

Pages → Searchable PDF (→ archive.org)

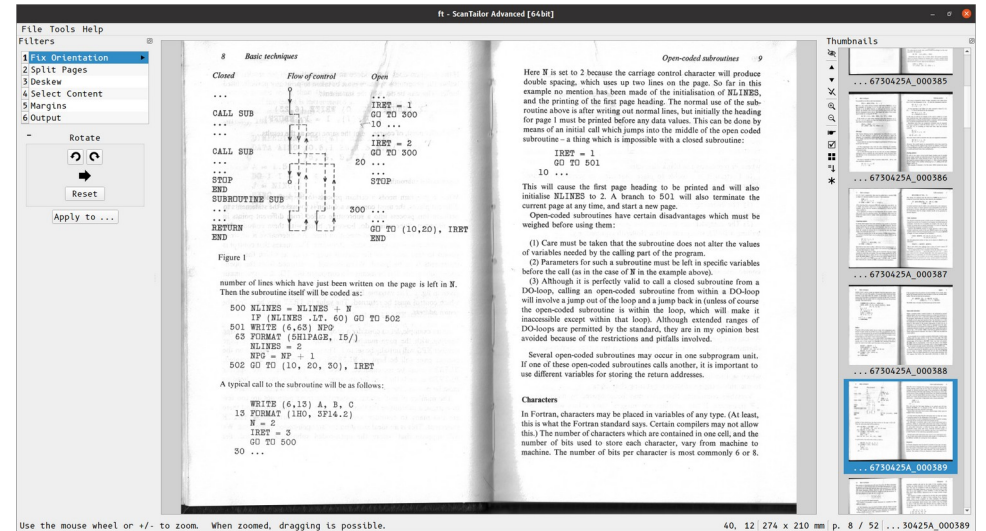
Stewart Russell
scruss@scruss.com

Background

- In April 2017, I spoke about *Scanned Paper You Can Find*
- Some tools got better since then
- If you have documents of interest, the Internet Archive is a useful place to archive them

ScanTailor Advanced

- ScanTailor original team moved on
- ScanTailor Advanced is maintained
- github.com/4lex4/scantailor-advanced
- You'll probably have to build it yourself



ocrmypdf

- Does exactly what you think it might do
- In most distros already
- Outputs (archival, open) PDF/A by default
- Might be worth looking at additional (lossy) optimization tools
- Uses all the cores!

```
ocrmypdf [-h] [-l LANGUAGES] [--image-dpi DPI]
          [--output-type {pdfa, pdf, pdfa-1, pdfa-2, pdfa-3}] [-c]
          [--sidecar [FILE]] [--version] [-j N] [-q] [-v [VERBOSE]]
          [--title TITLE] [--author AUTHOR] [--subject SUBJECT]
          [--keywords KEYWORDS] [-r] [--remove-background] [-d]
          [-i] [--unpaper-args UNPAPER_ARGS] [--oversample DPI]
          [--remove-vectors] [--threshold] [-f] [-s] [--redo-ocr]
          [--skip-big MPixels] [-O {0,1,2,3}] [--jpeg-quality Q]
          [--png-quality Q] [--jbig2-lossy] [--pages PAGES]
          [--max-image-mpixels MPixels]
          [--pdf-renderer {auto, hocr, sandwich}]
          [--rotate-pages-threshold CONFIDENCE]
          [--pdfa-image-compression {auto, jpeg, lossless}]
          [--fast-web-view MEGABYTES] [--plugin PLUGINS] [-k]
          [--tesseract-config CFG] [--tesseract-pagesegmode PSM]
          [--tesseract-oem MODE] [--tesseract-timeout SECONDS]
          [--user-words FILE] [--user-patterns FILE]
          input_pdf_or_image output_pdf
```

Same tools, slightly sharpened

- **tesseract** – still the only worthwhile free OCR package.
Still gets confused by tables sometimes
- **img2pdf** – *lots of images* → *one PDF*
Some options have changed since 2017
- **poppler-utils** – simple, solid tools like:
 - **pdfseparate, pdfunite** – split/join PDF pages
 - **pdftotext** – extract all the text from a PDF
 - **pdfimages** – extract embedded images from a PDF

What to upload to archive.org

- **The official line:**

Please contribute books, audio, and video files that you have the right to share. The Internet Archive, a non-profit library, will provide free storage and access to them. We reserve the right to remove any submitted material.

- **My approximate guide:**

- Is it out of print (and unlikely to come back in print)?
- Might someone else find it useful?

- **<https://archive.org/create/>**

Upload a book scan

- You can upload a zip file of scanned pages:

fortran-techniques-day_images.zip:

fortran-techniques-day001.jpg

fortran-techniques-day002.jpg

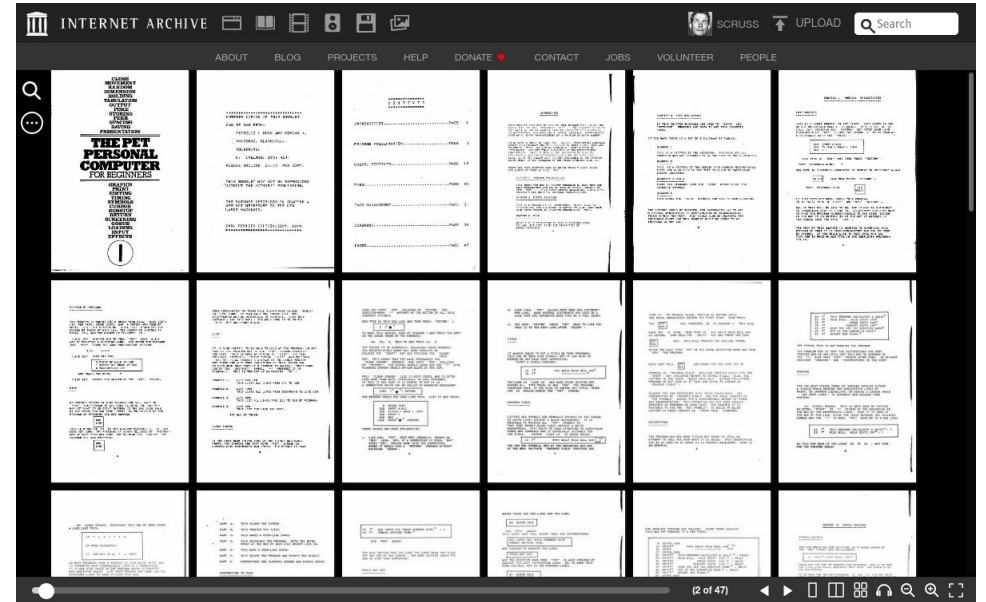
fortran-techniques-day003.png


...




- If it has a LoC *MARC* record, upload that too
- The more metadata, the more people will find it
- Internet Archive's magic, *almost-entirely-reliable* robots will do the rest ...


One I uploaded earlier

- It was donated by a GTALUG member
- While print quality is terrible (ALL CAPS DOT MATRIX), it's a great example of the local tutorial booklets that got people started with microcomputers
- archive.org/details/ThePetPersonalComputerForBeginnersBook1



 The Pet Personal Computer For Beginners Book 1
by Petifolio

 Edit Publication date 1979-09

If you have a lot to upload

- Internet Archive's **ia** Python tool can be scripted:

```
ia upload tbl105kRnd tbl105kRnd_images.zip tbl105kRnd_marc.xml  
--metadata='mediatype:texts' -metadata='lang:English'  
--metadata='subject:Sampling (Statistics)' -metadata='date:1949'  
--metadata='creator:Interstate Commerce Commission'  
--metadata='description:<p>A statistical table of 105,000 random  
decimal digits published in 1949. The data were compiled by  
processing numeric fields from interstate trade waybill punched  
cards.<br />Reference for method used:<br />Horton, H. Burke,  
and R. Tynes Smith III. "A direct method for producing random digits  
in any number system." The Annals of Mathematical Statistics  
(1949): 82-90.</p>'
```