Notes on using ABBYY FineReader for OCR -- based on FineReader 11 Professional.

abbyy_finereader.txt last modified 3/5/2014
This file is from http://archivehistory.jeksite.com/download/download.htm.

I have found ABBYY FineReader to be the best option for OCR accuracy; however,
it is weak on handling the format of output word processing files.

Basic steps for using FineReader are:
1. Set the options;
2. Obtain images of pages by either scanning or using existing images from a camera, previous scanning, or pdf file;
3. Analyze images to determine different types of areas (text, tables, pictures);
4. Read the areas to make the actual content;
5. Manually make corrections in FineReader;
6. Save the resulting output to a file;
7. Check the output file and make additional corrections and adjustments.

NOTE: Re-doing Analyze will cause manual corrections to be lost.

KEY OPTIONS

Click dropdown menu Tools>Options or Options icon (wrench) on toolbar.

Advanced>Font Matching specifies which fonts will be used for OCR output. By default, FineReader often uses unusual fonts. These are embedded in pdf files and make very large files. Font Matching can be set to use only Acrobat PDF fonts that will not be embedded (Arial, Times New Roman, Courier New). Settings from the previous run of FineReader are used by default.

Save>RTF/DOC/DOCX> has checkboxes for
- Set output page size
- Keep headers and footers,
- Keep page breaks,
- Keep line breaks
- Retain text and background colors.
- Picture settings (print=300 ppi, screen=200 ppi, web=96 ppi)

To maintain same formatting in the output file, keep headers and footers, page breaks, and line breaks. Retain background colors may be useful in some cases but generally not as default.
Can control ppi of images by the picture settings.
Optimal picture quality is set as:
  Resolution: 300 dpi or Original
  Quality: Quality loss not allowed, which makes PNG output files.
    Alternatively, Quality loss allowed, with quality slide set to 80% or higher gives good results with JPEG output--but is subject to more loss from later processing

Save>PDF
- Set output page size
- Save Mode set to "Text and pictures only" does not keep original page image and makes output file much smaller, but fonts in output may not exactly match original fonts.
- Enable Tagged PDF allows a PDF file to be flowed to a mobile device.
- Font Settings set to "Use predefined fonts" only uses the fonts that are in Adobe Acrobat (Times New Roman, Arial, Courier New).
- Clear Embed Fonts checkbox to not embed the fonts.

View>Font used to display plain text is also the font that is used by default for entering text after an Exact Copy has been imported into

Word. Default is Arial Unicode MS, which will result in embedded fonts in PDF. Change to plain Arial to avoid that.

Scan/Open tab has checkboxes for whether pages are automatically rotated and split.

Document tab has buttons for color or black and white processing. Use color in most cases, including grayscale. Black and white is two colors: completely black or completely white.

Unfortunately, the feature in FineReader Version 10 to save sets of options for re-use was apparently dropped in Version 11.

OBTAIN IMAGES OF THE PAGES

The original image of a page for OCR can come from scanning, from a digital camera, or from a pdf file. For pdf, FineReader regenerates the OCR and does not use the text in the file (if there is any text).

The best accuracy is obtained if scanning is done with FineReader; however, this can require preview and final scans for every page image if the crop area is adjusted for each image. In that case, scanning with VueScan can be done with only one scan per page. However, FineReader tends to handle bold type poorly if OCR is from a previous page image (jpg vs tif does not seem to matter).

Processing is easiest if the text is a consistent distance from the edge of the paper on all pages. Putting a marker (e.g., stickynote) on the scanner edge can help consistent alignment.

FineReader can automatically handle rotations and dividing into separate pages if two pages on source image.

Minimum recommended page image resolution is 300 ppi (dpi). Scanning at higher resolution (400 - 600 ppi) can help handle small type (9 points or smaller) or a page with fading or distortion. Using color scans works best for low contrast documents and for any pictures.

FineReader has an option to Edit Image, but this is rarely needed. Most of the optimization modifications are done automatically as well as can be done. However, occasionally a split page may be missed by FineReader and need to be split manually with Edit Image.

ANALYZE IMAGES

Page headers and footers, tables, references, etc. will be handled better if they are made their own text box. Redoing the analysis step will loose manual changes.

FineReader often makes odd shaped recognition areas (text, pictures, tables). It is much more reliable for subsequent processing if each recognition area is a simple rectangle. This may require deleting and re-creating some recognition areas.

In FineReader, click bottom of panel or tap Ctrl-F5 to toggle display of an enlarged area that can be used to precisely set areas and make corrections.

READ AREAS

- Clicking on Read icon at on top toolbar reads or re-reads all pages.
- Clicking on Read icon on the Image toolbar reads or re-reads only the

displayed page.
- When an area in the image is selected (clicked with the mouse), right-clicking the mouse button gives the option of reading the specific area. Right-clicking when mouse is over part of the page not defined as an area gives the option of rereading the full page (not all pages).

CORRECTIONS IN FINEREADER

Getting the text, picture, and table boxes located properly is important for making output for word processing. These are initially set automatically but can be adjusted in the Image window. Click on an area to select it. The sides can be dragged to change dimensions. Tapping the Delete key will delete a selected area. New areas can be drawn by clicking the icon for the type of area (text, picture) and then dragging a new rectangular area. Right-clicking the mouse gives more options, such as changing an area to a different type.

In general, for word processing the headers, footers, body text, and picture captions should all be in different text boxes.

By default, FineReader excludes lines, such a separating footnotes from the main body of text. Can manually adjust the text box to include the lines and they will be in the output file.

When creating word processing output files, the easiest practice is usually to make minimal corrections to content and formatting in FineReader and do most corrections with the word processing program (e.g. Word).

FineReader uses Styles and typically makes dozens of styles. These are often variations and the same type of item in the document (e.g. body text or a certain level of heading) will have different styles at different places. This means that several different styles need to be changed to change the formatting for that type of item--which basically defeats the purpose in using styles. The styles are not named clearly (most are Body Text (N), where N is a number) and font size is handled separately--all of which makes the FineReader styles very difficult and inefficent to use.

The styles are transferred to Word. Font size in Word may be slightly different (usually smaller) than in FineReader. My experience has been that FineReader often makes the fonts too small and the margins too big, particularly when using page images from a camera. Adjusting these is difficult in FineReader.

If formatting is modified in FineReader, the properties of the styles can be changed with Tools>Style Editor. Fonts and boldness often need modification. A different style can be applied to certain text by selecting the text and then choosing the style from the dropdown list on top toolbar.

The paragraph icon on the upper right of edit window shows nonprintable characters such as paragraph markers, manual line breaks, and tabs. Shift-Enter adds a manual line break without starting a new paragraph.

FineReader often makes a mess of any table with more than trivial complexity, particularly with small fonts and/or footnotes. The fastest strategy is often to set the table as a picture/image and just handle it as an image in Word or pdf. However, tables can be modified in FineReader by right-clicking the mouse over the table to get options such as making or deleting lines for columns or rows. Also, selecting one or more cells (click on a cell and Shift-click to select adjacent cells) and then right-clicking allows splitting and merging cells.

FineReader identifies and highlights cases when the OCR was questionable. Can sequentially go through them with either the Verification Window or by

clicking the Next and Previous Error icons--or by just going through each page. The questionable cases can often be handled more quickly in Word.

OUTPUT

Output for word processing can have varying degrees of matching the exact formatting of the original document--ranging from Exact Copy, to Editable Copy, to formatted text, to plain text. Editable Copy is often a useful balance.

Exact Copy attemtps to make the output match the orginal document, but uses Word Frames and makes it difficult to adjust fonts and line up margins for different pages. See section below on Exact Copy.

Editable Copy is easier to make adjustments, but tends to less precisely match the original document and pictures can be very confusing to work with.

PDF output can be made directly from FineReader, or FineReader can output to a word processor and the word processor used to make a pdf file. Output to a word processor usually works best for cases to be made into pdf. There are usually corrections that will need to be made and can be done easily in with word processing program. With images from a camera, FineReader usualy makes the fonts too small and margins too large. The fonts and margins are much easier to adjust in word processing than in FineReader, and Editable Copy is the fastest option by far for this.

If pdf output is made directly from FineReader, the pdf output can be either the letters and words from the OCR or can be the original page images with the OCR text hidden (but searchable). Keeping the original page images makes the pdf files much larger, but exactly matches the original document.

Output to txt will have line and page breaks (if those options are set) even though those do not show on the display in FineReader.

Can save FineReader document that has original images and settings to make revisions later.

WORKING WITH OUTPUT FILES IN WORD

The FineReader option to keep line breaks puts a manual line break at the end of each line in Word (same as Shift-Enter). This breaks a line without starting a new paragraph. The line breaks and paragraphs can be seen by turning on view paragraph markers (Home> paragraph icon).

In Word, display the list of styles by clicking the lower-right arrow on styles section of toolbar. Typically the dozens of styles created by FineReader have many variations and different styles are used in different places for the same type of documtent item. This means that format changes require that either many styles are modified, or the styles need to be re-assigned.

Making format changes outside of the styles is usually the best option when it can be done. Selecting text and directly changing font properties will override the styles.

If styles are wanted for certain document items, such as headings and indented lines, styles can be modifed or created and then applied to each of the relevant items. Right-clicking on a style in the style list gives the option to modify the style. Clicking on an item in the document shows the currently applied style as highlighted in the style list. However, for documents from FineReader, the highlighted style is often not the modified

style. To apply the modified style, click on the item in the document and then click on the desired style in the style list--even if it appears to be the same style.

The picture or image options in Word are not optimal from FineReader. The Word option to compress the pictures upon saving is turned on and needs to be turned off if high resoloution pictures are wanted. This should be done before the Word file is saved and is done for a Word document with: Office button (upper left)> Save As> Tools (button)> Compress Pictures> Options (button)> clear the checkbox Automatically perform basic compression on save.
In addition, the Word output files from FineReader do not have the option to Change Picture (when the picture is right-clicked or under the Picture Tools Format> button for Change picture). Pasting the content from FineReader into a new blank Word document will make the Change Picture option active.

FineReader often makes Word Frames for enclosing text and pictures. Frames can be modified and removed with the Frames command that is not available by default in Word (2007). Clicking the mouse inside a Frame shows the Frame. Double-clicking on the edge of the Frame brings up the Frames command. Alternatively, the command can be put on the Quick Access Toolbar by selecting Office Button> Word Options> Customize> Choose Commands From> All Commands. Then find the Frame command and put it on the Quick Access Toolbar.

NOTE that tapping the Delete key when a Frame is highlighted deletes the Frame and its contents. However, clicking the Remove Frame button on the Frame control, removes the Frame without deleting its contents--but may move, resize, and reformat the contents in unpredictable ways, particularly when columns are involved.

Pictures can be very confusing to work with. Some complications include:
(a) Pictures are usually put inside a Word Frame. In these cases, the picture is moved by moving the Frame rather than the picture inside the Frame. The Frame usually can be seen by clicking just to the left of the picture. The Frame may (or may not) determines how far text is from the picture. The caption may be in the Frame or in a separate Frame. The Frames may overlap and the main Frame may need to be moved to access the caption Frame.
(b) In some cases, the easiest option is to get rid of the Frame. (However, Frames may also be useful in some cases.) The anchor for the picture is inside the Frame so deleting the Frame with the delete key also deletes the picture. Use the Frame command to Remove Frame. In addition, the position of the picture is locked or grayed out so it cannot be changed. To undo this, click the picture to bring up the tab Picture Tools Format> Position. Then click one of the icons to put the picture someplace on the page. This moves it out of the Frame. Then drag it back where it should be. It can be relocated as needed after this.

EDITABLE COPY OUTPUT

Editable copy is usually the easiest to work with. If page and line breaks are both kept, it will match the original file well. But, pictures and graphics can be confusing to work with--see below. Editable Copy allows good control of margins (unlike Exact Copy) but can be unpredeictable and difficult to work with when the document has pictures and columns.

FinePrint puts headers and footers in Word as Text Boxes in the Header and Footer section. The Text Boxes are relative to the page and so do not change position when the margins change. Need to manually move them or set them relative to the margins if the margins are changed. If the margins are changed before the text boxes are moved, the header and footers end up buried in the text and can be moved by opening the headers and footers in Word and then clicking on the buried header or footer and moving the text

box.

FineReader also makes many different sections, and the headers, footers, and margins have to be modified for each section. For simple documents with the same headers, footers, and margins, the easiest practice is to eliminate the different section.

The basic steps in Word for adjusting the output for Editable Copy are:

1. In FineReader, make sure that:
   - headers and footers are identified as separate text blocks,
   - split pages were done on all page images with two pages,
   - complex tables are set as images, or checked carefully and fixed.
2. If the document has pictures that should not have the resolution reduced, set the document to not compress pictures (as described above). Other issues when handling pictures are described below.
3. If the headers, footers, and margins are the same for all the document, use the Replace function to find all section breaks (^b) and replace them with page breaks (^m). This should leave the headers and footers and make the document one section. Alternatively, if different sections are needed, the section breaks can be searched and replaced individually to leave different section.
4. Set the margins for the whole document (or the relevant section).
5. The most reliable practice is to delete all the FineReader styles and then attach a template and apply appropriate styles manually. A Word macro for deleting all user created styles is described in the section below on Word Macros Used with Abbyy FineReader below.
   Alternatively, the FineReader styles can be kept and worked around by:
   Under font, select the entire document and then:
   - set the font type and size,
   - set the scale to 100% and spacing and position to normal,
   - if there is no bold type in the original also toggle bold on and off to get rid of extraneous bold type.
6. If using the FineReader styles, under paragraph, select the entire document and then:
   - set the right and left indents to 0 if the document has no indents,
   - set the justification if it is constant for the document,
   - set the line spacing to single or the desired line spacing,
   - leave the other fields alone (notably first line indent and spacing before and after paragrapshs--unless these are constant for the entire document).
7. Open the headers and footers (i.e., double click in the header area):
   - click on the header text (which will often be buried in the body text) and then move it to the header position. This will typicially need to be done for both the odd and even pages and for each section.
   - If page numbers did not take, add page numbers--also typically for both odd and even pages and for each section.

Some common issues with pictures and editable copy include:

(a) Pictures may jump around unpredictably when moved, or cannot be moved. This can usually be stopped by clicking the Frame and then the Frame Tool and clearing the checkbox for "Move with text". It can be usefull to clear this checkbox on all pictures--but it comes back on sometimes.

(b) Some paragraph markers may be impossible to delete if they are the anchor for a Frame or picture. Clicking on a picture displays the anchor as a symbol. The anchor can be clicked and dragged to another paragraph.

(c) Formatting can be lost when cutting and pasting. In particular, Frames that normally span columns can be reduced to fit in a column. These need to be manually resized.

(d) The picture may need to be resized within the Frame.

EXACT COPY OUTPUT

When the output is set to Exact Copy for Word, it makes Frames in Word and puts a section break between each page. Frames can be moved as an object and the edges can be dragged to make lines wrap more like the original. Margins are more difficult to manage with Frames.

NOTE: Increasing font size, line spacine, or adding text will typically require that a frame be expanded and adds significant effort to making corrections or adjustment.

By default, frames are defined relative to the page, which means they do not move as the margins or text outside the frame are changed. Margins are also set to zero. In Word, Frame can be defined relative to a paragraph (or margin), which makes them move as text and margins change. The position of the frame are set separately for horizontal and vertical. These are all set with the Frames command.

NOTE: If the default of centering the content between the horizontal and vertical edges of the paper is not wanted, each Frame has to be moved or reset to be based on the margins rather than the page.

Applying a paragraph or heading style inside a Frame causes the paragraph or text to jump out of the Frame to the top of the page. A character style can be applied or the assigned style can be modified without jumping out of the Frame. Textboxes in Word are similar to Frames but paragraph and heading styles can be applied to textboxes.

For a document made with Exact Copy, by default new text added with Word outside of a frame is only 1 point high so does not show. Set font size to desired size before entering text. Displaying paragraph markers can also help. The default font for the new text is the font set in FineReader for font used to display plain text under the Options>View> tab.

In Word, paragraphs within a Frame have left and right indents that may need to be changed to get the line lengths the same for different paragraphs.

One way to get rid of frames and keep some of the formatting is Ctrl-A and then Ctrl-Q.

The optimal practice is to convert the Frames to Text Boxes and move the pictures out of Frames. Word macros for doing this are described in the next section.

LESSONS AND METHODS FROM USING FINEREADER TO REPUBLISH A BOOK

I wanted to help an author republish a book on the history of schools that was out of print. The book was originally printed with two columns of text on each page and typically several pictures of various sizes. The author did not have an electronic copy of the final text or electronic copies of many of the pictures. Developing the layout had taken substantial effort by the previous printer. Under these circumstances the best option appeared to be to do OCR from an original copy of the book and then update pictures that were available.

Experiments quickly revealed that using FineReader Exact copy was the best option because it would retain the original layout. Editable copy would basically require re-doing the layout and I did not have time for that.

For consistency, efficiency, and the ability to do the book in sections that could be combined later, it was essential to get rid of the dozens of junk styles from FineReader and apply the small number of styles that was actually

needed (basic text, a few headers, and picture captions). Efforts to use standard Word macro commands to delete the styles created by FineReader found that the junk styles produced errors. Experiments revealed that the styles could be deleted by repeated runs of the standard macro commands.

I also found that paragraph styles could not be applied in the frames used for text with the Exact copy option (using Word 2007). The Frames needed to be converted to Word Text Boxes for styles to be applied.

Aligning the text boxes and pictures with the margins would be very time consuming if done manually. FineReader centered the output on a page in Word and I could not find FineReader settings to get the mirrored margins needed for a book.

Word macros were developed to do most of the needed processing. The macro to convert text and pictures in frames to text boxes and to standard pictures, and then to align the textboxes and pictures with newly set margins was relatively complicated, but the project would not have been feasible without it. The final steps are described below.

1. The original book was scanned with VueScan to create an image of each page. VueScan was used because it has a Newspaper setting that descreened the pictures (very important) and because VueScan can do scanning like this with one scan (rather than the usual process of a preview and final scan).

2. ABBYY FineReader was used to do OCR on the page images from above, making an Exact Copy output. The FineReader options were set to keep page breaks, but not line breaks. Custom picture settings kept the original resolution and quality loss was not allowed. The most important step in FineReader was to set the text and picture identification boxes carefully--deleting any that FineReader made odd shaped and replacing them with simple, non-overlapping rectangles. The text boxes were combined when possible into one larger box.

3. In Word, a series of macros was run to apply standard formatting. Those are described below. Then the styles for headers and picture captions were applied using Quick Styles. Finally, the location of a few text boxes or pictures was adjusted if needed--often by using keyboard shortcuts with macros for moving the objects a few points (as described below).

WORD MACROS USED WITH ABBYY FINEREADER

Macros PJ_Init1 to PJ_Init5 initially adjust document from ABBYY FineReader.

Macro PJ_init1: (Ctrl-Alt-Shift-1)
- Displays counts of various types of pictures, tables, textboxes etc. (calls macro ObjectCounts)
- Clears the Word setting to automatically compress pictures on save. This is done by feeding Word keystrokes. I could not get it to work reliably so it stops on the dialog box after clearing the checkbox. The user must accept the setting and cancel out of the dialog boxes.
- Removes section breaks so there is only one section to start with.
- Sets margins.

Macro PJ_Init2: (Ctrl-Alt-Shift-2)
- Removes all user created styles (calls macro remove_user_styles). This includes styles created by FineReader. The macro typically encounters errors due to FineReader creating styles that Word considers duplicates. The macro loops several times to sequentially handle these error cases. When it finishes, it displays a notice that tells whether all user created styles were successfully removed.

Macro PJ_Init3: (Ctrl-Alt-Shift-3)

- Attaches the PJ_Book template, applies the normal style to all text, and sets quickstyles.
- Sets the font theme.
- Replaces all tabs with a space.
- Removes any cases of bold fonts.
- Sets the Courier New 1 point font size set by FineReader for the text layer (section breaks, paragraph markers, etc.) to Times New Roman 11 point.
- Sets header and page number format--different on odd and even pages.
- Adds a paragraph marker before and after each manual page break (every page) and an additional one at the end of the doc. This allows textbox and picture anchors to not be on last paragraph of a page (which can make objects jump unpredictably).

Macro PJ_Init4: (Ctrl-Alt-Shift-4)
- Calls macro FineReaderFrames, which does the following:
  - Displays counts of various types of pictures, tables, textboxes, frames, etc. before processing is started. (calls macro ObjectCounts)
  - Converts frames with text to textboxes (because paragraph styles cannot be applied within frames). All readable text is in textboxes (none on text layer).
  - Converts pictures to normal pictures (shapes) in the drawing layer (the original pictures as inlineshapes in frames cannot be moved or replaced easily--both of which are needed).
  - Sets textboxes and pictures to be in front of text (otherwise wrapping makes paragraph markers with textbox and picture anchors tend to move off page and cause major problems).
  - Leaves tables in frames (because attempting to convert them to textboxes caused problems in the position and structure of the tables).
  - Aligns the textboxes, pictures, and frames relative to the margins and along the margins when appropriate (margins differ for odd and even pages).
  - Increases the contrast of pictures by 5%, which also converts pictures from grayscale to color and produces much better results when converted to pdf.
  - Makes the remaining frames with anchors (for textboxes and pictures) small and moves them to the left margin 3 inches from the top so they can be seen if needed. Removes any text from the remaining frames with anchors.
  - Displays counts of various types of pictures, tables, textboxes, frames, etc. after processing is completed. (calls macro ObjectCounts)

  NOTE: For reasons unknown, this macro sometimes deletes the contents of tables if the tables are not on the last page.

Macro PJ_Init5: (Ctrl-Alt-Shift-5)
- Does formatting for standard tables. It makes the lines adjust to content, sets paragraph properties (no first line indent, single space, left aligned, etc.), and first row set to bold. This is optional and would only be needed if there are some standard tables. Tables with special formatting may be better handled separately.

All of these macros and others mentioned below are available in a file at:
http://archivehistory.jeksite.com/download/download.htm

A file with notes and explanation about using Word (including styles, templates, macros, etc.) is also available at the download webpage noted above.

KEYBOARD SHORTCUTS * indicates JK custom (listed by frequency of use)

Ctrl-Alt-Shift-1 to 5 -- Run PJ_init macros to initialize ABBYY FineReader output*

Arrow -- with Picture or Textbox selected, move it 1 point in direction of arrow
Alt-Shift-Arrow -- move selected Frame, Picture or Textbox 4 points in direction of arrow* (Pictures may need nudge with simple arrow key initially)

Ctrl-2/3/4/6/7/8/9/0 -- Set space after paragraph in points. 7=add 6, 8=12*

```
Alt-2/3/4/6/7/8/9/0  -- Set space before paragraph in points. 7=add 6, 8=12*

Ctrl-Shift-5 (num)   -- Select paragraph*
Ctrl-Alt-Shift-Enter -- Replace all manual returns with one space (usually selected
                        area)*

Alt-Ctrl-1/2/3   -- Apply heading 1 style or 2 or 3
Alt-Ctrl-N       -- Apply normal style*  (Word default for this is Ctrl-Shift-N)

Ctrl-/           -- Show paragraph markers*
Ctrl-E           -- Center paragraph
Ctrl-L           -- Left justify paragraph
Ctrl-R           -- Right justify paragraph
Ctrl-J           -- Full justify paragraph

Ctrl-]/Ctrl-Shift-> -- Increase font 1 point
Ctrl-[/Ctrl-Shift-< -- Decrease font 1 point

Ctrl-Alt-Enter   -- Combine the next paragraph or manual line break*
Ctrl-Shift-Enter -- Change next paragraph mark to manual return (Shift-Enter)*

Ctrl-1           -- Single space paragraph
Ctrl-<           -- No indent for paragraph*

Ctrl-Shift-Down  -- Add next paragraph to selection
Ctrl-Shift-Up    -- Add prior paragraph to selection
Ctrl-Shift-.     -- Select to end of sentence inside the period. Can repeat.*

Alt-Down         -- Go to next object
Alt-Up           -- Go to prior object

Alt-Shift-Home/End  -- Go to previous/next section break*
Alt-Ctrl-PgUp/PgDn  -- Delete previous/next manual page break*
```