

Use OCR, instead of retyping documents



John Deans

paper and uses software logic to figure out which markings are what letters in the alphabet, numbers, punctuation and even symbols.

The end result should be an exact copy of the contents on the sheet of paper, but in an editable text format so you can modify and update it. This comes in very handy when you need to recreate

a large amount of digital text from a book, magazine, or just dozens if not hundreds of paper sheets.

I had a client couple of months ago that told me she was going to spend the weekend retyping an employee manual by hand from the only hardcopy she was able to find. She thought it was the only way to recreate the electronic version since the original MS Word file of the manual was lost years ago.

After I regained my composure with the thought of all that manual work, I then showed her how to utilize the OCR capabilities built into the scanner software she already had installed on her workstation.

Within a few minutes, she had it scanned in and OCR'd all that information into a new MS Word formatted employee manual.

This could have come in very handy a couple of weeks ago when the Congressional joint committee came up with a 1,100 page hardcopy document of the \$780 billion dollar stimulus bill. These plundering legislators saved the file in a PDF formatted document which only contained images of each page without any searchable text.

Odds are this was done on purpose to inhibit due diligence from reading, searching, and deciphering the largest spending bill in the history of the world.

Since they had to rush this unsearchable bill to President Obama's desk within 24 hours, hardly any of our congress people actually read the document. Lord, help us all.

If I were up there in D.C. and was delivered an 1,100-page PDF file with nothing but pictures of text, I would have immediately found a high speed scanner with a fast OCR engine and converted it back into plain text within an MS Word file.

But what do I know since I'm just a country computer consultant and not some Washington pinhead bent on bankrupting this nation.

OCR got its start in 1929 when Gustav Tauschek of Germany built a mechanical device utilizing templates and a photo-detector. When the character

would line up perfectly with the template the light would be detected by the photo-detector.

When computers started making a difference in the early 1950s companies like Standard Oil, *Reader's Digest* and Ohio Bell Telephone used OCR in unison with their large mainframe computers for converting text on paper to computer information.

The U.S. Post Office uses OCR in a big way to quickly scan and read the address to be delivered to several times faster than any human can.

The banks now are scanning in every check and deposit slip we write and hand over to the teller. Numerous times the Chase check deposit scanning and OCR system has caught my bad math when I am depositing checks.

As long as the original hardcopy is clean with crisp text in a standard font, most OCR software is around 99 percent accurate. Just to be on the safe side though I would read it over carefully and verify what was on the sheet of paper matches the digital text in the MS Word file.

This goes double if it is some sort of financial or legal document since you would not want a decimal in the wrong place or something even worse.

OCR can sometimes be confused

with the dynamic and real-time process of Intelligent Character Recognition or ICR. OCR converts a static marking on a paper to a digital character whereas ICR does a simultaneous translation of handwriting on a sensing surface like a tablet or even a smart phone.

My old Treo 700 had this ICR capability within its PalmOS software. I could use the stylus to scribble a name or number and it would be automatically converted to text for me on the fly.

I have a client that uses ICR on his tablet PC everyday instead of writing down his notes on a paper tablet.

OCR has become both affordable and highly accurate thereby enabling home users to implement digital character capture from older paper documents on their home computers. Canon makes a great \$60 scanner that has the OCR software with the installation CDROM.

Bottom line: Next time you start to type something from an existing hardcopy, stop and OCR it.

Next week's column: Moves, adds and changes.

John Deans of DeansConsulting.com is a Brenham area computer networking consultant who can be reached at 289-2233 or John@DeansConsulting.com for questions and comments.

Most every office I visit has some sort of optical scanning capability. The scanner could be part of a large copier/printer system from Sharp or Kyocera, a standalone flat bed scanner on someone's desk, or a scanner built into a personal copier/scanner/fax/printer device made by HP or Lexmark.

These scanners are usually used to obtain a digital image of a hardcopy picture. I use my Canon scanner once a week when I scan in this article after I cut it out of the paper

Depending on how its layed out in the paper, I have to scan it in at least two parts. Utilizing LViewPro I parse it together and crop out the other articles and white space.

From there I save it to a JPG image file, print out a hardcopy for my archives, and then save a PDF version for the DeansConsulting.com/columns.htm Web site. Finally I stick the original cutout article in an envelope to be mailed to my mother in Houston.

Once in a while I will use the scanner to acquire an editable MS Word document from text on a hardcopy. This process is called Optical Character Recognition or OCR for short.

OCR scans the characters on a sheet of