ABBYY OCR and pdf software

Scan in paper documents, create searchable pdfs.

I am astonished at how fast accurate and easy it is to scan in old paper documents with ABBYY Finereader. I am using revision 11.

I heard about ABBYY from Walt Jung, the famous author and engineer that used to work at Analog Devices. He is undertaking the conversion from paper to electronic format of his own work, as well as many articles from the past. While I was leery of saving documents in pdf, it turns out it is really the simplest way. If you try to make a Word or even an RTF document, you are only asking to spend a day a page rearranging things and suffering. You have to scan at 600dpi and then cut out the images and handle them separately.

I have found many colored magazine drawing love to be downsample as gif images with only 2 or 3 colors. They get more readable, and much smaller. So great, you can make some type of HTML file, but it is just not worth the work and it will never look like a printed page.

So I gave in to the forces of evil (Adobe) and just scan directly to a pdf inside ABBYY. It works great, and the page you get is an image of the page, with all the formulas and stuff perfect, and best yet, ABBYY puts the searchable OCR (optical character recognition) text "behind" the image of the text. So you can highlight it and copy and paste it into a plain-text or Word document or HTML page one day. Oh, use no Adobe products, ever, as they are Agents of Satan. Use Foxit free pdf viewer and their paid writer to edit pdf files.

I find 300dpi scans are OK, but 600dpi are best for OCR. What is astonishing is how small the files are. ABBYY lets you download a demo-- feel free to give it a try.

I use ABBYY in conjunction with my 8.5x11 Canon MF4890dw which will do color scans despite being a black-and-white printer. It has a sheet feeder that is remarkably reliable. For 11x17 I have a Brother MFC-J6710DW inkjet multifunction center. For really nice flatbed work, I have a Canon Canoscan 8800F flatbed USB printer. The other two are on the network, which is nice.
Here are some notes from Walt Jung, on his experimentation with ABBYY:

> 1) The first sample I sent was done page-by-page using a grayscale setting on the scanner, and combining JPG images in FR11. The problem with this approach is that the background must be lightened, and if you have a picture, it washes out. A workaround is to process a page as a text file, setting the background with contrast and/or gamma. Mixed results. The above is a giant pain, and takes hours. As noted, mixed results...
>
> 2) The second sample was done by scanning directly from FR11, and letting FR11 do the processing. It is much, much faster, and looks great. A downside is that you don't have the source page files as JPGs. There may be a workaround for this, but I'm not aware.

I agree that just letting Finereader do its thing has great results. If you really value the document, you can always run it through the scanner again to make a set of jpeg files. I do suspect that you can do a screen capture of a Finereader pdf and get whatever you need. What is astonishing is how much smaller the Finereader pdf files are, when they are, in essence, an image, with the OCR data hidden underneath.

Walt dropped another note about getting the most from ABBYY Finereader:

> 1) When you save a project file, it creates a long folder tree. In that tree each page of the document has a number/Image/data file (*.dat) structure. Among the latter is a "grayComponent.frdat" file. This is typically a huge file, like 16megs. Opening it with Iranfanview you get a msg that it is a renamed BMP file. It should open with any graphics editor, but I've only used two, both worked OK.
>
> 2) You may need to edit this file *directly*, if you need to cut and paste anything into the particular page. There does not appear to be any cut/paste ability with the built in FR11 editor. 3) Once you have cleaned up a particular page file, you can add it back into the set of pages as viewed in FR11. Just drop it onto any blank area, and it then loads at the end. You will then need to renumber it into the proper place, and delete the original page it is replacing.
> It seems as if there should be a more direct way to do all of this, but I haven't found it, as yet. I'm now getting excellent results of book chapters for my "Audio IC Op Amp Applications 3d Ed", but it has been a long and very hard struggle. I think I can see and end to it all, hopefully.

So now Walt has figured out how to do little edits on the pages before ABBYY does its pdf creation magic. He had more advice recently:

> I have been working on this project for some time, and it is frustrating for all of the setbacks. FR11 is a big help, but, ironically introduces almost as many problems/questions as it solves! I've mentioned the inability to directly edit the working page files, except via the workaround trick. There is a related limitation on the work files. They don't get backed up unless you set out to do this on your own, and you can easily lose them. I lost a set of edits on one chapter inadvertently because of this. Now I am zipping all of the work files as a belt-suspenders step (the entire DIR is also backed up to Skydrive).
>
> Glad that the editing suggestions helped you on your contract documents. But, the FR11 pgm does have a lot of hidden quirks. I just discovered the "Outline" option when generating a PDF. This really is a nice touch within the final PDF, as it offers a panel which highlights chapter/subsection headings at the left. But it seems haphazard in action. I've only been able to get *some* headings to be recognized. Sigh... another ticket to open up with ABBYY.

I should mention that one option you have is to send things out for scanning. I told Walt about onedollarscan, a company in San Jose that started out scanning any book for a dollar. That was a bit tough, so now they scan 188 pages for a dollar, and if the book is shorter than 188 pages you pay the dollar anyway.

> Yes, I do recall your mentioning that book scanning service. I may contact them yet. There are a couple of reasons why I haven't done this, so far. One is getting enough resolution into the final page files. 600dpi looks better than 300dpi. I have actually assembled all 200 pages of this book in a quick draft, it was around a 30 meg file. At 600 dpi, with an OCR layer. All of these things aren't std with one-dollar scan, but maybe they could be negotiated. Don't know.

When I was negotiating to scan all the EDN magazines from 1974 to 2000 with One Dollar Scan, I did get them to agree to scan at 600dpi, and to provide jpegs instead of searchable pdf files. They had to up the price to 4 dollars a magazine. Recently I just asked for cheapest and they offered 2 dollars a magazine. That would be about 1400 bucks.