OCRing Books: Rebuilding from Scratch

by Mordenkainen [March 2002]

This is a brief instruction manual to successfully OCRing books with reasonably complex layout/formatting and/or inlaid pictures. Please note that many other techniques exist and simpler books can usually be OCR'd with good results by directly outputting to PDF. Rebuilding from Scratch is a time (and patience) consuming way to OCR books. These are some of the things I've figured out and not an absolute bible to follow blindly.

Final Note: Although this is presented as a series of steps they don't need to be exactly as presented. You can scan batches of pages, work on them, OCR them and then scan another batch and repeat the process.

SOFTWARE

There are many software apps that do the same (or more) than the apps I'll be referring to in this document. I use Paint Shop Pro, ReadIRIS, Microsoft Word and JAWS PDF Producer.

Whenever you work with image files, you'll probably work with JPGs. You'll need to find a compression rate that gives good quality without making the file size too big. I usually prefer better quality than lower file size but beware if your book has lots of images. You can always save important pictures at a higher quality where the, what is called in the industry as, filler-art can be a little more sacrificed in terms of quality.

SCANNING

A general advice is to find the correct brightness/contrast. The only way to do this is by experimentation. What you want is being able to see the page, without having to adjust the monitor settings. Also, remember that it's always easier to fix if the image is brighter in the first place rather than brightening a dark image.

The covers are almost always in color so you should scan them at 150 dpi color unless you want to OCR any text in the cover (like author's name, etc.); if so then scan it at 300 dpi color but most of the times a straight image for the covers is alright. Don't forget the back cover. Since the scan will provide the source you'll want to align the book correctly to avoid rotation but what you really should avoid is slanting the page (for instance, by pressing too hard on the book in one end while scanning) since you can fix rotation later in the image program.

Now, for scanning the book per se, examine the pages. If you see any repetitive margin graphics present throughout the book you'll need to scan a page (or odd and even pages) for the express purpose of getting those margin graphics. Choose a page(s) that can provide the better source. This depends on the book. In some books, this is right at the beginning (or end) while others is smack in the middle of the book. Like in the covers, avoid rotation and slanting: remember, these graphics will be used throughout the whole book so they should look good. The dpi should be 150.

TIP: Don't forget web enhancements. These usually have the same look as the book and you can save yourself a lot of trouble by simply screen capturing the images from the web enhancement.

Don't concern yourself with the rest of the page, just concentrate on the margin graphics. Once done, open the images on Paint Shop Pro and carefully crop the relevant parts into individual files. If a book has a margin graphic that is present on top, bottom and side of the text you should split it into three files to avoid redundant image space.

With margin/repetitive graphics done you'll want to scan the book. Page-by-page, if the page just has pictures (other than repetitive graphics) then you should scan it at 300 dpi color (if the image is in color of course) or 300 dpi greyscale. Only scan in Black & White if the picture really only has 2 colors. When in doubt go with greyscale.

If the page only has text (aside from margin graphics, etc.) you'll want to scan at 300 dpi greyscale OR 300-600 dpi B&W (depends on background, font, etc. - experiment until you find the best way).

Okay, now you have all pages scanned according to whether they have pictures or not. With this in mind, load all the pages with pictures into Paint Shop Pro and carefully crop around the pictures, extract the picture and save it into a separate file. Depending on the book you might need to reduce the size of the image (see below).

OPTIONAL: Since you already have the repetitive graphics as separate files (right?) you can open all pages in your image program and remove the margin graphics from all pages. The benefit is when you feed the pages to your OCR program you can pretty much let it auto detect the page layout.

ROTATION/SLANTING: If any pages are rotated you can fix it with Paint Shop Pro. Most OCR programs can deal with up to 4 degree rotations (which are pretty severe) but you'll be wanting to keep it down to 0.5 degree rotation to avoid OCR problems later (especially if the book is formatted in double columns). Slanting or skewing is harder to fix and it's just better to rescan the relevant page(s).

OCRING

You'll need to do a bit of experimentation here (to see if 300 dpi greyscale works better than 600 dpi B&W for instance) to get optimal results. Feed batches of pages to the OCR program. If it supports learning USE IT! See if the auto detection system works correctly, fixing the problems that might come up. What you want is to keep the text/table boxes as close to the actual text/table as possible. Choose the .TXT output.

REBUILDING

A prime consideration is fonts. You should have the same fonts (if possible) or very similar ones. This will save you much grief when you're trying to make each page resemble the original scan.

TIP: If the book has web enhancement you can open it in Acrobat and (if without security settings) use the Text Touch-Up tool to see which font is used for each piece of text.

Start a Word (or whatever) document. You'll want to create a template to save you time and effort and also to reduce final file size. The template is composed of one or more pages that have the margin graphics set into position. You can also put text boxes with the page number to it change automatically. Set the margins as closely as possible to the original (use a ruler on the book). If the book uses different margin graphics for odd and even pages make sure you choose "different odd and even pages" in the header/footer options. Set columns, etc.

Now, enter header/footer mode. Insert the margin graphics and any page number, chapter number, etc. text boxes here. This way, each page will automatically use the same graphic lowering the file size and saving you the trouble of doing it.

Now when starting to build the book, two important things must be on the look out. If there's a page without the usual repetitive graphics (the front and back covers for instance, choose INSERT > BREAK > SECTION BREAK this will create a new page without the margin graphics.

Start inserting the text from the .TXT into the word document. Keep Paint Shop Pro opened in browse mode so you can quickly open each relevant page. When a page has a picture, insert it. Double-click and try to set the correct size. Make sure it's around 100%. If the image is bigger than needed open it in Paint Shop Pro and reduce it accordingly.

Keep formatting the text using the correct fonts/colors and with a similar line spacing (under paragraph properties).

TIP: Text Styles is your friend! For example, if the book divides the body text by headers using Verdana font at 16 pt size, Red color, Bold, create a style with this properties so in every header you can just select the appropriate text, click the "my_book_style1" and voilá.

For pages which some text (like tables) defy the general look you can use text boxes. Repeat until book is done.

PRODUCING THE PDF

I use JAWS because it always creates the lowest size files for me. You need to create a job by choosing the options. The one I use is like this: (if not stated, it's assumed all other options are cleared/disabled).

GENERAL: PDF file format: v1.3

Thumbnail: None (many people like this but for me it just adds to the file size without any real benefits - I print the books, I don't waste my eyes trying to read books on the computer screen. For this same reason I don't make bookmarks)

Resolution: 72 (this is only used for gradients. If you use a lot of them or large ones increase this to 150 to improve printing quality)

Advanced
Transfer Functions: Apply
Convert CMYK to RPG
Convert divide independent...

COMPRESSION
Color Images: Bicubic, 150, JPG low compression Greyscale: Bicubic, 150, JPG low compression
Monochrome: Subsample, 300, CCITT Group 4 Compress Text

FONT EMBEDDING
Never Embed: Tahoma, Times New Roman, Wingdings Embed all fonts (except base 14 fonts)
Subset fonts

COMMENTS

Finally .RAR the file! If you use JAWS there won't be much improvement but it hardly any work for you and can help a lot of people. Just as an extreme example, the Dark Sun Revised Campaign Setting takes 22.5mb as a .PDF but is only 14.5 when RAR’d!

Anyway, Practice makes Perfect. This technique demands time and hard work but it can ultimately achieve results rarely possible with direct to PDF OCRing. Just make sure the book you're working on really warrants rebuild from scratch.

If you have any other suggestions or just have a question about the above you can get in touch with me on DalNet's #RPGbookz or Nullus's #BW-RPG. Have fun and good luck on your projects.