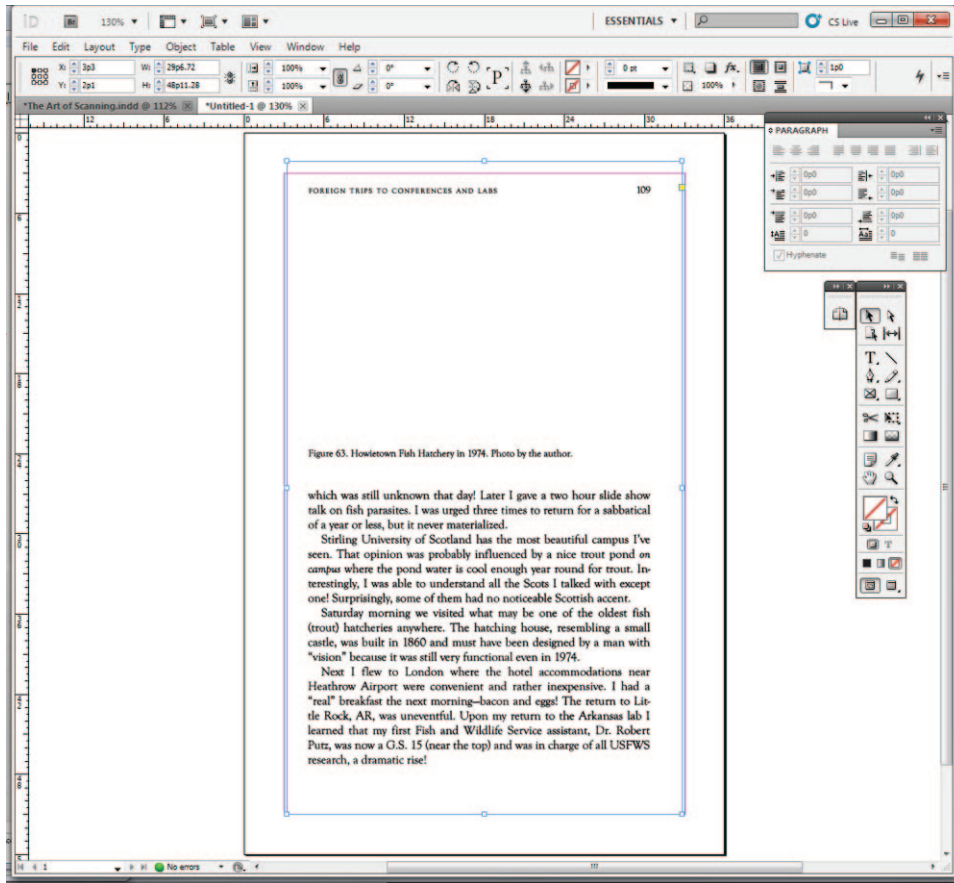
Figure 18. Place the (straightened if necessary) text page TIFF file into the InDesign document (File > Place or Ctrl+D). Use the red margin lines to help situate the type on the page: either as centered or (better) with slightly larger outside than inside and bottom than top margins.
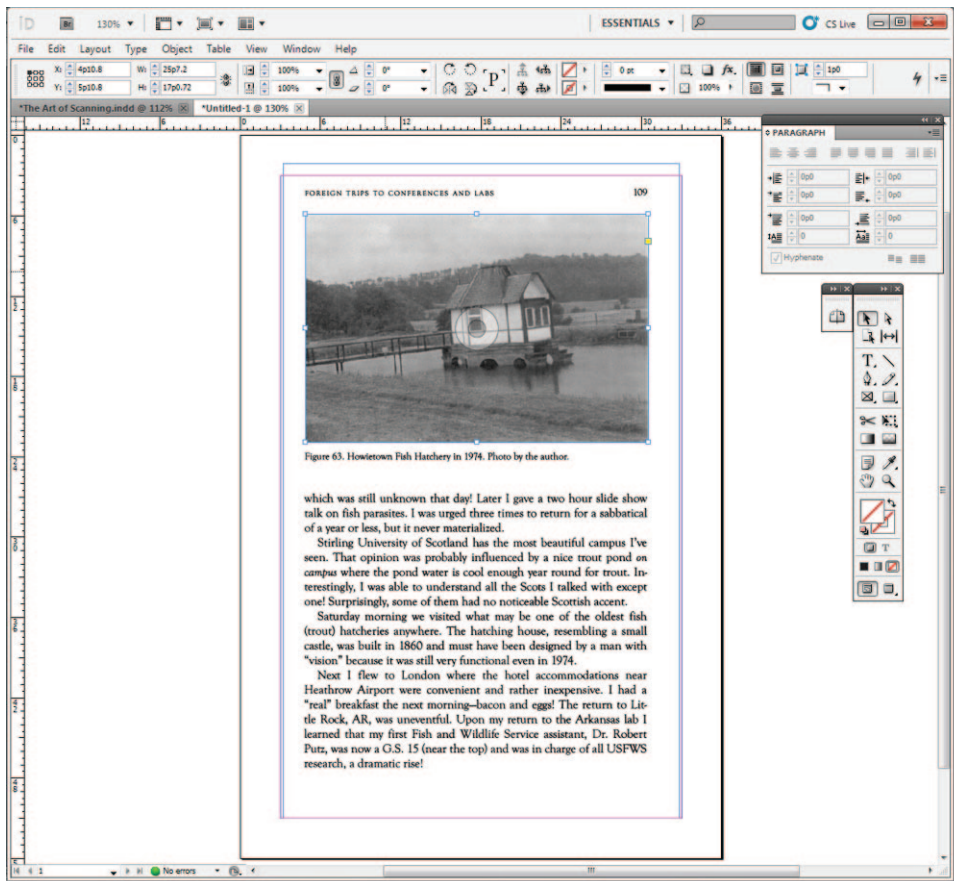
Figure 19. Place the (straightened if necessary) image JPG file into the InDesign document (File > Place or Ctrl+D). Situate the image in the hole you created by deleting it from the line art/bitmap version of the page.

Phase 4: Adobe Acrobat

- Combine the PDF files, either by inserting the single-page type-only files one-by-one into the PDF file with the illustrated pages, or by extracting all the pages in the PDF file with the illustrated pages into single pages (Document > Pages > Extract > as single pages) and then using the "Combine" function to assemble them in proper order.

- Page through to verify that all the pages are there and in the correct order.

- Run OCR (optical character recognition) on the entire document. Make sure to select the proper language. This will also straighten any pages that were not completely square before.

- Use the "Crop pages" command to set all the pages to the same size.

- Save the file. You are done. You can delete all those single-page TIFF and PDF files and the JPG files of the illustrations and the InDesign file.

6. Various Issues You May Encounter

Scanning Issues

The shadow from the gutter creeps into the type block

Some books are bound so tightly and the type is so close to the spine that the volume will not lie flat enough to properly scan the type that is closest to the gutter. There are two possible remedies for this:

1. Photocopy the page and scan the photocopy. Copy machines splash so much light on the page that the gutter shadow is usually reduced or eliminated. In especially problematic cases, be sure the spine of the book is oriented at a right angle to the copier's light source.

2. Scan the page as grayscale (at 600 dpi) and save as TIFF. Open in Photoshop and convert to bitmap/line art (Image > Mode > Bitmap). You may need to adjust the "levels" to reduce the darkness of the gutter shadow first, or use any of Photoshop's image adjustment techniques to push the gutter shadow below the white/black threshold. Whatever you do, be sure that on the conversion to bitmap you select the "50% threshold" option and not one of the "dither" methods (see below "The type has holes in it !").
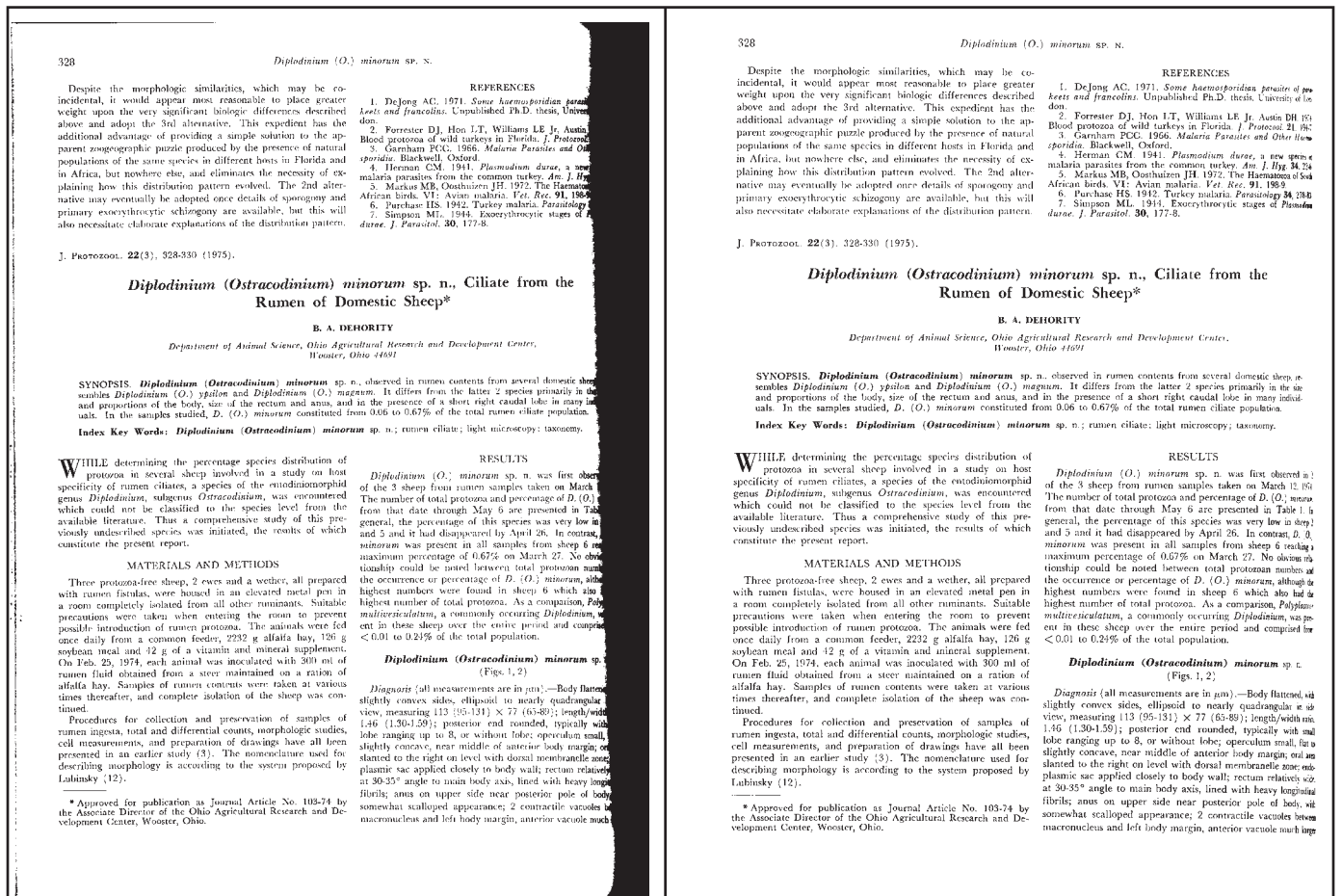
Figure 20. Left: Page scan showing gutter shadow. Right: Same page scanned as grayscale, with levels adjusted, highlights "burned," and shadows "dodged" in Photoshop, then converted to bitmap/line art. See Section 7 for "Adjusting Scanned Images."

The type closest to the gutter bends downward at the top and upward at the bottom.

This is a result of the book not lying flat and the distortion is created by the perspective. This can be cured using Photoshop. Convert the bitmap/line art image to grayscale. Use the selection rectangle to select the type that is so affected (usually the first several letters of all the lines). Go "Edit > Transform > Skew" and boxes will appear at the corners of the selection rectangle. Pull the upper and lower corners to distort the rectangle (into a trapezoid) so that the lines of type run straight. Click on the selection rectangle tool, and accept the transformation. Convert the file back to bitmap (being sure to use the 50% threshold option).
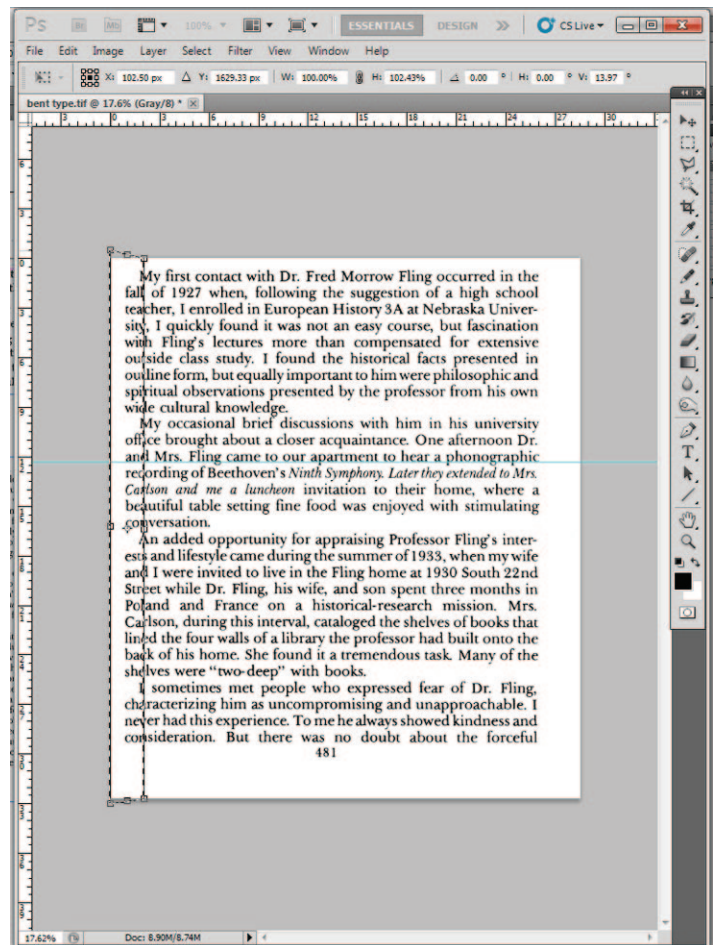
Figure 21. Left: Page scan showing bent type due to gutter distortion. Right: Same page in Photoshop, converted to grayscale, with affected area selected and transformed with "Skew" function. Convert back to bitmap/line art before saving.

The type has holes in it !

The image was converted from grayscale to bitmap/line art using one of the "dither" options instead of the "50% threshold." In the Image > Mode > Bitmap command sequence, there is an "Edit" button at the last stage. Click this and select "50% threshold" instead of one of the "halftone" or "dither" options (which try to fake a grayscale image by removing dots from black to give representation of shades of gray, but actually wind up making your type look as though it was shot with a shotgun). Once you change this setting, it should remain (unless someone changes it or Photoshop gets re-installed).

The image has swirling or checkerboard patterns.

It was a printed halftone that was not "descreened" in scanning and so is showing these patterns called moire ("mo-ray"). You can re-scan it with the "descreen" function turned on, or you can use the Photoshop Gaussian blur (Filter > Blur > Gaussian blur) to eliminate the patterns. Usually a radius setting of 1 or 2 pixels is enough to eliminate the moire without damaging the sharpness of the image.
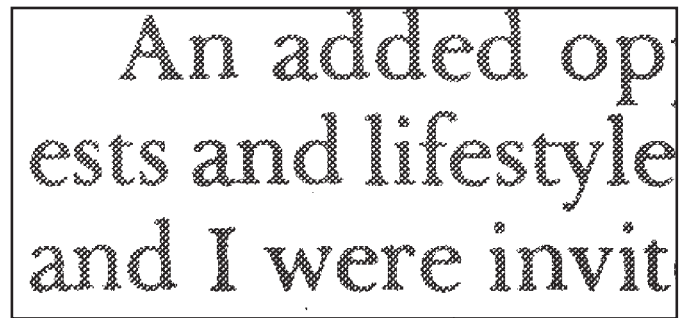
Figure 22. "Shotgun" or "Swiss cheese" type effect produced by converting gray type to bitmap/line art using a "halftone screen" or "dither" method rather than the 50% threshold filter in Photoshop.

The thin parts of the type are disappearing.

You need to lower the threshold.

The type is filling in.

You need to raise the threshold.

The image has horizontal lines in it from the type on the reverse page showing through the paper.

The best way to fix this problem is to re-scan the image with a black sheet of paper behind the page.

Figure 23. Type and figure from the following page are showing through the paper and are visible in the image. The best way to correct this problem is to rescan the image with a black sheet of paper behind it.

Acrobat Issues

Some pages do not resize when I try to make all pages the same size.

> Acrobat will not change pages sizes if the original page is larger than the new target size.

> Sometimes a page was not cropped down enough to fit the new size. Try a slightly larger size.

> Sometimes this happens when the OCR program turns a sideways page right-side-up (according to the type orientation, but not the book orientation; i.e. it was originally a "turn page" in the book). To fix this, go to that page and make only that page square by the larger measure (e.g., make it

9 × 9 if you are shooting for 6 × 9). Then use the cropping functions to chop off from the top and bottom so the page is the right size, though a different orientation.

> If you want to avoid having Acrobat rotate these "turn" pages, then do not OCR those pages. Select the range of pages before and after for OCR.

There are pages out of sequence.

> Go to that page and extract it: Document > Pages > Extract > delete page after extracting. Save the extracted page, then insert it back into the PDF file in the proper place, using Document > Pages > Insert pages ...
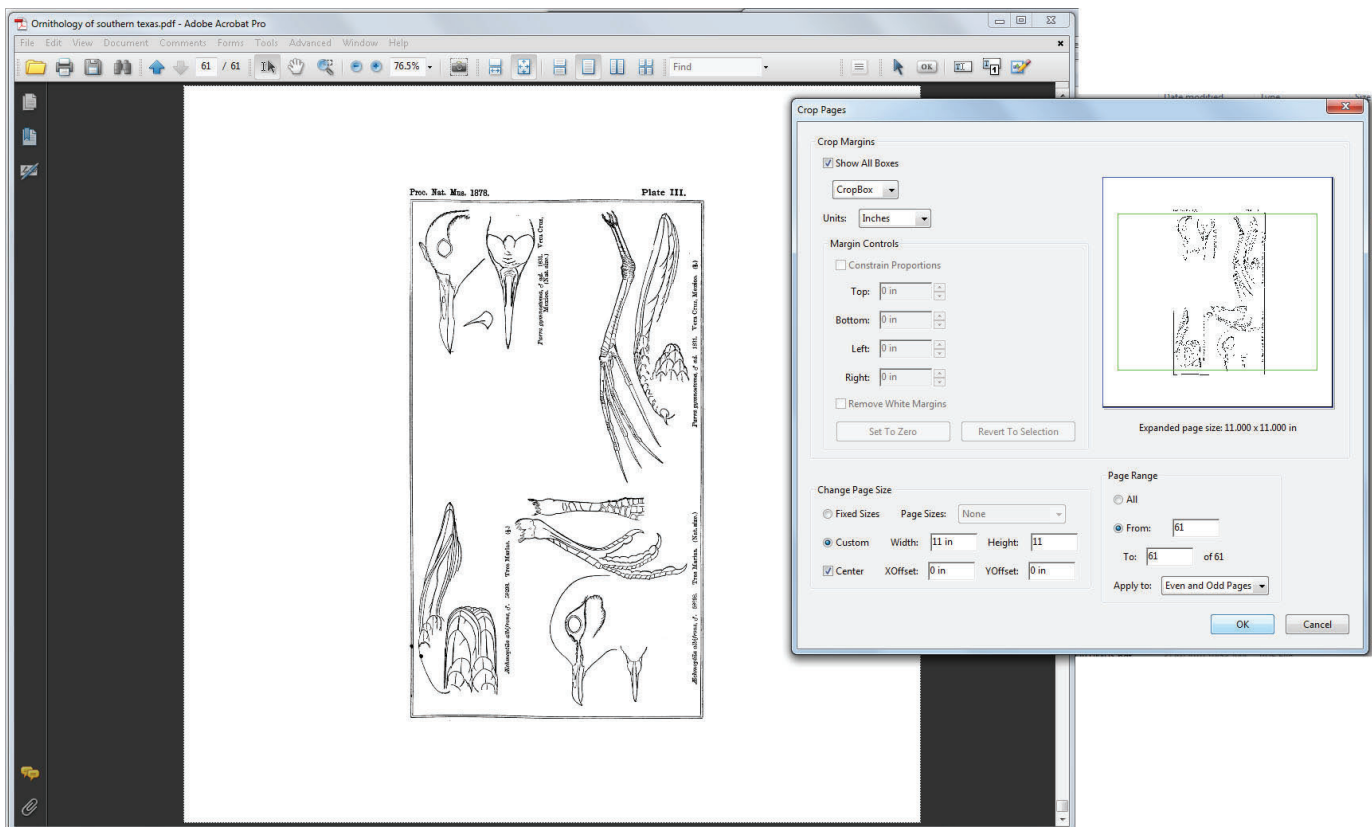
Figure 24. Acrobat cropping dialogue screen used to re-crop a "turn" page. Set the new page size as the square of the longer side, then come back and use the upper Margin Controls to crop down to the desired size.

The file is too gigantic (i.e., over 100 Mbytes).

Acrobat has an "Optimizer" function that will decrease file sizes by eliminating unnecessary hidden elements and reducing the resolution of image components. It is currently under the "Advanced" menu.

There are six different sets of settings (for images, fonts, transparency, objects, use data, and clean up); the settings for images (shown below) will usually have the largest effect.

I normally save the optimized file under a new name, so that the original is retained, just in case ...
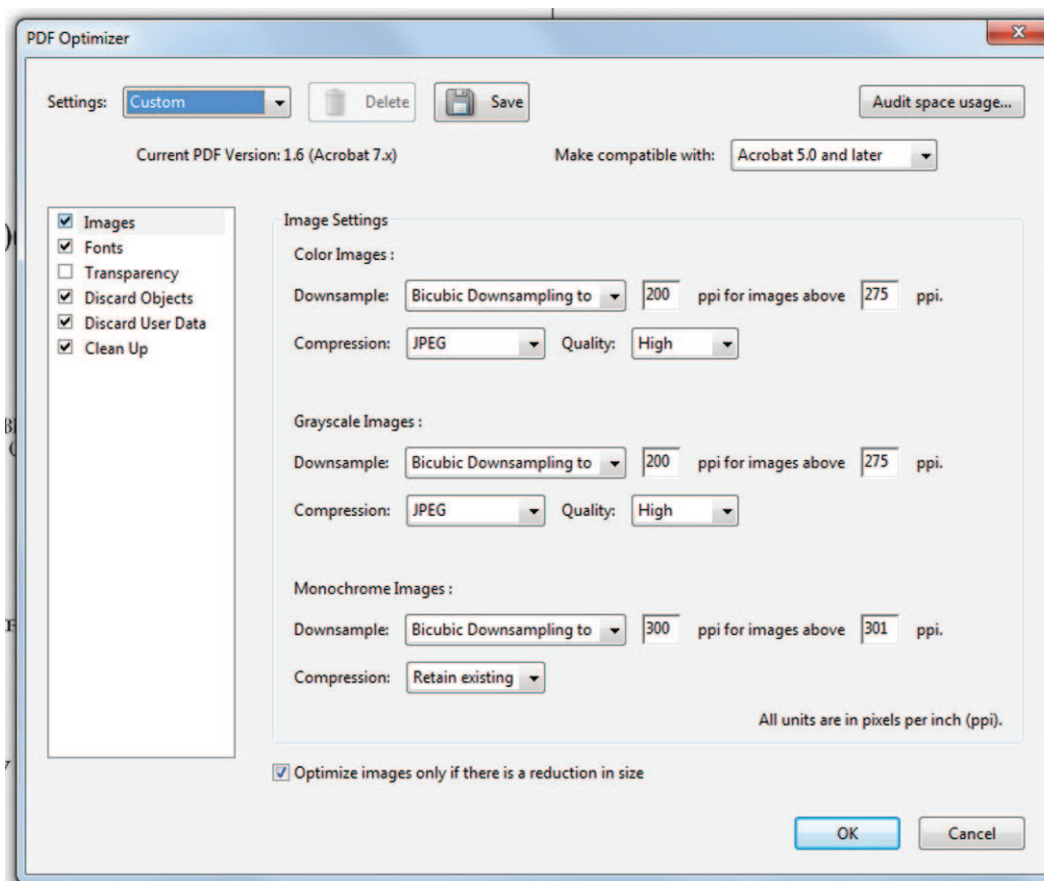
Figure 25. One of Acrobat optimizer dialogue screens. This one (for Images) usually has the largest effect on file size.

7. Adjusting Scanned Images

Many books have been written on using Photoshop to adjust and manipulate images, and some of them are very good. This section will only treat some common and elementary methods of improving the appearance of images scanned for publication in PDF documents.

Straightening and cropping should be done in that order: straighten first, crop second. To straighten, enlarge the upper left corner of the image and pull down a horizontal guideline from the top ruler bar to the top left corner of the image. Then go over to the top right corner and see how far off the guideline is from the top of the image. Rotate the image clockwise or counterclockwise in small increments (usually .5 or .25 degrees) until the guideline and top right corner of the image are aligned.

Crop by using the selection rectangle and with the image enlarged. Start at one corner and go to the opposite corner (such as upper left to lower right); then hit Alt + O P.

The **moire** issue has already been addressed, twice, so we won't repeat what has already been presented, except to say "use the descreen function of the scanner or the Gaussian blur function in Photoshop to cure the problem of swirling or checkerboard patterns."

The next most common problem with grayscale or color images scanned from printed materials is that they are **too dark and murky**. This is usually best addressed by adjusting the "Levels" in Photoshop. Think of a grayscale image as a collection of pixels with darkness ranging from 0% (white) to 100% (black). A scan, however, will usually produce an image with the pixels ranging from about 10% to about 70%, so the optimal range of 100% is reduced to about a 60%difference. Usually such an image needs to be "stretched" so that the darkest areas become 95-100% black and the lightest 0-5%. In Photoshop, select Image > Adjust > Levels and you will see a box pop up with what is called
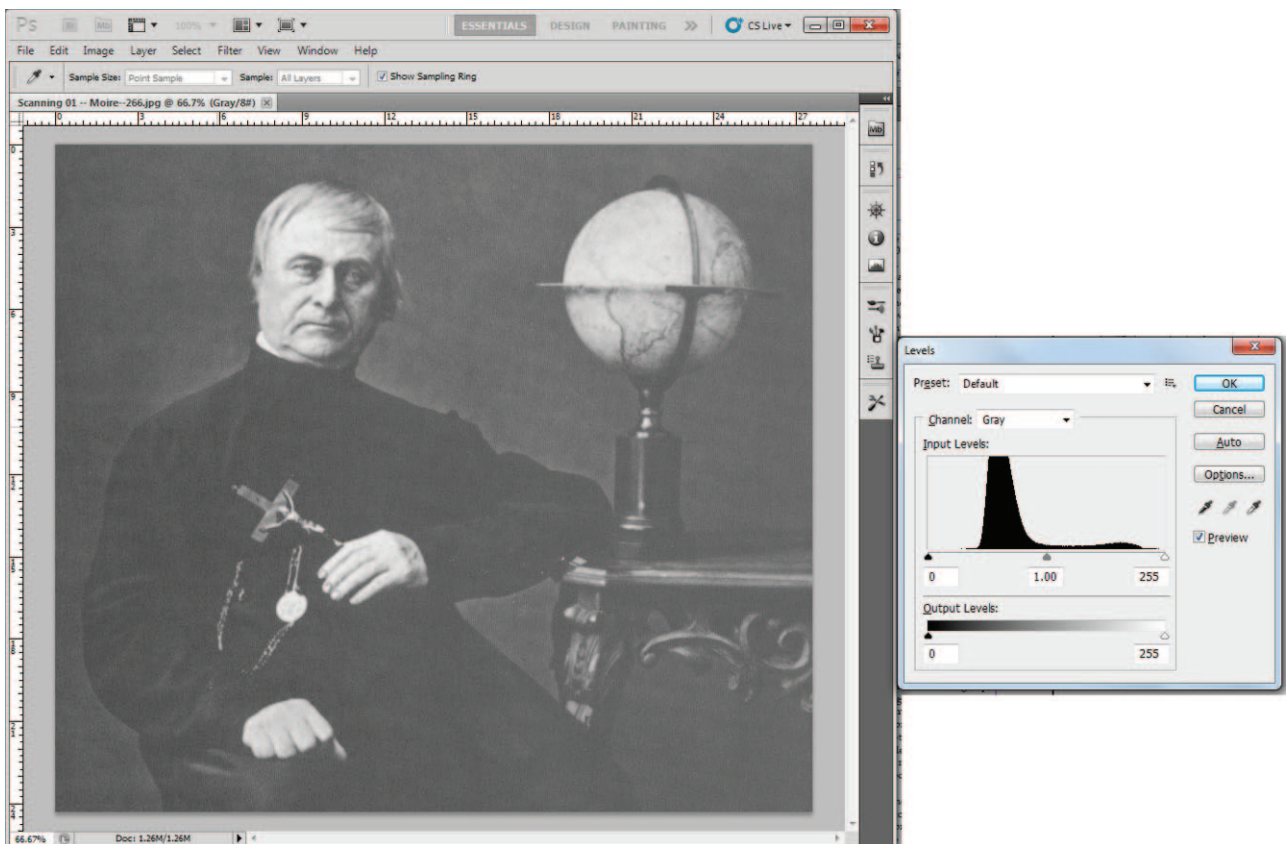
Figure 26. Image and histogram in Adobe Photoshop; Commands = Image > Adjust > Levels (or Ctrl+L). Note that most of the pixels are concentrated in the area between 50% and 75% black.
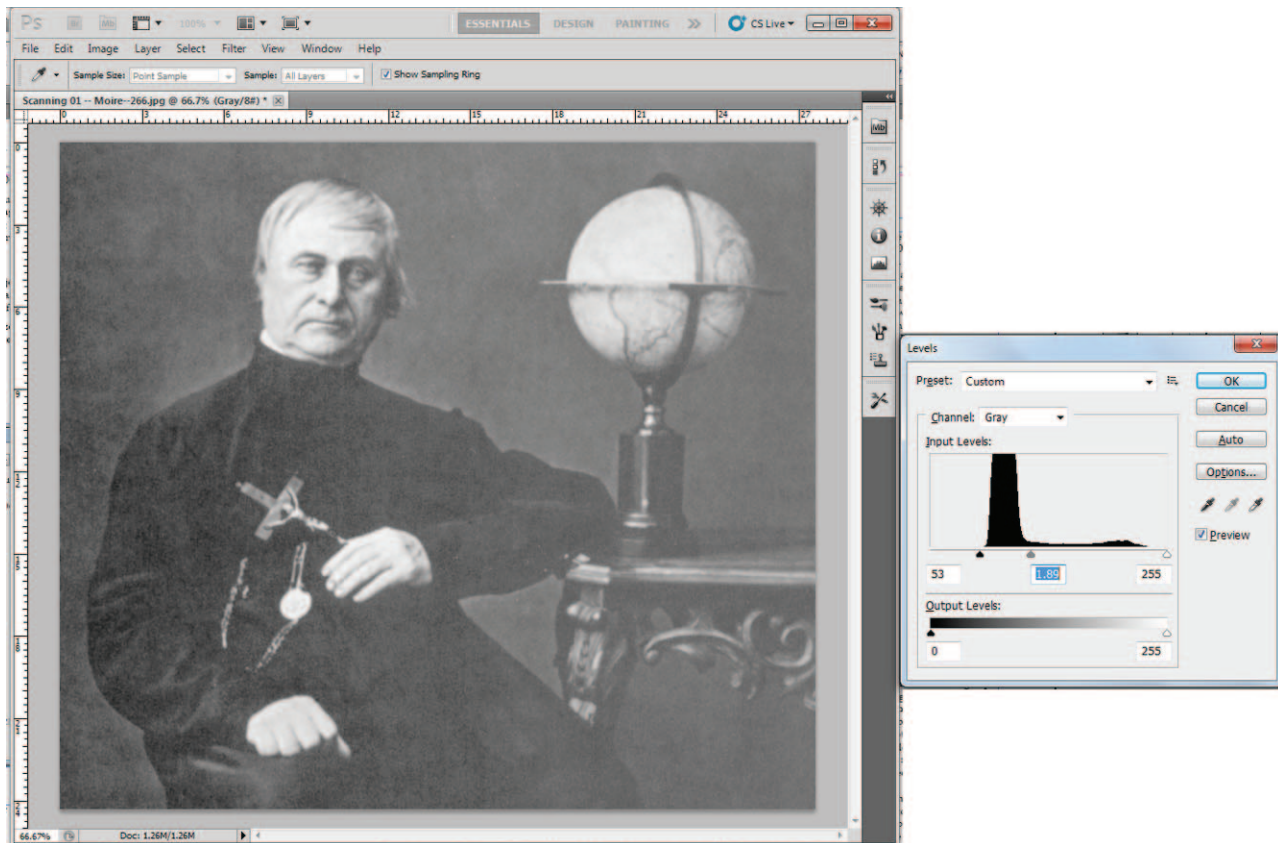
Figure 27. Image and adjusted histogram in Adobe Photoshop; the Input Levels have been adjusted to "stretch" the distribution of most pixels across the range from approximately 60% to 95% (instead of from 50% to 75%). Note that the "Preview" box is checked so that the effects on the image are seen as they are made.

a histogram of the image. This is a graph (called "Input Levels") showing the distribution of pixels along a continuum from 0% to 100% or level 0 to level 255. Note that most of the graph falls toward the middle. Below the graph you will see 3 sliding triangles, representing the 100%, 50% and 0% levels. Grab the left-most triangle (the one for 100% black) and move it to the right until it just begins to touch the area where the pixels are graphed. Normally the right-most (0%) slider does not need to be moved. Areas that show 0% black look burned out, so you don't want to create more of those. Now grab the middle slider (the one for 50% black) and move it gradually to the left to lighten up the image. Stop when the image is good and contrasty. It should "pop" in a way that it did not before. Be sure the "Preview" box is checked, so you can see the results before "accepting" them. A bit of practice will help build confidence in working on images: don't be afraid to try to improve them. After you click "OK" to accept the new image, save it. If you don't like the transformation, use Ctrl + Z to reverse the previous step. (Ctrl + Z is an invaluable tool — it always undoes the previous operation, whatever that may have been.)

A similar effect can be achieved with the Image > Adjust > Curves command, but that is somewhat more complicated and more appropriate for use on images that are to be printed rather than displayed onscreen.

Photoshop also has an automatic levels adjusting command (Image > Adjust > Auto Levels) that applies a preset filter to the image, based on its reading of the histogram. Try it if you like; if you don't like the results, just hit Ctrl + Z.

Photoshop also has a Brightness/Contrast control that can be used to adjust images. Again, play with it if you like; you can always Ctrl + Z if you don't like the results.

The Levels, Curves, and Brightness/Contrast controls can also be applied to color images. However, often the problems with these involve the color being skewed in one color direction or another—too yellow, or too blue,
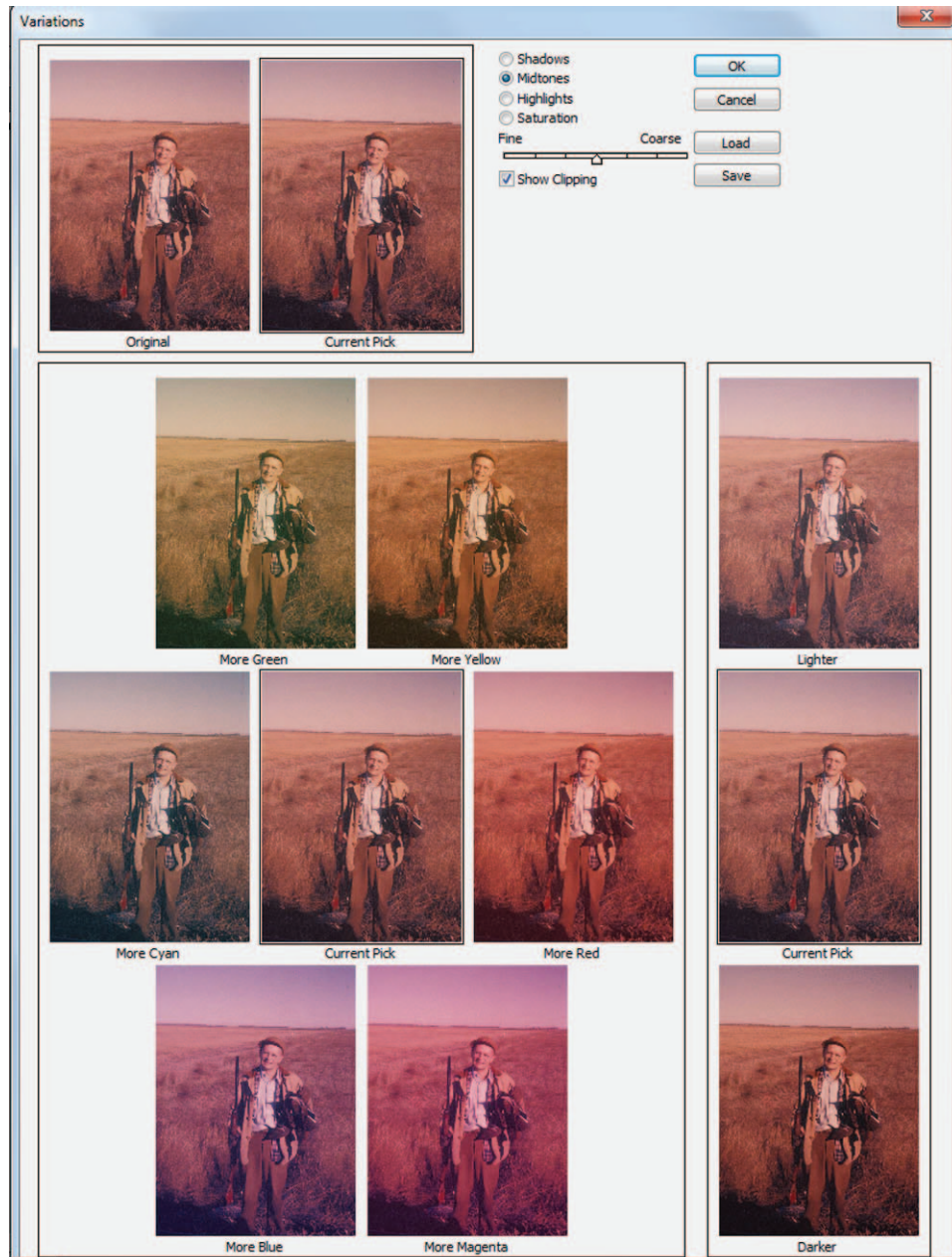
Figure 28. The "Variations" box in Adobe Photoshop. The colors of this old photo print had migrated over time, so the scan needs to be adjusted back to more natural-looking tones, in this case by selecting the "More Cyan" option and then probably the "Lighter" option as well.

etc. The easiest way to address these is through the Image > Adjust > **Variations** command. This brings up a screen that shows the current image plus variations for darker, lighter, more red, more magenta, more cyan, more blue, more green, and more yellow. Pick the one that looks the best, and accept the transformation. Note that the variations can be selected to apply to the darkest areas, lightest areas, or mid-tones. Selecting the mid-tones has the greatest effect.

Sometimes only parts of images need to be fixed, and here the "**Dodge**" and "**Burn**" tools are useful. The dodge tool looks like a lollipop or a circle on a stick; rub it over an area to reduce the darkness. The burn tool (which alternates in the same tool square as the dodge tool) looks like a hand making a "OK" sign with the thumb and forefinger. Rub it over an area to increase the darkness. Both of these tools have 2 types of settings. One setting applies the effect to either the

shadows, midtones, or highlights only (usually the midtones is what you want to affect). The other setting controls how strong the effect is—usually a setting of about 20% is appropriate—higher ones can quickly go too far.

Figure 29. Photoshop's "Dodge" and "Burn" tools and their effects. Top: Unretouched scan, showing location of Dodge and Burn tools. Bottom: Highlights of the forehead, cheeks, neck and lace have been "Dodged" (made lighter); shadows of the nose, lips, eyes, and lace have been "Burned"(made darker).

8. When You Didn't Do the Scanning, but …

Sometimes you need to work with files or images that you did not scan. Perhaps they came from Interlibrary Loan or from a well-meaning author or from an online source that didn't do a particularly good job of scanning or did not make PDF files that are up to your standards of quality. With Acrobat and Photoshop you can usually improve them, or at least solve some of the worst issues.

If the images are okay, and it is just a matter of running OCR and cropping or standardizing the page sizes, these operations can be done in Acrobat.

Sometimes, however, a header or watermark has been placed on the page images and then OCR cannot be run, because Acrobat finds there is already "live" text on the page. Sometimes you can use Acrobat's Object Touch-Up Tool to delete the header or watermark or text box, so that OCR can then be run on the pages.

A PDF file can always be converted to a set of TIFF files using Acrobat's File > Save As command and selecting TIFF as the format. Each page of the PDF is saved as a separate file. You get to select the output type—gray-scale, color, monochrome (=bitmap/line art)—and resolution. You can then work on these pages in Photoshop and recombine them into a new PDF file. (*Note:* Some encrypted files do not allow saving as TIFF; but they do allow printing to MS Office Document Writer format, which can be exported to a non-encrypted PDF file that can be saved as a set of TIFF files).

Pages saved as TIFF files can be opened in Photoshop and re-saved as PDF files and then re-combined with Acrobat into a single PDF that will run OCR.

One common problem with "found" PDF files is that the type is **too light** and faint or is **full of holes** (the Swiss cheese or shotgun effect). In the first case, it was probably bad scanning, and in the second, scanning as grayscale and converting to bitmap/line art with one of the "dither" options. The solution for these two problems is similar:

Working in Photoshop from a grayscale 600 dpi version of the page image, use the "magic wand" tool to select any black area—such as a letter. Do Select > Similar to select all black areas (i.e., all the type). Then do Select > Modify > Expand and expand the selec-
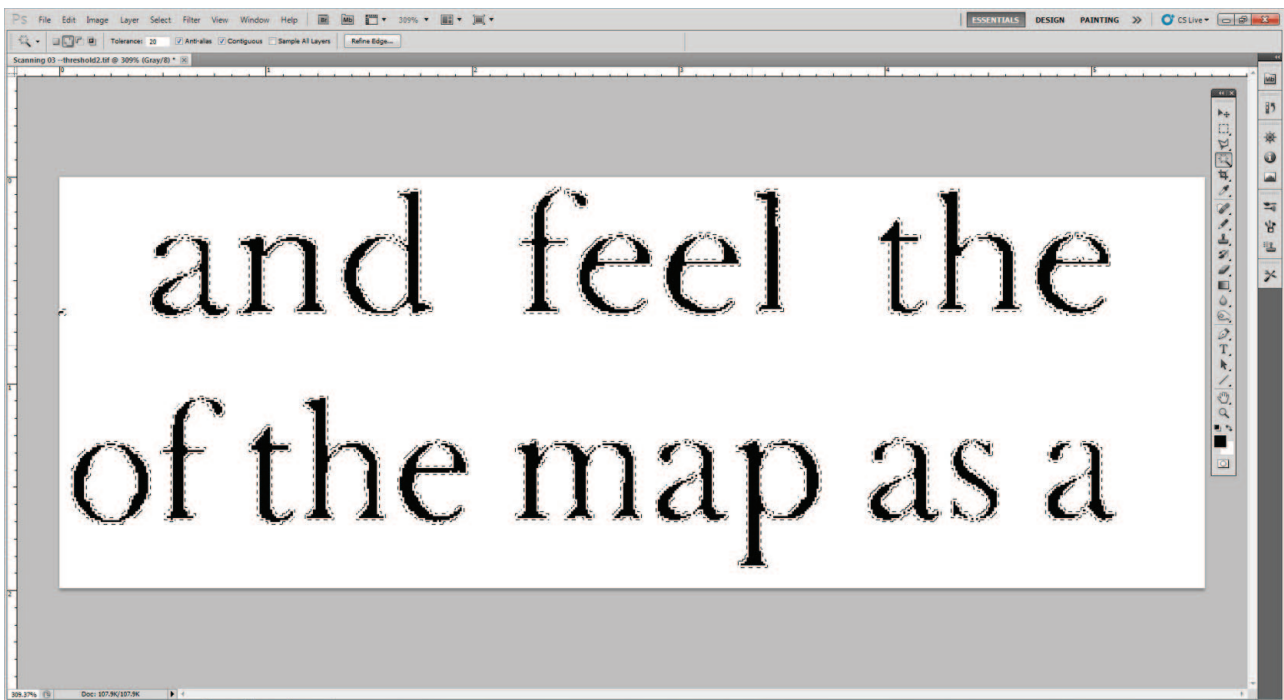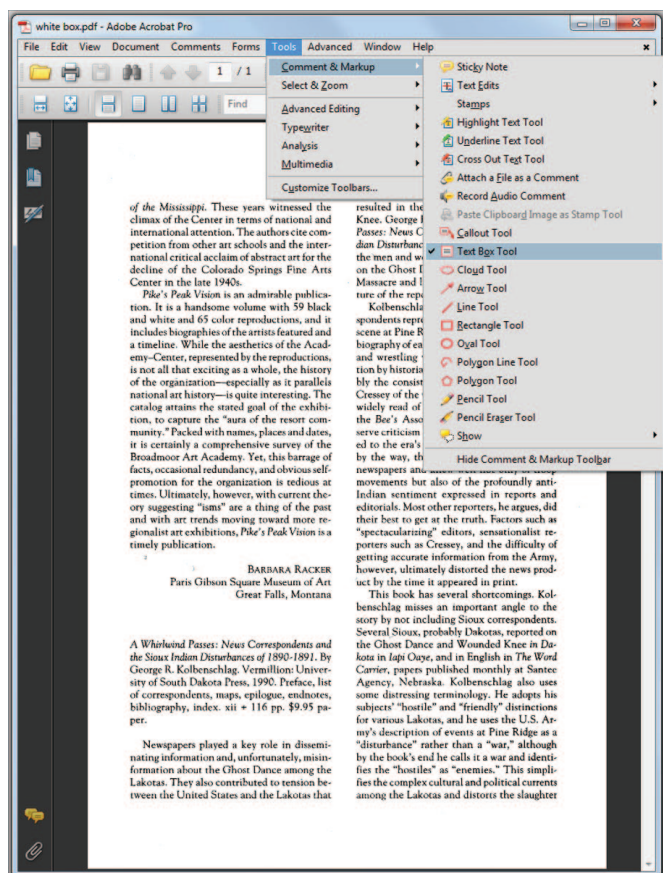

Figure 30. Photoshop's selection area for fixing faint or "Swiss cheese" type. In a 600 dpi grayscale file, use Magic wand to select one (black) letter; then Select > Similar; then Select > Modify > Expand > by 1 pixel. Next fill the area (Edit > Fill) with black, and convert back to bitmap/line art.

tion by 1 pixel. Then do Edit > Fill > Black to fill the selected area. You have just expanded all the type by 1/600th of an inch; any single pixel holes should have been patched and the type should look stronger and blacker without filling in. Convert to bitmap/line art and save as Photoshop PDF.

Figure 31. Here the selection area has been filled with black, and the image converted back to bitmap/line art. The type is not perfect, but it is much stronger and more readable.

◻   ◻   ◻

Sometimes a PDF file of an article will contain the last lines of the previous (unwanted) article or the first part of the following (also unwanted) article. This is especially common in the Book Reviews sections of journals. The solution here is to cover the unwanted portions with a white text box in Acrobat. In current releases, this is found in the Markup menus; select "no border" and make the background color white.

Figure 32. Adobe Acrobat's Text Box tool used to "white" out unwanted portions of the PDF page. Left: Location of the Text Box tool (Tools > Comment & Markup > Text Box Tool). Right: Page with unwanted portion obscured; the blue outline of the box shows only because it has been selected with the TouchUp Object tool to show its position for this illustration; normally there would be no border and the white box would blend with the page background.

□ □ □

Some publications fail to put their name, citation, and copyright information on the first page of an article. In such cases, we try to add this information in an Acrobat header or footer. You cannot style the type for italics, so we put the publication name in all capitals. Be sure to select the Page Range Options and declare page 1 only, or it will repeat on every page.
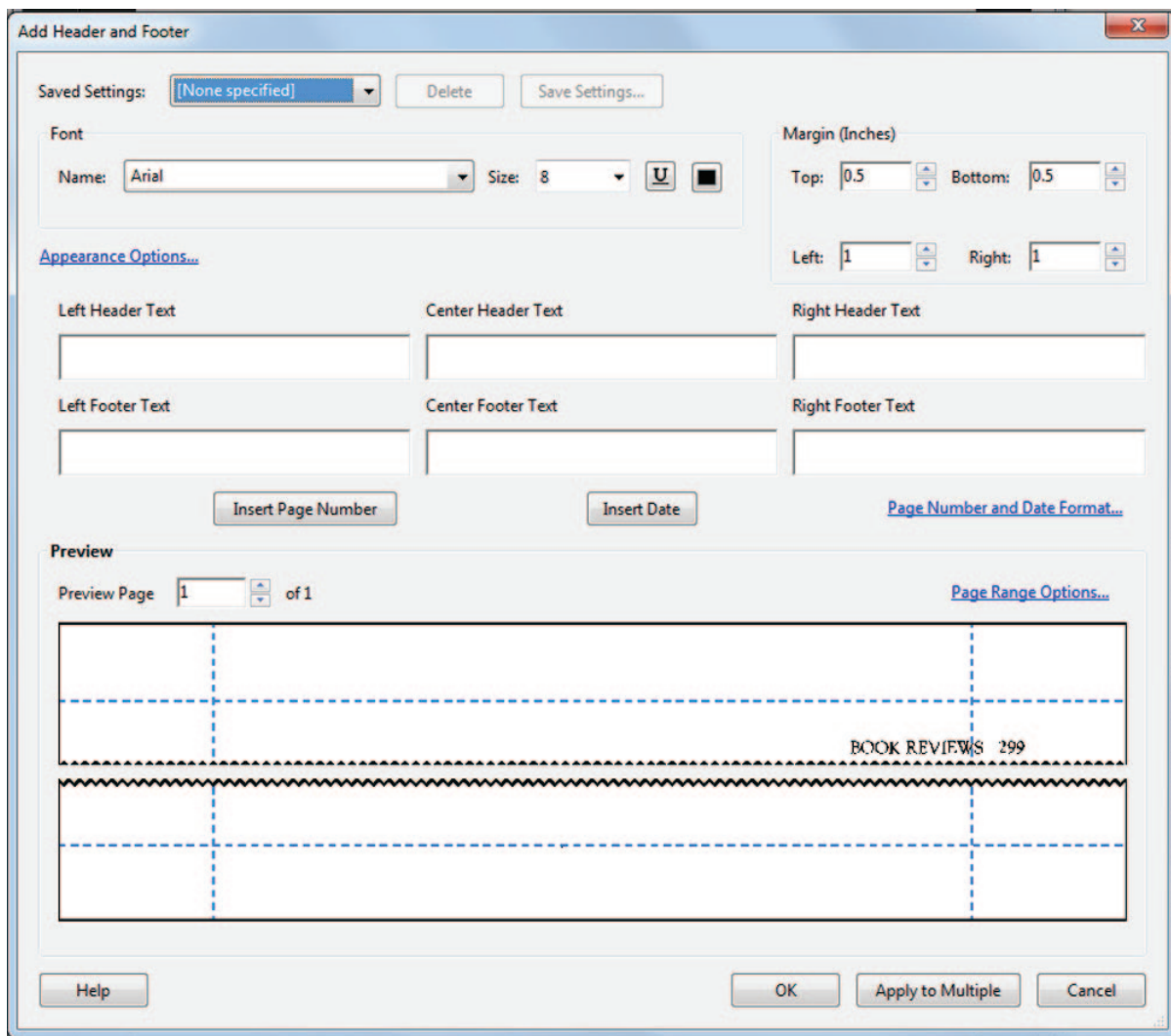
Figure 33. Adobe Acrobat's Header and Footer dialogue box and setting. You can control the typeface and size, position, and margins. This can also be used for adding pagination if needed. The "Page Range Options" allows you to put the header (or footer) on the first page only, or on all pages.

9. OCR Scanning

Sometimes you do not wish to reproduce the image of the page but you want to capture the text for purposes of revision, typesetting, etc. In such cases, OCR (optical character recognition) at the scanning stage is the most useful approach.

You will need a scanner that has the OCR function; if your scanner does not have this, then scan the pages as line art, compile them into a PDF file, and run Adobe Acrobat's OCR function (as outlined in Section 4). You can then copy the text and paste it into a word processing file.

OCR scanning has improved greatly over the past 15 years: where it used to be about 95% accurate (50 errors per 1000 characters), it is now closer to 99%(10 errors per 1000 characters). Given that an average page has about 2000 characters, that's a reduction from 100 errors to 20 errors per page.

Most OCR software will work reasonably well on text in **columns**, except when a page starts out as 1 column and then switches to 2 columns, or when there are large vertical breaks in the columns. In such cases, it may help to photocopy the pages and cut or fold them into single columns; the scanning may take longer but the improvement in accuracy makes the effort worthwhile. Alternatively, one can use the scanner's selection function to scan each area on the scanner bed separately, but this requires previewing each page.

OCR has difficulty with **odd or unusual letter forms**: roman characters do better than italic ones; oldstyle figures are often problematic; and Greek characters usually don't work at all.

OCR programs use a dictionary to help recognize words; they verify the reading against the word list to improve the accuracy. It is important for the program to know what language it is reading, and unfortunately text with multiple languages (such as English with occasional French or Spanish terms) will show lots of errors on the unrecognized language. Thus French "thé" (tea) will be recognized as English "the." If you change the language to French, then all the instances of English "the" will be recognized as "thé".

Spanish "ó" is usually recognized as "6" (unless Spanish is selected as the base language). Greek "α" (alpha) is recognized as "a", and "β" (beta) as ß (German double-s) or as "13"; and so forth. So clearly, spell-check *and* proofreading are required on all text that has been OCR scanned.

Normally, OCR scanning works best when the **image type** is set for line art, but if type is especially faint, or is printed on a background, or is reversed out of a dark color, or the gutter shadow is a problem, then using the grayscale setting may improve the results.

Usually, a **resolution** setting between 300 and 600 dpi works best. Very small type requires the higher setting, but a setting that is too high will increase the misrecognition of every dirt spot and fly dropping on the page.

Old fashioned typewriter text sometimes shows a unique problem: the letters are spaced so far apart that the OCR engine thinks each one is a separate word. In such cases, I find it is easier to search-and-delete all the spaces and then divide the words by hand than it is to delete each unwanted space one by one.

At the risk of repetitiveness, I will re-state two comments from the Scanner Software section:

- Some OCR scanning programs allow scanning **multiple pages into one file**, and this is a nice feature. Be careful, though, not to overload the system—to have it crash on the 38th page of a 40-page document will force you to start all over again.

- Some OCR programs offer MS Word or RTF word processing format as an **output type**, and this is tempting in that it may capture italics, bold, and other text characteristics. However, some programs attempt to reproduce the page layout by placing everything in separate text boxes, which are awkward and unwieldy to handle. I have found it best to go with unformatted Unicode text, producing a single text stream, and to manually restore italics, boldface, etc. as necessary.