Creating PDF Documents to Upload to FamilySearch Memories

FamilySearch accepts documents in the PDF format with a maximum size of 15 Mbytes.

PDF stands for Portable Data Format and was developed by the Adobe Corporation using printer driver technology. Adobe originally provided a free reader (the Adobe Acrobat Reader) to view the files. PDF has since become an international standard. The church distributes documents using the "Searchable PDF Format".

If you have an electronic version of the document you wish to upload to FamilySearch, the best approach is to convert it to a PDF file directly (without scanning it). In Microsoft Word you can save a DOCX file to the PDF format by saving the file and selecting PDF in the "Save as type:" drop-down menu. In Libre Office Writer, the word processor on the Family History Center computers, a file can be converted to PDF through an export operation. This provides excellent quality and the smallest files size.

If you are starting with a printed copy, you can create a PDF format using the HP Scanjet G3110 flatbed scanner with HP Scanning (G3110) software.

Definitions

Image: a photograph of a document that does not recognize text within it.

OCR: Optical Character Recognition. A technology that recognizes text within a digital image. It is commonly used to recognize text in scanned documents and images.

PDF (filename.pdf) Portable Data Format. There are two types of PDF files: The first is an **image** file, which is a compressed photograph of the text. The second uses OCR software to locate text in the image. The second type is called a **Searchable PDF** file. With a Searchable PDF file, using the Adobe Acrobat Reader, you can search for text in the file and copy text and paste it into another document. But the file cannot be edited.

Text File (filename.txt) a plain text file without any formatting. Microsoft has a text editing program called Notepad that can be found at Start -> Window Accessories -> Notepad

RTF File (filename.rtf) stands for Rich Text Format. An RTF file contains limited formatting. Microsoft offers an rtf editor called WordPad which can be found at Start -> Window Accessories -> WordPad.

HP Scanning offers several options that can be selected from the scan shortcuts:

1. Document to PDF file - Document as an image

2. Quick Document to PDF - Document as an image. The scan is faster and the file is smaller but the quality of the image is worse than the previous option.

3. Document to Searchable PDF File - Searchable Text (OCR)

4. Text (OCR) to RTF File – Editable Text (OCR). An RTF file cannot be uploaded to FamilySearch, but it can be ported to a Word processor where it can be edited and saved as a PDF file.

5. Text (OCR) to WordPad – Editable Text (OCR). The RTF file is opened in WordPad where you can edit it and save it.

6. Document to TIF File (Tag Image File Format) an alternate to JPEG format that doesn't use any compression to reduce the file size. The consequence is the file is much larger than the JPEG equivalent. We recommend using JPEG.

The fourth option creates a saved RTF file while the fifth option opens the RTF file in Microsoft WordPad which you can edit and/or save.

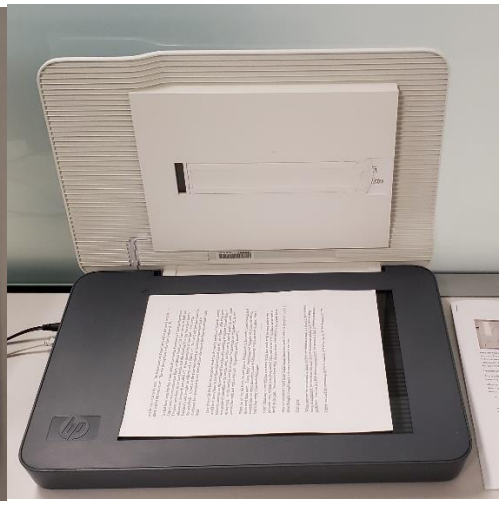
I created a file in Microsoft Word that filled a single 8.5 by 11 page with text to try all of these options and compare them. The results are in the table below. The file size is a function of the document features. Your results will be similar but not identical.

	Format	**File Size**	**Page Count**	**Comments**
	Original Word File, docx format	30K Bytes	NA	The original single page file
	Saved as a PDF file	98K Bytes	153	Saved from the word file
1	Document to PDF file (Image)	650K Bytes	23	Image Scan
2	Quick Document to PDF (Image)	279K Bytes	53	Smaller file, reduced quality
3	Doc to Searchable PDF (OCR)	780K Bytes	19	More useful, but larger file
4	Text (OCR) to RTF File	7K Bytes	NA	153 pages possible
6	Document to TIF File (Image)	8,129K Bytes	1	Excessive size

The first line represents the original Word docx file which was 30K Bytes in size. The second line is the result of saving the docx file as a pdf file rather than scanning it. This process was described at the beginning of this guide. This increased the document size to 98K Bytes. Given the FamilySearch file size limit of 15 Mbytes, we could potentially create a 153-page document in Word that would fit within the FamilySearch limit when saved as a PDF. In the table above, the maximum page count is reported in the 4th column within the 15Mbyte limit.

The number in the first column of the table corresponds to the number in the HP Scanning option listed above. The last five lines in the table list the scanning options from HP Scanning software. Options 1, 2, 3 and 6 create a document that could be uploaded. The PDF format uses compression to reduce the file size. The TIF format is uncompressed. While acceptable to FamilySearch for photos, the 15 Mbyte limit would restrict TIF files to a single page document which isn't very useful. With the second to last option, Text (OCR) to RTF File, the RTF could be loaded into a word processor, edited and saved directly as a PDF file, matching the 153-page estimate on the second line. This process is described later in this document.

Creating PDF files from Printed Documents using the HP Scanjet G3110 Flatbed Photo Scanner



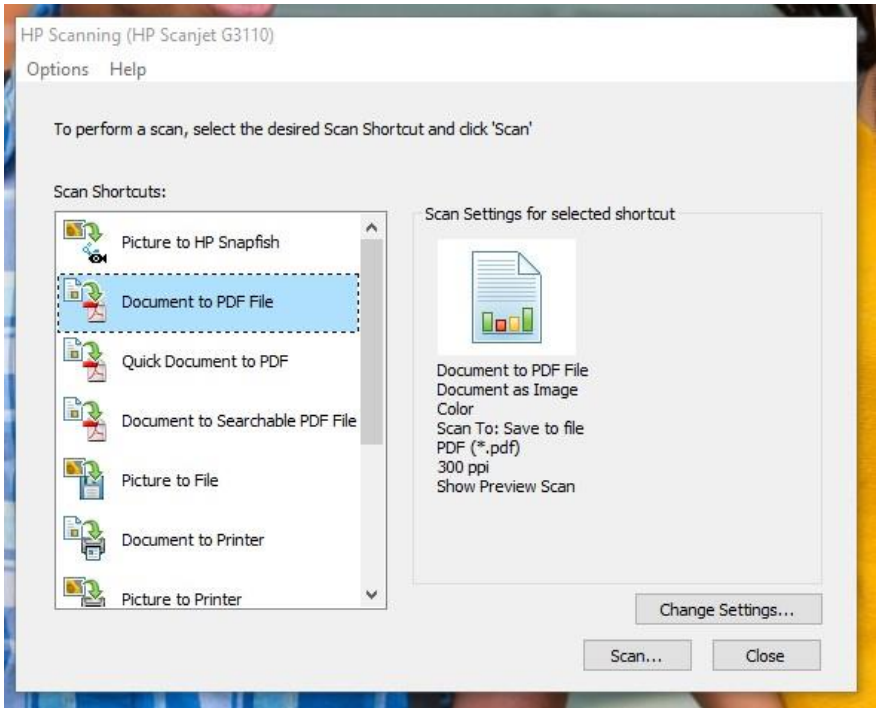
Here are two pictures of the HP Scanjet G3110 scanner. On the left, as it appears with the lid down and on the right with the lid up and with an 8 ½ x 11 paper on the scan glass ready for scanning.

To scan a document, place the first page on the scan bed glass with the top of the page near the HP logo and flush against a corner of the scan bed with the text to be scanned facing down. Then close the lid.

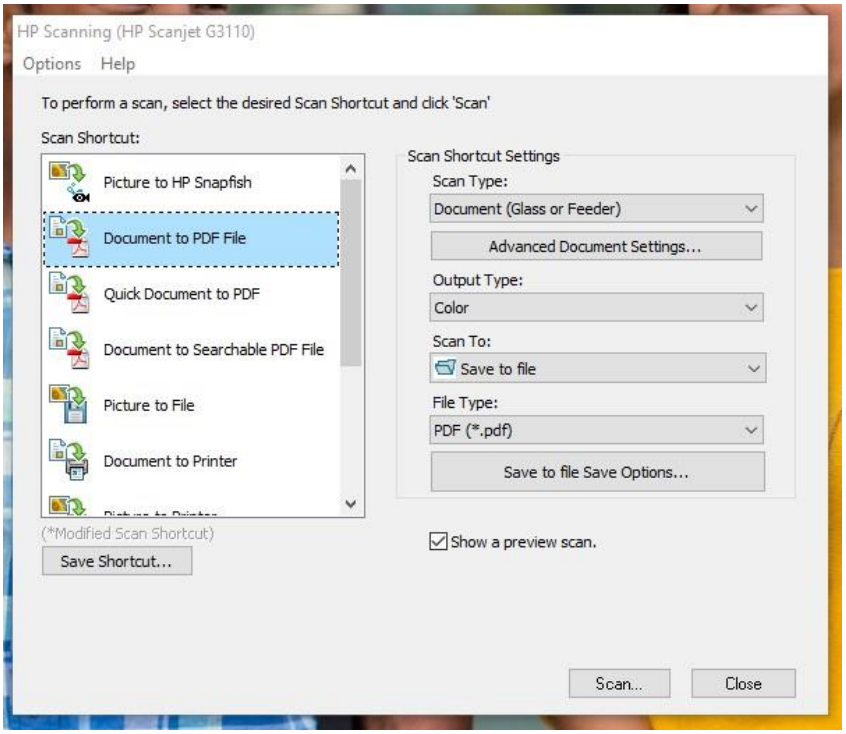
Start the software by clicking the HP Scanning (G3110) icon on the desktop of the computer. (Not the HP Copy (G3110) icon!)

After selecting the icon, the scanning window opens as shown on the next page.

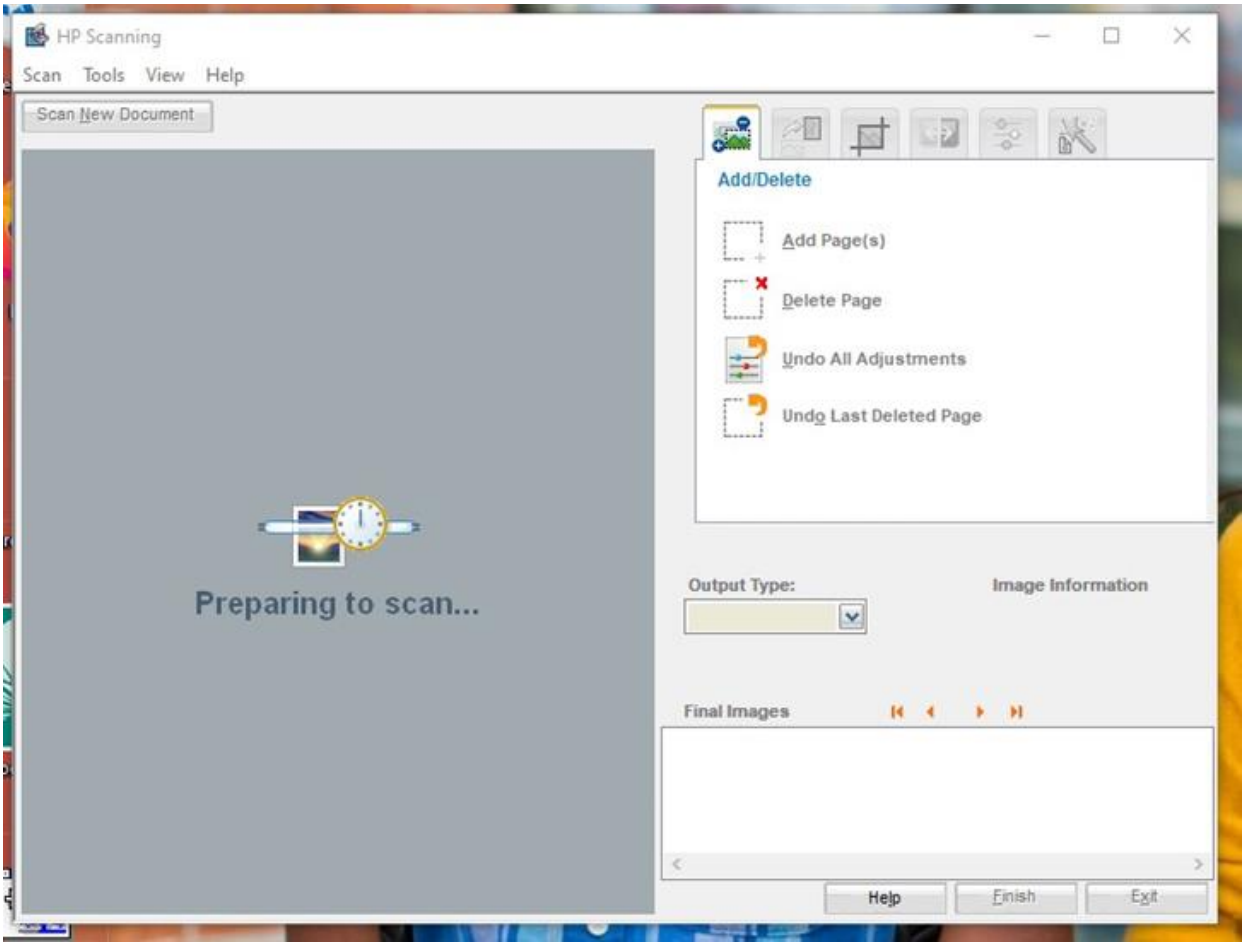
In the Scan Shortcuts list on the left of the window, scroll to find the “Document to PDF File” and select it as shown in the picture on the left. Alternately you could choose “Quick Document to PDF” or “Document to Searchable PDF File”. The process for these three options is the same. The differences in the results are discussed above.



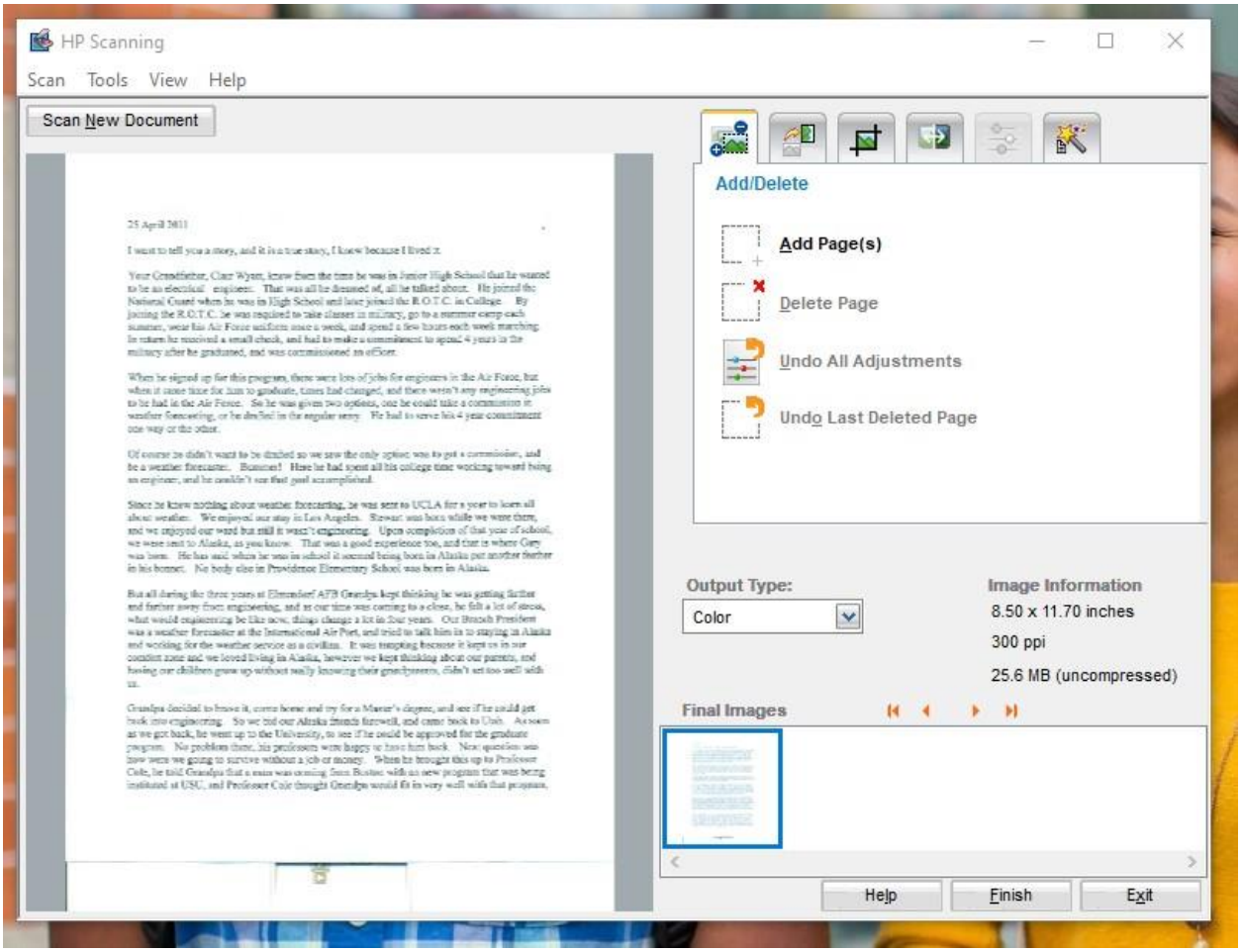
Then select the “Scan …” or “Change Settings…” button at the bottom right of the window. The “Scan Shortcut Settings” appears in the right of the window. For the “Output Type:” you have three options: Color, Grayscale or Black and White. Results may vary with the type and quality of the original with each selection. The Color setting often works well on documents that are mimeographed or where the original text isn’t dark black. The other defaults are fine.



Press the “Scan …” button again. The HP Scanning window opens. The window on the left provides a status message, “Preparing to scan…” There is a significant pause before the scanning starts. The scanner lamp turns on which is visible around the edges of the scanner lid. The motors that move the scanning bar start up.

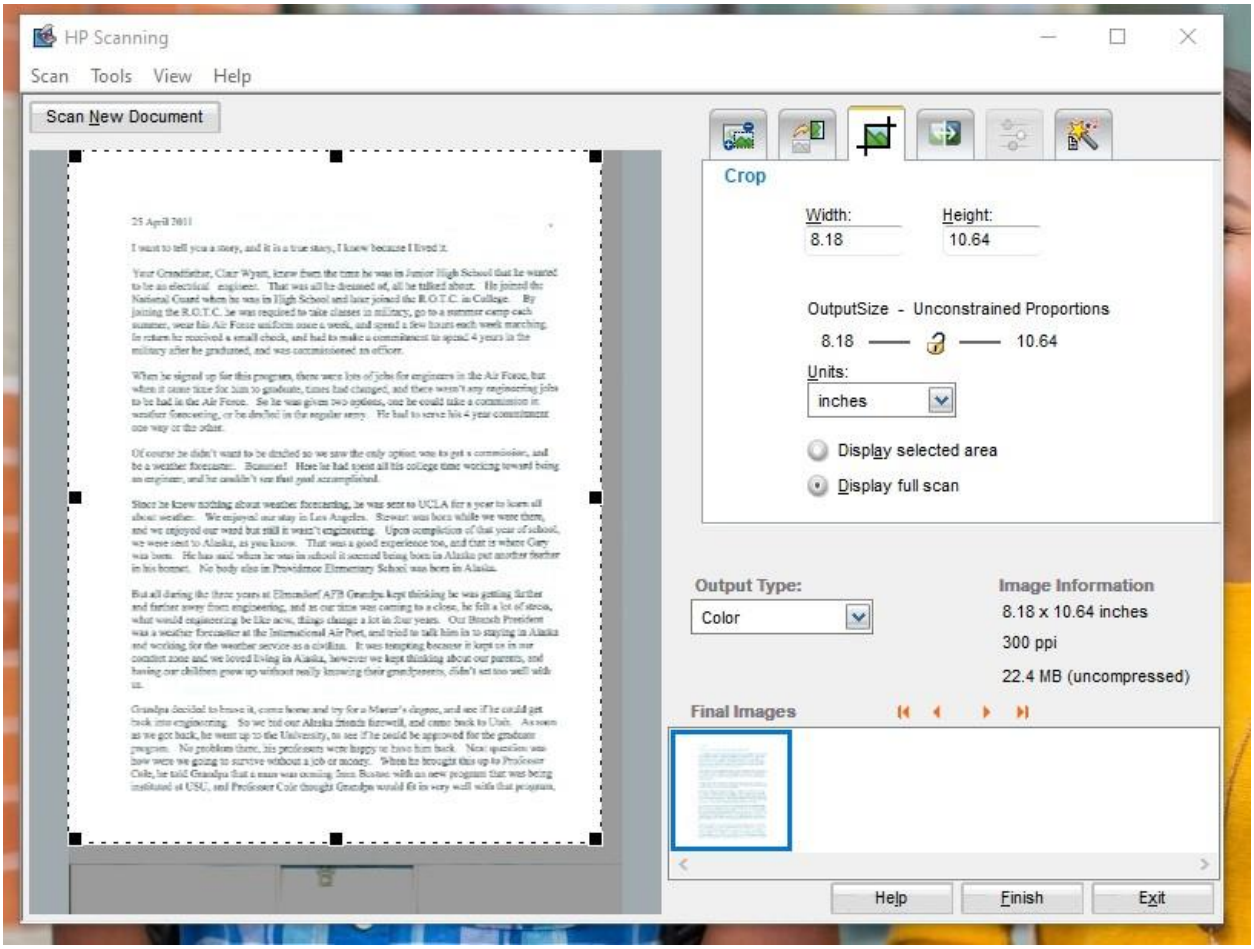


When the scanning completes the scanned image is displayed on the left of the HP Scanning Window. Since the scan bed is 9 x 12 inches, it is larger than the 8 ½ x 11 inch paper, the scan artifacts at the bottom of the image are from scanning the lid of the scannerbeyond the edge of the paper. Note that the window shows the most recently scanned image on the left and all of the scanned images as thumbnails at the bottom of the right side of the page.



There are 6 tabs in the top left window. The first tab was automatically selected when scanning was started and is labeled "**Add/Delete**". The remaining tabs provide editing options to improve the appearance of the images as you scan them.

Selecting the **second tab** opens the "**Rotate/Flip**" page with the options to "Rotate Left", "Rotate Right", or "Flip" (which mirror images the scan output.) Occasionally this can be useful. We have seen a case where the scanner rotated an image for no apparent reason. Such problems are easily repaired here.



The **third tab** is the "**Crop**" tab. This is really useful and, optimally, should be used after scanning every page. The scan image is outlined with black dashed lines and the small black squares are handles to grab with your mouse so you can set the boundaries of the scan image. Note that in this example, the edges of the image have been cropped to match the original size of the scanned paper which eliminated the scan artifacts at the bottom of the image. You can also match the margins on the right and left side of the text. The crop becomes final when you select a different tab.

The **fourth tab** is "**Lighten/Darken**" and allow you to adjust the highlights, shadows, midtones and gamma characteristics of the scanned image. Normally you don't have to make any adjustments here. If the original is a really bad copy, these adjustments may improve readability by making the text more visible.

The **fifth tab** is inaccessible during this sequence. (It is greyed out.)

The **sixth tab** is labeled "**Correct Document**". It enables two effects which it states are only evident in the final scan. The first is **Sharpen** which is described as "Focus edges to increase clarity in an image". The second is **Descreen** which is described as "Reduce undesirable patterns

in scans of printed items." These aren't normally needed in a text scan but can improve the quality of images embedded in a text document.

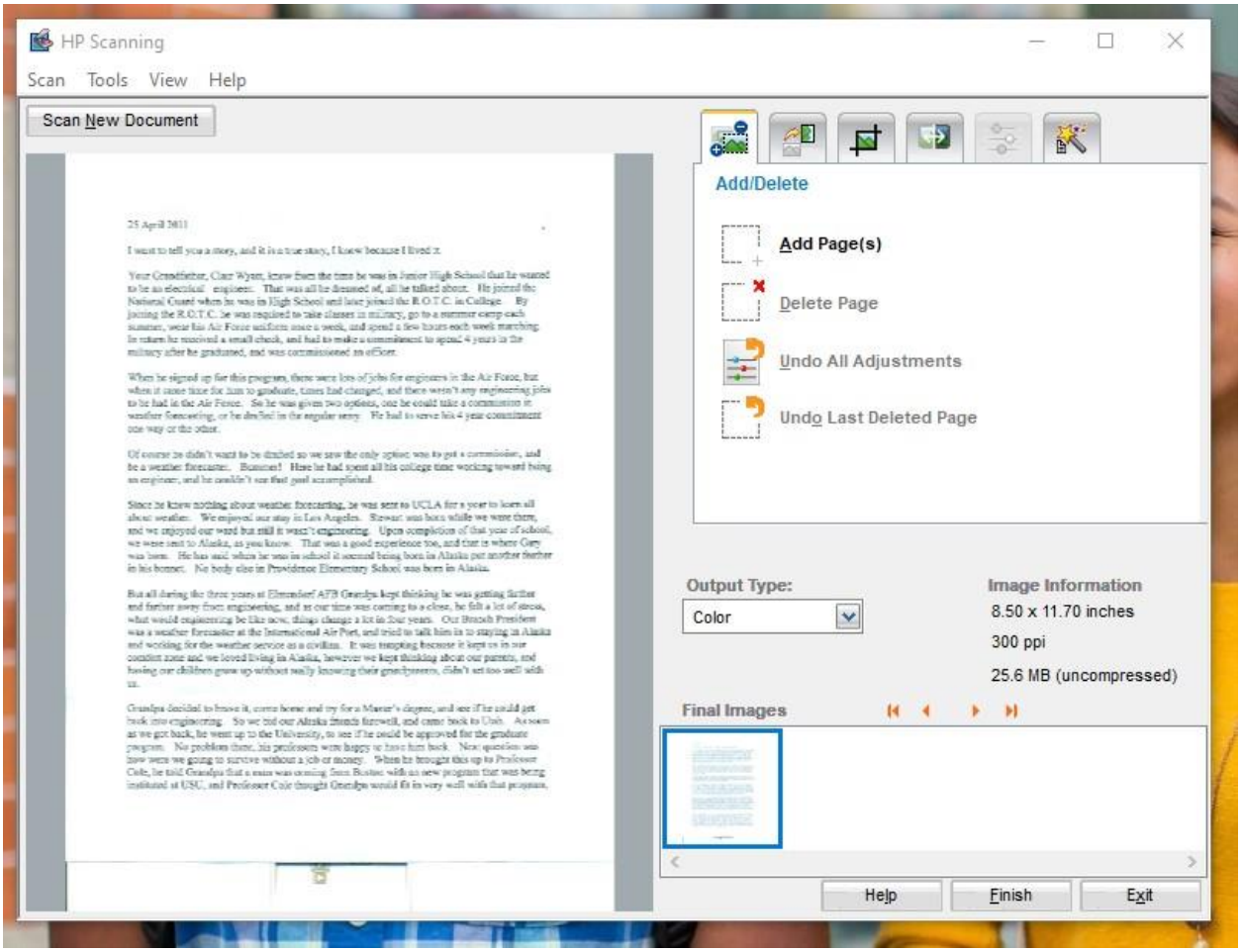
The first, Image sharpening, refers to an enhancement technique that highlights edges and fine details in an image. Image sharpening is widely used in printing and photographic industries for increasing the local contrast, sharpening the images and reducing blur.

The second, descreen, is used to remove Moiré-pattern artifacts when scanning halftoned printed images (found in books, newspapers and magazines).

Here is an example from the internet. The left picture exhibits Moiré-pattern artifacts and the right picture has been run through a descreen software algorithm.

If you are scanning a history that includes images from newspapers, this could be very helpful to get an acceptable looking image.

Once you have determined that your scanned image is acceptable as originally scanned or enhanced by these software options, return to the first tab, "Add/Delete". You have a number of options there. If you want to scan another page into the document you are creating, first replace the previously scanned page with the next page to scan on the glass, then select the "**Add Page(s**)" button. The next page will be scanned. If editing the page wasn't successful or you scanned the wrong page, choose "**Delete Page**" and the most recently scanned page will be deleted. If the adjustments didn't work out, select "**Undo All Adjustments**" to get back to the original scan. If you deleted the page by mistake, you can re-add it with "**Undo Last Deleted Page**". You can also trash your work and start over with the "**Scan New Document**" at the top left of the window. When you have scanned all the pages you need, select "**Finish**" at the bottom of the page.



Note that there is some image information displayed about the scan. While 25.6 MB is a huge file size, (way in excess of the 15 Mbyte limit at FamilySearch) the program will compress the file at completion and final pdf will be much smaller in size. Note the results in the table on page 2.

When you select “Finish”, the software creates a final document that includes all the pages you scanned, names it “scan0001” and saves it at C:\Users\Patron\Document\My Scans. It will automatically open Windows Explorer to display that folder. You move the file off the computer onto your flash drive, delete the original and rename your copy with a meaningful title. Then you are done!

Scanning a document with OCR capability to create an editable document.

The “Document to Searchable PDF File” uses OCR to create a PDF file that appears much like an image PDF file. The Searchable PDF option allows you to copy text from the document and paste it into a word processor file. However, you cannot edit or make any changes to the file.

To make an editable file use the “Text (OCR) to RTF File [Editable Text (OCR) Save to file Rich Text File (*.rtf)]” option which uses OCR to recover the text from a scanned image.

Why would you want to create an editable file from a printed document? There are a lot of good reasons. If you have a large document that you would like to upload to memories, recovering the text, opening it in a word processor and saving or exporting it to the PDF format would increase the number of pages within the FamilySearch limit by a factor of 3 to 7. (See the table earlier in this document.) If you have a printed document that you would like to update, correct, re-edit, combine with other material or reformat, then you can either manually retype the entire document into a word processor or you can scan and OCR it.

The process is to scan the file, with the OCR software creating the RTF file. Then open the RTF file from your word processor. There you can edit it as needed, save it in the word processor format, then save or export to PDF to upload to FamilySearch Memories as described earlier.

There are two challenges with this process. The first issue is scanning accuracy. The OCR software works quite well with high quality printed documents where the font is clear and the contrast between the symbols and the background is good.

The symbols used in English can be difficult to distinguish. This becomes more apparent with poorer quality documents such as a mimeograph document, a photocopy of a photocopy of a photocopy, a torn or dirty document or a typed document. A human reader with intelligence can discern the differences by context and read accurately without much effort. The OCR program doesn't have that capability. Some examples of challenges for the OCR software are distinguishing between a lower case "L", an upper case "I", the number "1" and the exclamation point "!". Another example is between an upper case "S" and the number "5" or between the letter "O" and number zero "0". There are many more.

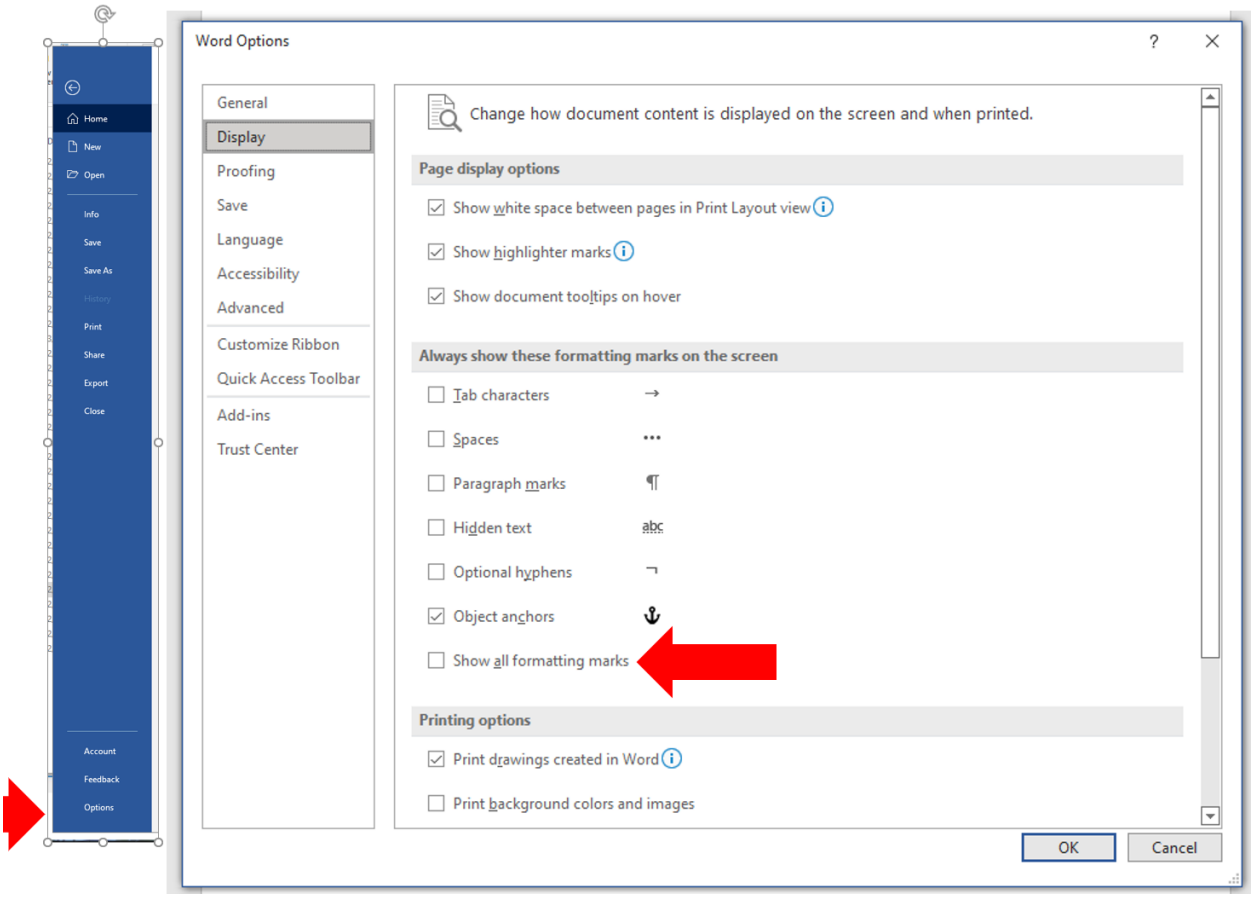
With a marginal document, the OCR software may produce occasional errors. When read into a word processor, these often result in a spelling error and are flagged by the word processor making them easy to find and correct. As the quality of the original document gets worse, the number of errors increase and may reach the point where it is easier to retype the original than to repair scanned rtf file.

The second issue is with line breaks. The RTF format uses line breaks to mark the end of a line of text. This is in line with the goal of providing an accurate copy of the original document with minimal formatting. The line breaks are a problem for a word processor like Word or LibreOffice Writer. The word processor will see the line break as a forced new line and will create some ugly formatting issues. What you would see when you view or print the document is a paragraph with "runt" lines where a new line starts at an unwanted place because of the new line characters inherited from the RTF file.

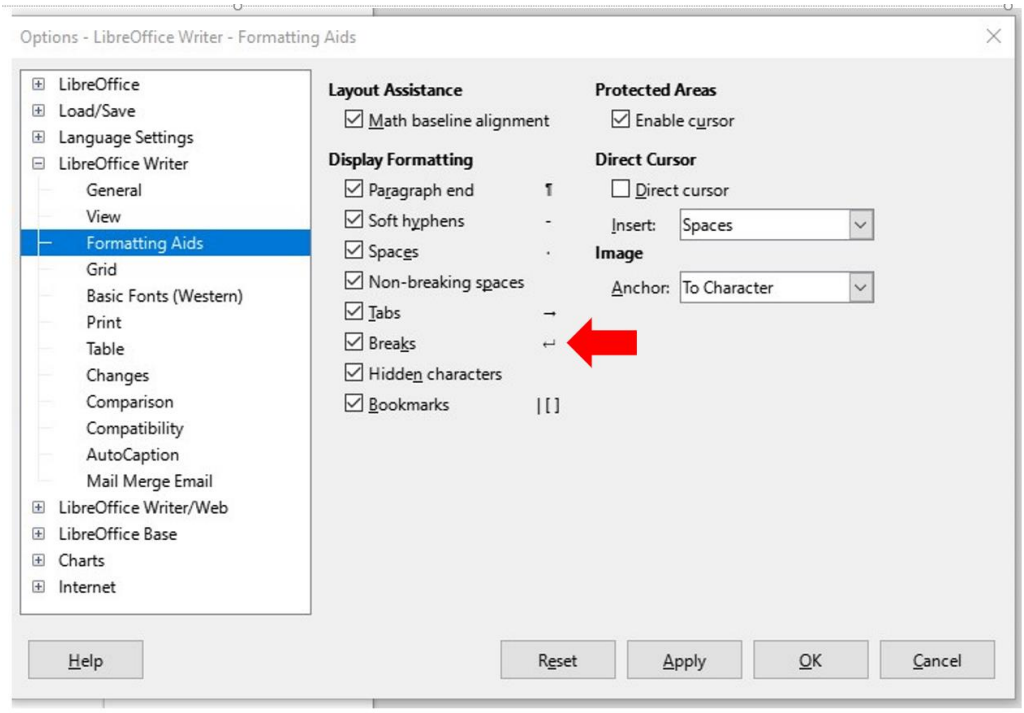
These new lines need to be removed carefully without removing the paragraph breaks. One approach is to save the RTF file as a TXT (text) file. This process removes all formatting including the line and paragraph breaks. Then re-add the paragraph breaks.

I like to the line breaks in a word processor. The line breaks are not visible, but like other characters, they can be deleted. An easier way to remove them is to enable the word processor to display formatting characters. Then they can easily be seen and deleted.

Here is how to do this in Microsoft Word: Choose the “File” tab (top left) to open the file menu. At the bottom of the menu select “Options”. The Word Options page opens with the “General” tab selected on the top left. Select the “Display” tab and check “Show all formatting marks” from the menu. (Once you are through editing, you need to uncheck this option!)



In LibreOffice Writer, select Tools->Options->LibreOffice Writer->Formatting Aids – the formatting Aids window opens:



Select Paragraph end, spaces and breaks. (When you are finished remember to uncheck these choices.)

In either case the paragraph break appears as a stylized backward "P", the spaces appear as a raised period and the line breaks appear as a left pointing arrow.

Using·a·program·called·Optical·Character·Recognition,·
which·is·↵
abbreviated·as·OCR,·you·can·scan·a·document·and·
create·a·text·file·↵
rather·than·an·image·file.·¶

This·capability·enables·you·to·recover·a·text·file·from·an·
image·file·that·↵
can·be·edited,·corrected·and·reformatted·to·meet·your·
needs.·¶
¶

Here is an example of an RTF file read into a word processor with formatting marks visible. The margins were reduced which created three runt lines where the line breaks were present in the scanned RTF file. These need to be removed.

Search through the document, removing all the line breaks (left pointing arrows) then save the file in the word processor format. Now you edit, update, reformat, combine, etc. with your electronic copy of the document!

You can save or export it to PDF and upload your document to FamilySearch memories.