Notes on using Adobe Acrobat--including PDF file types, modifying text, adding links, optimization, and OCR.

acrobat.txt last modified 9/27/2014
This file is from http://archivehistory.jeksite.com/download/download.htm.

This was developed for Acrobat 9 Pro, but the basic principles described here are applicable to the newer versions of Acrobat Pro. The newer versions have improved pdf editing, pdf exporting, and OCR, as well as simpler, more intuitive menus. However, my experiments indicate that ABBYY Finereader does significantly more accurate OCR than the newer versions of Acrobat. Also, Acrobat Pro version XI produces a link error when a Word document is converted to pdf and the document has named or anchored links that open at a particular spot on a webpage--which makes Acrobat XI unusable for my purposes. Acrobat 9 and X did not have this bug.

PDF CONTENT

A pdf file can contain text, formatting info, and images. The formatting information can include embedded fonts that are the font specifications used by the computer operating system to display the font. Typically, certain standard fonts that are available on all computer systems are not embedded (Helvetica (Arial), Times (Times New Roman), and Courier (Courier New)), but other fonts are embedded. The embedded fonts can significantly increase the file size. The file may not display properly or at all if fonts are not embedded.

The default font embedding is the subset of characters actually used, which can result in files that cannot be edited because the font specification for other letters are not embedded.

Some fonts have copyrights that prohibit embedding and others have restrictions on whether the embedded fonts can be used by others to modify the file (often described as "Print and preview" licenses). Font licenses that are "Installable" or "Editable" can be embedded without concern.

Images are typically in a compressed format such as JPEG. A pdf file from a scanner is typically an image of the page without text or formatting.

PDF files can be very large and need to be optimized for internet use. This includes downsampling and compressing images and using only the standard fonts that are not embedded.

Adding and modifying text and hyperlinks can make a pdf file much larger. A simple Save As with the same file name will sometimes eliminate lots of residual stuff after a pdf file has been modified.

PDF FILE TYPES

Standards have been developed for the following types of special pdf files. PDF files that do not comply with these standards are also widely used.

PDF/A is for archiving. It embeds all fonts and keeps lots of meta-data.
PDF/A-1a is for documents with text and formatting, and possibly images.
PDF/A-1b is for documents that are images without text, such as from scanning. It is also used in other cases when documents cannot be successfully converted to PDF/A-1a.

PDF/X is for printing. There are several different versions with different features. Typically embed fonts, remove internal comments and identifiers, does not compress images. Some will handle color management.

PDF/E is for engineering -- 3D stuff embedded.

These pdf file types can be selected from the Save As command.

VIEWING INTERNAL INFORMATION ABOUT A PDF FILE

Key options for viewing the internal information about a pdf file are:
File> Properties
Document> Examine Document
Advanced> PDF Optimizer -- then click the button for Audit Space Usage
Advanced> Preflight -- then click on side arrow for PDF analysis, and select "List page objects, grouped by type of object". May want to limit pages processed, which is set at the bottom of the screen. Click Analyze button. Look at Results to see resolution and compression of images, embedded and unembedded fonts, etc. Preflight can also be used to check and correct compatibility with different pdf standards (described later).

PDF FROM WORD

Acrobat> Create PDF. Acrobat> Preferences to set pdf settings for output file. Note the checkbox to create PDF/A-1a output file for archiving.
NOTE: I could not get Bookmarks to work properly using Word headings or styles.

MAKING SMALL PDF FILES (typically for display and sharing, not archiving)

In Word, set Acrobat> Preferences to Smallest File Size or to jk_smallest_6. jk_smallest_6 is same as Smallest except it embeds fonts like the Standard option so there are fewer problems.

A pdf file can also be made smaller by opening in Acrobat and then Save As> File Type> Adobe PDF Files Optimized. Click Settings button to see the options.
Can get to the same place with Advanced> PDF Optimizer.

Once on the PDF Optimizer screen, click the Audit Space Usage button to see how big the output file is and how much space is being used for images, text (content streams), and fonts.

Under Images, the options include altering the resolution and compression of the output images.

Under fonts, the fonts that are embedded and the fonts that are used but not embedded are displayed and can be controlled.

Experiment to see what works in a given situation. In general, larger file sizes should be expected for archival pdf files and smaller file sizes used for email and internet displays.

The option Document> Examine Document also displays certain internal components of a pdf file and allows individual components (such as hidden text) to be removed by setting the checkbox for the component in the list of components and then clicking the Remove button.

PDF A WEB PAGE

In Internet Explorer (IE) click Convert icon. Down-arrow icon has other options for settings etc. If Acrobat icons are not shown, right-click on an unused area of the toolbar and select Adobe PDF.

Click the Select icon to select only parts of the web page to convert. The blue line around a section toggles on and off with each click. Multiple areas can be selected. The red rectangle jumps around and does not do anything useful that I have found. I could not keep Selection working reliable. (Google Chrome has pdf extension that work better)

Can also make a pdf from a web page within Acrobat. Create PDF> PDF from Web Page. Can click Capture Multiple Levels icon to get linked pages.

DISPLAY PROPERTIES WHEN PDF FILE IS OPENED FOR VIEWING

Set under Files> Properties> Intitial View.

SHORTCUTS

Ctrl-Z to undo.

ADDING NEW TEXT

Tools> Typwriter> Show Typewriter Tool Bar is easiest way to add text. Clicking at a point sets the cursor ready to enter text. Set the font from the toolbar.

Tools> Advanced Editings> Touchup Text Tool is another way to add text. After toggling this on, put mouse cursor at point to enter text and then Ctrl-click that point. A dialog box opens to select font and then a text box opens at the selected point.

Tools> Comment & Markup> Text Box Tool adds a text box that can have a border. Click twice to edit text rather than border. To set font properties etc. Show properties toolbar View> Toolbars > Properties Bar.

MODIFYING/REPLACING TEXT

Tools> Advanced Editings> Touchup Text Tool can be used to modify text. After toggling this on, put mouse cursor at point to change, delete, or enter text and then click that point. The text can be changed, but a warning may come up that the exact font is not available.

Move a text box with Tools> Advanced Editing> Touchup Object Tool.

Redaction is another way to remove or cover text.
View> Toolbars> Redaction or right-click unused area of toolbar and select Redaction. Set redaction properties. Set Redacted Area Fill Cover to white to make text or images disappear. Redaction and Typewriter can be used when Touchup Text Tool does not work because needed fonts are not available.

ADD HYPERLINKS

Tools> Advanced Editing> Link Tool or activate Advanced Editing Toolbar.

LINK TO SPECIFIC PDF PAGE

In link to a pdf file, put e.g., file.pdf#page=3 to open the pdf file on page 3. Then drag the mouse pointer to make a rectangular area that is the link button. Dialog box comes up to select properties, including invisible rectangle. The text for the link is set separately and will usually exist or be set first. Set the mouse cursor to the hand tool to turn of the link tool. Then test the link.

BOOKMARKS

Can add from document structure or manual assignment. Easiest is to use the bookmark icons on the left frame of Acrobat. The gears icon is used to define new bookmarks.

STANDARDS COMPATABILITIY

Advanced> Preflight is used to check compatibility of file with standards for the pdf file type. This is very useful when working precisely with the PDF/A format. Select the profile for the file type by clicking the right arrow. Then select the specific processing (verify or convert) and click the Analyze button or Analyze and Fix button. The output is on the Results tab.

Some problems cannot be corrected at pdf stage and need to go back to the original source file such as Word document.

NOTE: Save As output includes different file types. The output file is adapted to the file type without going through the Preflight process.

TAGS

Tags are markers similar to html markup, and are useful for bookmarking and for accessibility. Can tag in Acrobat or Word. In Word, Acrobat> Preferences> Settings> Enable Accessibility and Reflow with Tagged PDF. This makes tags Styles.

In Acrobat, View> Navigation Panels> Tags. Check for tags with Advanced> Accessibility> Quick Check.

OCR

Acrobat can do optical character recognition (OCR) when scanning a document and also on PDF files that have not previously had OCR done. OCR with Acrbat tends to have many errors (more than some other OCR software) and does not have a way to correct the errors within a pdf file. (The OCR text can be output to a word processing file and corrected, and then a new pdf file made from the word processing file.)

Scanning or image resolution of 300 ppi is recommended, and 600 ppi is better for small fonts or noisy originals.

The default OCR method is Searchable Image, which does some cleanup of the image and adds an invisible overlay that has the text as letters. The font HiddenHorzOCR is embedded for the letters and has associated fonts Helvetica (Arial) and Times (Times New Roman). The file size can be very large. The file can be Saved As to a word processing file or text file that has the letters and words from OCR. The OCR method Searchable Image Exact does the same thing, but without any cleanup or modification to the original image.

An alternative OCR method is ClearScan, which creates special fonts from the image and then uses these fonts rather than the original image. The file size can be much smaller and the file display and printing looks essentially identical in my experiments. Special fonts that have names starting with Fd and then a number are created and embedded in the output pdf file. These fonts show on the File> Properties> Fonts display, but not on the Advanced> PDF Optimizer> Fonts display.

When a file with ClearScan OCR is Saved As to a word processing file (.doc or .rtf), it is output as an image of the page rather than with the letters

and words recognized. The letters and words from the ClearScan OCR can be output and viewed with Save As "Text (Accessible)". (Save As Plain Text does not work for me.)

Scanning is initiated with File>Create PDF>From Scanner>. The main options for scanning are Black & White Document, Grayscale Document, and Color Document. Normally Black & White Document is for text only documents and would have OCR run automatically. The other options are for documents with pictures and would not have OCR run by default.

OCR during scanning is controlled by File>Create PDF>From Scanner>Configure Presets. The Input section also sets the output file properties. The Color Mode is normally set to Grayscale for both Black & White and Grayscale scanning. Resolution controls the resolution of the output file and greatly affects the output file size. 300 DPI (ppi) is a standard setting for OCR. The setting for Optimization quality has little impact on final output file size for text.

OCR is handled with the checkbox Make Searchable (Run OCR). The options button controls the OCR process. PDF Output Style is set to ClearScan here. The option Downsample Image resolution apparently reduces the resolution if it is above the specified value. The default of Lowest (600 dpi) is safe, but 300 dip could be used with little loss of OCR accuracy in most cases.

OCR of an existing file is done with Documents>OCR Text Recognition> Recognize Text using OCR. The Edit button brings up the dialog box for OCR settings (same as for scanning above).

After doing OCR with Acrobat, the pdf file can be saved as a .doc, .rtf, or Text (Accessible) file to see what the OCR produced. Also Documents>Examine Document will show the hidden text in some cases. Acrobat does not have a direct way for the user to correct the actual letters that are generated with OCR. Thus there can be, and often are, many errors that are unknown. The option Documents>OCR Text Recognition>Find All OCR Suspects is supposed to show questionable cases, but I cannot get that to work. It supposedly works with Searchable Image, but not ClearScan, but it does not work with either for me. Save As to a word processing or accessible text output file is an easy way to see what the OCR did.

(In general, ABBYY FineReader is much more accurate and reliable for OCR than Acrobat and also allows questionable cases and errors to be identified and resolved. To see the difference, compare .doc files made with Acrobat OCR and with FineReader for the same item. The final files are also often much smaller with FineReader Pro. Options in FineReader determine whether the original page image and/or the OCR text are kept and displayed. The fonts may not match the original fonts if the OCR text is displayed.)

OTHER FEATURES OF ACROBAT

Distiller can be set to automatically check a folder and convert any file in the folder to pdf. Open Distiller, Settings> Watched Folders.

Can make fill-in forms in pdf.

Can imbed video etc.

Lots of commenting, reviewing, printing options.

Can set page numbers and headers and footers.

Batch processing options. Advanced> Document Processing> Batch Processing.