Scanning and Saving a Collection of Documents
The Joys and Woes of Details

John C. Nash

nashjc@ncf.ca

2016-3-3

The Task

- A career of about 50 years has left many reports, documents, souvenirs, letters that it would be good to keep, but which fill too much space.
- Goal is to scan or otherwise obtain in machine-readable form all or most of the documents.
- Nice to have EDITABLE text, but images will suffice
    - Many documents have hand-written annotations
    - Many have images or drawings
    - Want all visible enough on a “normal” computer screen
    - Ideally in appropriate orientation, but ...
- Prefer to have efficient digital "size". Match quality to visibility, though it is easiest to use high-resolution and large files

Document properties

- Wide variations in intensity -- print, pencil, fountain pen, ball pen, coloured ink, faded photocopy, spirit duplicator, mimeograph, old typewriter ribbon, carbon copy
- Sometimes mixed in single “document” or ensemble that should be together
- Some colour, some greyscale, most B&W lineart
- Mixed orientations, often crooked thanks to sloppy photocopy
- Many reprints of journal articles, as well as tech reports. Some items are from my own work
- Many paper types and textures, some torn down wide listing paper, onion skin, A4, A3, letter, legal, ... Also some coloured backgrounds.

Hardware and Infrastructure

- Access at uOttawa to online journals
- Access at Telfer / uOttawa to Xerox WorkCenter document station with 2-sided document feeder that scans to email
- Samsung SCX 4521F multi-function printer
- CanoScan LIDE 60 flatbed scanner
- HP Deskjet 2132 inkjet printer with flatbed scanner
- IPEVO Ziggi HD document camera

E-journals at uOttawa

- Sometimes flakey when network busy or other issues (proxy sometimes seems to lock up)
- Downloads have different styles depending on service:
  - 2347021.pdf (JSTOR)
  - 1962_-_W_Spendley_-_SequentialApplicationofSimplexDesignsinOptimisatio[retrieved_2016-02-09].pdf (SIAM)
  - out.pdf (PROQUEST) PAINFUL!

Xerox WorkCentre

- Double sided document feed etc.
- When it works – superbly readable, small documents
- BUT
  - Many jams
  - “Resume” function seems to work after jam, but missing material or failure
  - “Forgets” email address
  - Sometimes decides not to send a document
  - Sometimes sends an unreadable document
  - Overall had to abandon this approach (but it used to work well)

Samsung SCX 4521F

- Purchased (2006) because it had a Linux driver

  .... that Samsung “withdrew” in a few months
- Messy to install scan function in Linux
- 1 sided doc feeder – works with Irfanview/Win7
- Jammed after about 6 documents, and jam so bad it broke a sensor arm that seems unfixable, though flatbed scanning possible in Windows with Irfanview. (Printing still OK)

CanoScan LIDE 60

- $99 purchase in 2006 to scan family album in Alberta that owner was reluctant to lend
- Formerly NOT in Linux compatibility list, but I've never had difficulties. Uses USB power.
- Reliable, if a little slow.
- Too small for legal paper.

HP Deskjet 2132

- Purchased in Florida when needed “cheap” printer (Walmart $39 US)
- Has flatbed scanner. Uses external power
- USB interface. Plays well with Linux

IPEVO Ziggi HD Document Camera

Ziggi Cam

- $150 CDN delivered ($99 US, but shipping extra)
- CCC21 bought one 2013, so useful I got one for myself in 2014 after move to Stittsville
- 2592 x 1944 pixels max – the WHY!
- Windows s/w only, but Cheese worked well with Linux Mint Maya in 2013/14
- BUT ...

Ziggi Downside

- UVC interface not standard – have to manually push focus ... CAREFULLY!
- Registration can be a nuisance. Need to use masking tape or jig to maintain straight image.
- Pantograph can be moved too easily
- Cheese or its infrastructure not stable over time

Ziggi / Cheese

Notes on Ziggi Cam with Cheese

Host	OS	Kernel	Cheese	Notes
EEE1225B	Mint Maya 13	3.2.0-77 #114	3.4.1	Keeps resolution and scanner settings, Spacebar fires camera Focus manually once in a while
UL30A	Mint Rafaela 17.2	3.16.0-38 #52	3.10.2	Loses resolution and scanner settings Different UI from 3.4.1 Change resolutions causes "There was an error playing video from webcam" Cannot recover except by closing. Got log msg "Streaming task paused; reason not negotiated (-4)"
UL30A	LMDE 2 Betsy	3.16.0-4-586 debian 8.3	3.14.1	Loses resolution and scanner settings Different UI from 3.4.1 Spacebar takes picture only 1st time. Seems to ignore dconf-settings (when trying to keep resolution and scanner)
UL30A	Bunsen Labs Hydrogen (Was #!)	3.16.0-4-amd64 debian 8	3.14.1	Loses resolution and scanner settings Different UI from 3.4.1 Countdown camera option keeps spacebar functioning
J6	Mint Rafaela 17.2	3.13.0-37-generic #64	3.10.2	Loses resolution and scanner settings Different UI from 3.4.1 Change resolutions causes "There was an error playing video from webcam" Cannot recover except by closing. But restart worked, with 1st time spacebar.

Software

Windows: (Used only to test hardware on Ziggi and Samsung)

- Irfanview: Actually useful enough that I run under WINE. Appears USB Webcam should be accessible under VirtualBox, but see next item.
- Ziggi IPEVO Presenter. Quite useful. Does not work under VirtualBox XP or Win7(32) (different Presenter versions)

Software: Linux

- ***cheese*** Webcam
- ***xsane*** scanning tool (e.g., sane-tools package)
- Other webcam s/w available, but I've not found so useful
- Potrace / ***mkbitmap*** – to clean up bitmaps
- ImageMagick – especially ***convert***
- My scripts – to automate scan and convert
- ***pdfshuffler*** – reorder, combine and reorient
- ***pdfjam*** – join, flip, 2-up

Rotation and reordering

- Start with three “pages” as jpg files
  - 1.jpg, 2.jpg, and 3.jpg
- Put together as pdf

  convert 1.jpg 2.jpg 3.jpg 123.pdf
- Use pdfshuffler and reorder and rotate
  - DEMO
- BUT ... different sizes

Resizing

- Can do this with *convert*

  convert 123.pdf -resize 1600x1200 123a.pdf

- Some issue to decide the sizing dimensions
- Can sometimes shrink the pdf file size, but it will often not work

Cleanup of image

```
convert 3.jpg 3.pnm
mkbitmap -x -f 20 -t 0.05 -o 3a.pnm 3.pnm
mkbitmap -x -f 20 -t 0.25 -o 3b.pnm 3.pnm
mkbitmap -x -f 20 -t 0.45 -o 3c.pnm 3.pnm
mkbitmap -x -f 10 -t 0.05 -o 3d.pnm 3.pnm
mkbitmap -x -f 10 -t 0.25 -o 3e.pnm 3.pnm
mkbitmap -x -f 10 -t 0.45 -o 3f.pnm 3.pnm
convert 3a.pnm 3a.pdf
convert 3b.pnm 3b.pdf
convert 3c.pnm 3c.pdf
convert 3d.pnm 3d.pdf
convert 3e.pnm 3e.pdf
convert 3f.pnm 3f.pdf
convert 3a.pnm 3a.jpg
convert 3b.pnm 3b.jpg
convert 3c.pnm 3c.jpg
convert 3d.pnm 3d.jpg
convert 3e.pnm 3e.jpg
convert 3f.pnm 3f.jpg
```

Scripts

- Use bash scripts with ***scanimage*** and ***mkbitmap*** and ***convert*** and ***rename*** to capture, combine and convert to pdf
- Works very well with flatbed scanners (HP and Canon) and is relatively fast and gives good registration
- Variants to allow for colour, faint print, differing levels of cleanup

Personal conclusions

- Document feeders wasted a lot of time
  - Too many paper jams for the type of documents I have, or feeders are just too fragile
  - Complicated by having to use software over which I had no control (Samsung and Xerox)
- Ziggi Cam with Linux Mint Maya (Ubuntu 12.04 base) and ***cheese*** good for awkward items where registration not too critical.
  - Fast
  - Experimenting with ***fswebcam*** for scripting --> pdf

Personal conclusions 2.

- Scripted scan with flatbed scanner works reliably if a bit slowly
  - Script choice by monochrome or colour, with possible *mkbitmap* cleanup
  - *Pdfshuffler* to rotate if needed
  - Occasional resizing with convert if needed
  - Shrinkpdf.sh attempt to reduce filesize

Thanks

Questions or Comments?