How To Get Editable Text From A Book Or PDF Files of Images

This article is for webmasters and technical people who use the Linux operating system on their PCs. If you're not one of them, you might want to give this one a pass.

Some of the information here was previously in the beginning of The Present Antichrist By Rev. Fred J. Peters. I moved it from that article to here because most of my visitors are not technical people.

The main purpose of this website now is to find rare books, especially materials the Jesuits and Roman Catholic Church doesn't want to you read, books either in hard-copy or PDF files which are hard to read from a phone, and make them more accessible to the public by converting them to HTML format which makes them easy to read from the small screen of a phone. This is how I do it.

If I start out with a PDF file that I can't easily get the text from by copy and paste, I first create a directory for the PDF file, then I copy the PDF file into that directory, and then I open Terminal in that directory. After that I convert the PDF file into images with the following command:

```
pdftoppm -png name-of-file.pdf name-of-output
```

This will create in the directory multiple files with the name of the output file with -01.png, -02.png, -03.png etc. added to the name of the output file name you gave it. This is especially good when working with PDF file with multiple columns. It will give you images of single columns only.

After that I use Optical Character Recognition (OCR) software to create editable text files one by one using the "for in do" command in Terminal (the Linux command line) to run the OCR software (tesseract) on each PNG file. I used the following command on my Fedora Linux Desktop PC:

```
for i in img??.png; do tesseract "$i" "text-$i" -l eng; done;
```

This created files, text-img01.png.txt, text-img02.txt, text-img03.txt, etc.

Then I combined them all together with the following command which created the file present-ac.html

```
cat text-img*.png.txt >name-of-combined-file.html
```

And then using a text editor, I proofread present-ac.html and corrected all the things the OCR software didn't get right which was very little.

And then I used an online service to remove all the needless line breaks.

If I start out with a hard copy book, I scan the pages of the book with xsane (or any available scanner software in Linux) to create png files with numbers – img01.png, img02.png, img03.png, etc. – of each scan of the open two pages of the book. It's important to save these files in its own new directory. After that I run the above commands to extract text from the PNG files.