The Old Shoebox Newsletter
Adventures in modern technology by Marlo E. Schuldt

Converting Printed Pages Into Text Documents

The goal of this newsletter is to inform you of a helpful software program that comes with most scanners. It's called Optical Character Recognition or OCR.

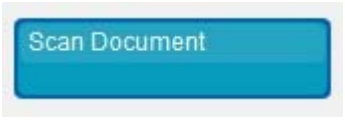
OCR scanning software makes it possible to convert a printed page into computer text that can be used with Word, WordPerfect or other text editing software. This makes it possible to reformat or use portions of a book, article or page in a family history.

You have two options when working with text:

1. Retype the page or book which could result in a lot of work.
2. Use an OCR program to scan and convert printed pages into text you could edit and use in your computer with other programs.

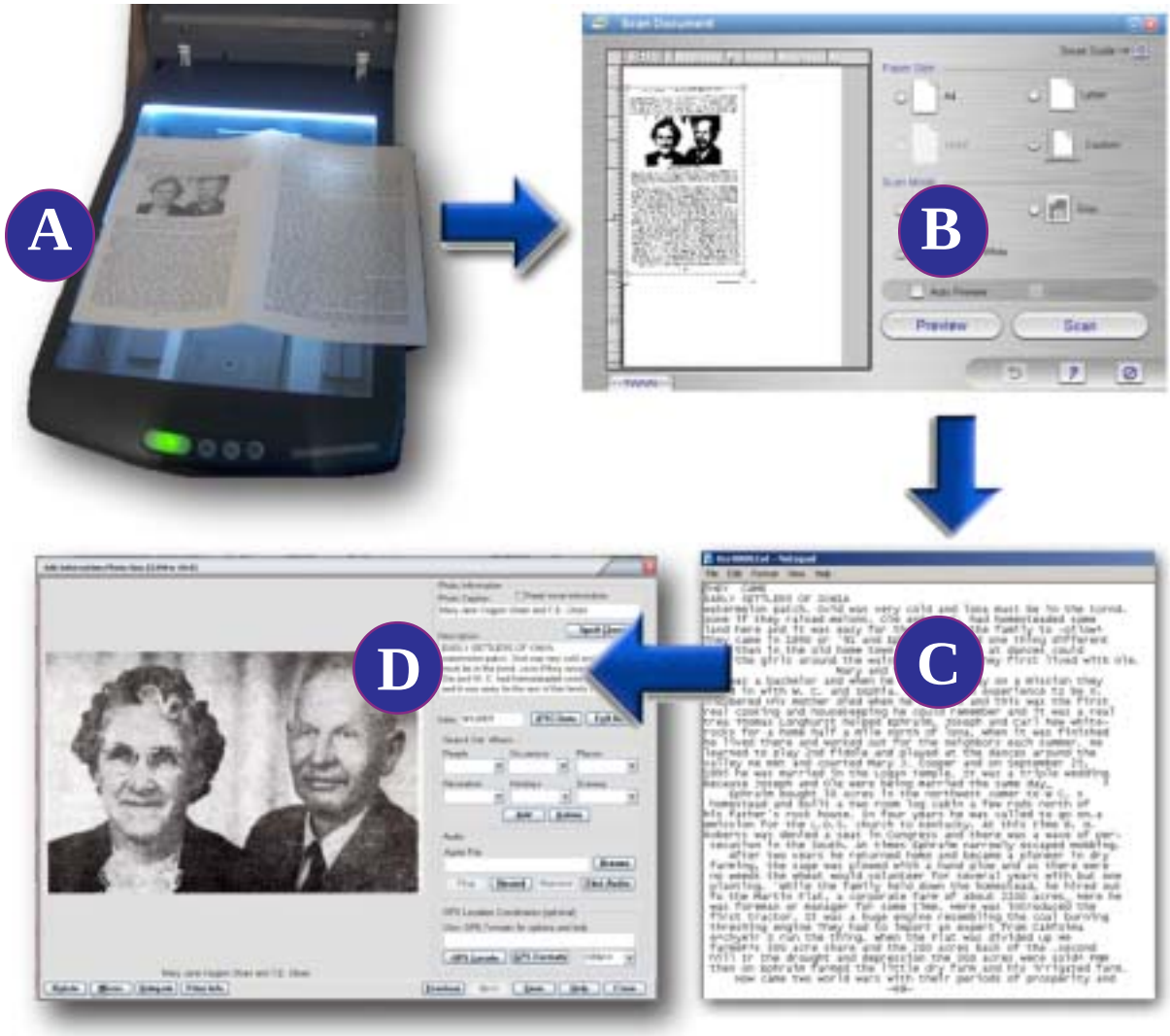
Using an OCR program may be very helpful and save you much time.

Click on the Epson OCR program icon in the Epson Smart Panel.



Click on the HP Solutions icon on your desktop. Next select the Scan Document button.

You will see the adjacent display. Click the Scan button located on the lower left of the display. *(Not shown in this screen capture).*



Here's How OCR Works

1. Place the page of text or book on a scanner. **(A)**
2. Start the OCR scanning program and scan the page. **(B)**
3. Save the page in the OCR program.
4. Let the OCR program "Recognize" or convert the scan into text. **(B)**
5. Save the converted text into a Word, WordPerfect or text file. **(C)**
6. Use the spellchecker in your word processing program to correct word recognition problems and spelling errors.

Use OCR Text With Heritage Collector

1. Copy text from the text file. Highlight the text to be copied by left clicking and dragging the cursor over the text.
2. Right click on the highlighted text and select copy.
3. Open Heritage Collector Pro. Right click on a thumbnail and select Edit Info.
4. Place cursor in Description box **(D)** and right click and select Paste. All the text will be copied into the Description box. You may need to scroll down to see all the text pasted into the Description field.

The example above is from the Epson OCR program. The HP OCR program is more automatic.

Don't get too excited and think your have found the panacea for all your old books, printed histories and other typed documents.

There are a few problems you need to know about before you start using OCR software.

1. **Poor quality of the original text**. Some old typewriters did not produce good, clean text. For example, it may be hard to tell the difference between the typed letters of an "e" and an "o." Sometimes this resulted when the ribbon was worn out. Other problems may have resulted when some of the keys became plugged or built up with old ink making the characters distorted such as a solid letter "o."

2. **Uneven printing**. Light or varied print quality will cause varied results since some of the text will be harder to be recognized by the OCR program.

3. **"Bleed Through."** Thin paper causes the images on the reverse side to show through onto the side you are scanning creating bogus words to appear in the margins and within the final text.

4. **Spelling Errors.** Unusual words and spelling errors in the original document may cause incorrect conversions.

5. **Smudges.** Blotches and other imperfections in copied documents or newspapers may be turned into erroneous letters and words.

Tips and Tricks

1. ***Darken Text.*** Light or poor quality text may be improved by taking the book or page of text to a copy store. Use darker settings so that the copy becomes darker which may improve the performance of the OCR.

2. ***Reduce the "Bleed Through."*** Place a black piece of paper (not glossy or shiny paper) on the back of the page being scanned. This will help absorb some of the black text or images showing through onto the page being scanned (OCR).

3. ***Experiment With OCR Settings***. Change the contrast and brightness settings in the OCR software. Scan one page and then evaluate the results. Hopefully you will be able to determine the best OCR settings for the text you are trying to convert. Unfortunately some text will not OCR with acceptable results. It may be faster and easier to retype.

4. ***Checking the Final Results***. Use the "find" option in your word processor to locate consistent errors.

   For example the OCR may mistakenly turn a capital "J" into a 3 (3im for Jim). Do a find or a search and replace (find 3 and replace with J). This will save you much time by letting the computer find consistent mistakes made by the OCR.

   Search other mistakes that consistently occur in the text document.

A Final Caution

An OCR program is essentially a glorified spellchecker. It makes “educated” guesses when recognizing and converting scanned images into words (text).

As we all know, a spellchecker can make some pretty funny and serious errors when it “checks” a document. An OCR program is no different and will make similar errors. Consequently you must always read or compare the original document with the converted OCR text. Get someone to help you compare the final document.

Copyrighted Material

Please remember the newsletters are copyrighted. If you would like to include a newsletter or excerpts in your newsletter or publication, please e-mail me for permission.