DIGTIZATION OF BOOKS USING FOSS TOOLS IN FEDORA

A. Mani

Member, Calcutta Mathematical Society
9/1B, Jatin Bagchi Road
Kolkata-700029 India
E-Mail:a.mani.cms@gmail.com
Homepage: http://www.logicamani.co.cc

FEDORA RELEASE EVENT 2011, GCECT, KOLKATA

ABSTRACT

In this presentation I will talk about building your own state of the art digitization systems using open hardware designs, capturing images, post processing for OCR and OCR. Aspects of open content licenses will also be covered.

Contents

1. Introduction
2. OPEN CONTENT LICENSES
3. EBOOK FORMATS
4. OPEN HARDWARE
5. POST PROCESSING IDEAS
6. SCAN TAILOR
7. SIGIL
8. OTHER FORMATS
9. OCR
10. REFERENCES

MYSELF

- Research in Mathematics (Algebra, Rough Sets and Logic)
- Some Independent Consultancy in KDD, Statistics and Specifications
- FOSS Activism: Ilug-Cal.Info, Fedora, LQ, No.440415,...

WHY DIGITIZATION?

- Access to eBooks and CCSA-licensed works
- Revolution in Education
- Kindle, Nook, iPad - too problematic for any education model.
- Already over 2 million free e-books are available
- A large work force needs to step in.

WHY DIGITIZATION? (Continued)

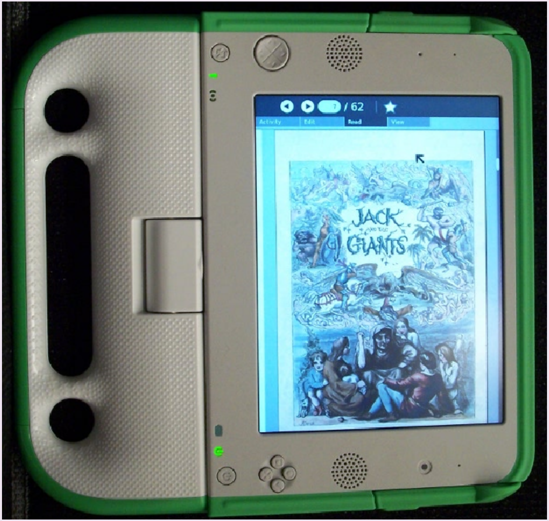
- Very few free e-books for beginning readers
- Many that exist are badly dated
- Advanced Academic Works - Out of Print
- Advanced Academic Works - Copyright & Proprietary Publication Regime
- Most of the free e-books are in English

WHY DIGITIZATION? (Continued)

- The General Publishers Copyright Regime
- Publishers Copyright ⇒ Low Readership
- Low Readership ⇒ Low Quality of Life
- Publishers Copyright ⇒ Out of Print
- Publishers Copyright ⇒ 'Knowledge is a Product'
- Digitization Can Partly Rectify This Scenario

WHY DIGITIZATION? (Continued)

XO Laptop with Screen Folded:



OPEN CONTENT LICENSES

- Content should be Freely Available
- Modifiable Subject to Attribution
- Redistribution
- Costs
- Clear Terms of Distribution
- No Publisher Copyright

COPY-LEFT LICENSES

- A License that is Free as in Freedom
- Requires Derivative Works to be Free
- Requires Derivative Works to be under the same License.

COPY-LEFT LICENSES

- GNU-GPL
- Share Alike?
- GFDL: Technical measures to obstruct further copying of your copies.
- GFDL: Printing
- Open Content License
- Open Publishing License

OPEN LICENSES

- Creative Commons Licenses
- Design Science License
- Free Art license
- FreeBSD Documentation License
- Open Content License
- Open Publication License

CREATIVE COMMONS LICENSES-3.0

- CC BY
- SA : Share Alike (Allow Modification only if user shares it)
- NC: Non-Commercial
- ND: No Derivative Works
- Compulsory: Attribution
- Optional: Restrictions or Waivers

FORMATS

- Commercial e-book reader formats - proprietary, Printing?
- Free eBook Formats - Many Formats, Many Purposes
- Plain Text, Rich Text
- PDF, Image-Container PDF, DjVu, PS
- CBZ, CBR
- EPUB - XHTML + XML Based

CHOICE OF FORMATS

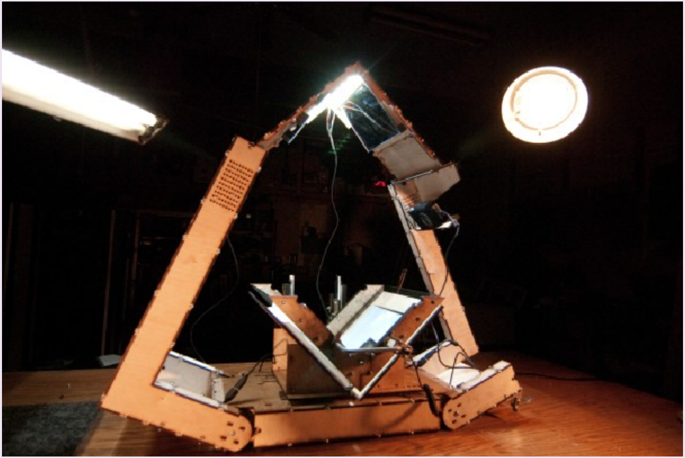
- Screen Reading + Printing (Normal Readers)
- Readers with Visual Disabilities: Plain Text
- Searchability
- Screen Reading Only
- Editability
- Size

BASICS

- Flat Bed Scanner: Not Usable
- Scanners
- Two Digital Cameras vs One
- V-shaped Cradle angled at 90 Degree
- http://diybookscanner.org
- http://www.atiz.com/: Commercial, Expensive, Closed

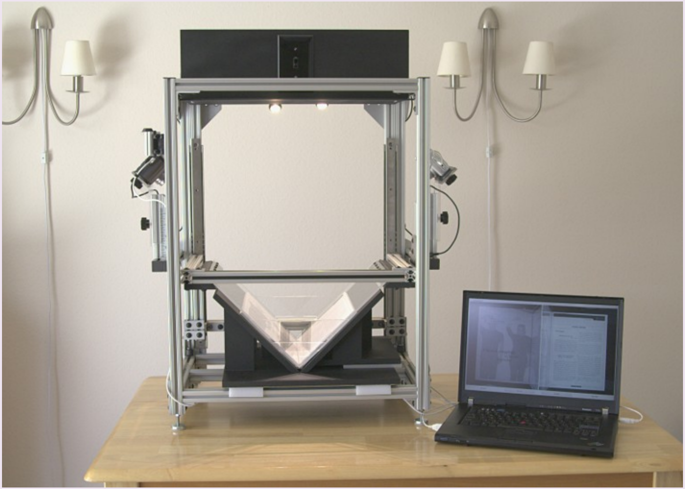
EXAMPLE-1

Open Hardware-1



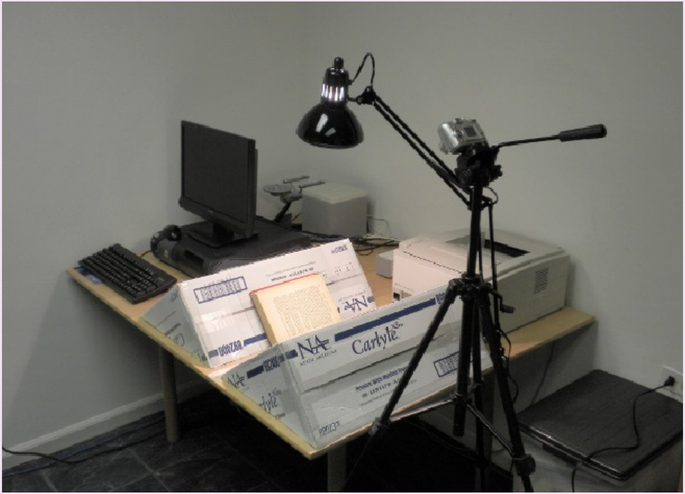
EXAMPLE-2

Open Hardware-2



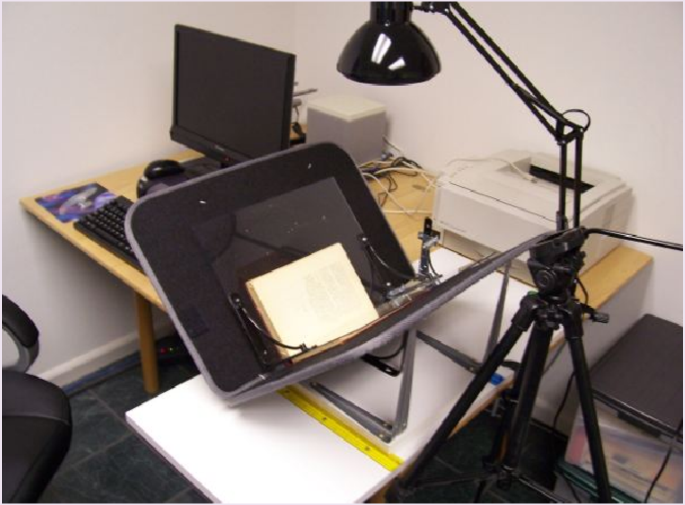
EXAMPLE-3

Open Hardware-3



EXAMPLE

Open Hardware

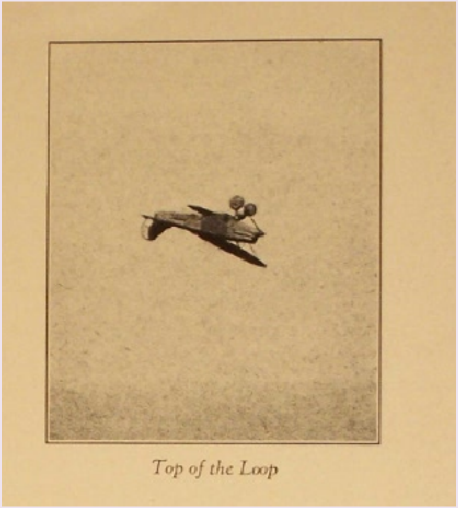


POST PROCESSING

- Images are Often Imperfect
- Imperfect Images cannot be Easily OCRed
- Rotation: for orientation
- Cropping: To remove portion outside book
- Skew Correction
- White Balance

POST PROCESSING (2)

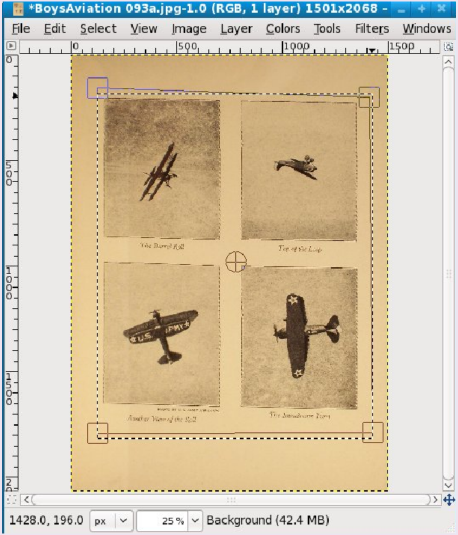
Keystoned Pages: Camera not parallel to book



Top of the Loop

POST PROCESSING (3)

Perspective Tool: GIMP



POST PROCESSING (4)

- Batch Processing: ImageMagick
- Crop 100 Pages: mogrify -crop 1298x1800+400+500 *.jpg
- 1298x1800 :original size in px
- +400+500: Offsets
- mogrify -rotate a *.jpg: Rotate by x degrees
- mogrify -scale 30% -format jpg -quality 90% -verbose *.jpg: Reduces Size

POST PROCESSING(5)

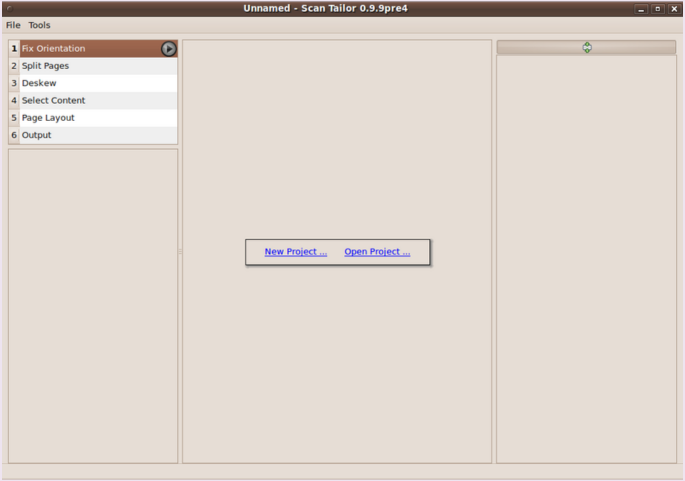
- Cleaning Pages: Grey, Yellow etc (White Balance)
- (A) Use Threshold Filter: GIMP, OK for OCR Only
- (B) Convert to grayscale and adjust Brightness Contrast
- mogrify -modulate 180,10,0 *.jpg: Brightness %, Saturation, Hue
- convert -v *.jpg abc.pdf: write Script for larger cases
- pdftk

SCAN TAILOR: THE EASY WAY

- Select 'Typical' Page
- Interpret Necessary Adjustments
- Use these in Batch Mode
- Undo if Necessary
- RTFM

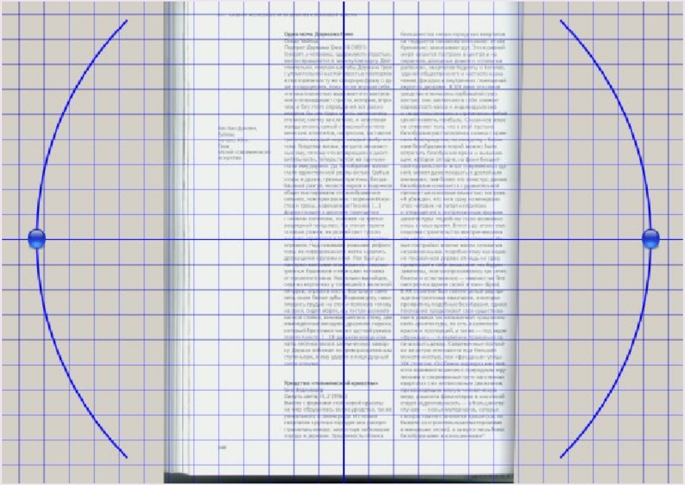
SCAN TAILOR (2)

Default Screen



SCAN TAILOR (3)

Removing Skew



SCAN TAILOR (4)

- Open ScanTailor & select 'new project'
- Browse and Select Images; Select Output Directory for Tiff output
- Set DPI Level: Use 300x300 or Estimate it First
- Set the orientation of the first image and 'apply to' and 'all pages'
- Click to 'split pages' (if required) or use Page Layout; Deskew
- In the 'output' stage use BW for simple text and line drawings etc: Always Preview

SCANNING TIPS

- Use Grayscale or Colour Mode, Not BW
- Do Not Scan at Resolution Less Than 300 DPI
- Do Not Scan to Lossy Formats like JPEG or Lossy-TIFF
- Scan to TIFF or PNG
- Use Lossless Compression Formats Like LZW
- Avoid scanning mode "Document" and Other Options

SIGIL

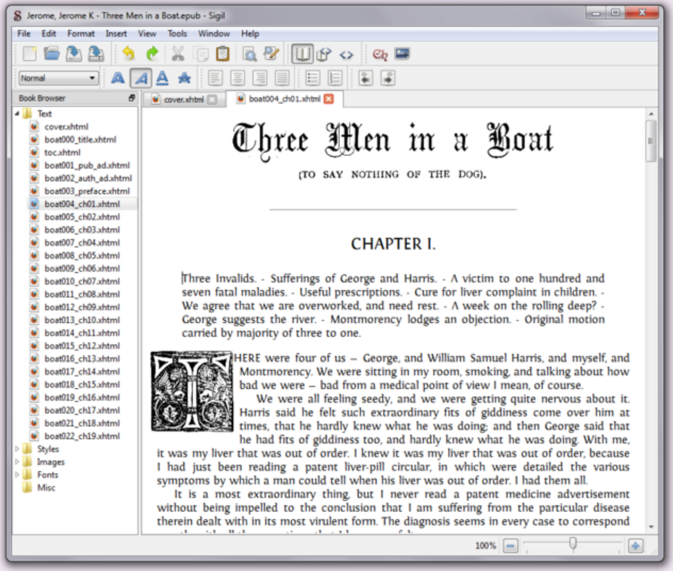
- A multi-platform WYSIWYG ebook editor
- Edit books in ePub format
- License: GPL Version-3
- Full Unicode support: UTF-16
- Full EPUB spec support
- Multiple Views: Book View, Code View and Split View

SIGIL (2)

- Metadata editor with full support for all possible metadata entries
- TOC Editor (XML)
- Multi-level TOC support
- SVG, XPGT (XML Page Template) Support
- Advanced automatic conversion of all imported documents to Unicode
- Embedded HTML Tidy; imported documents are thoroughly cleaned

SIGIL (3)

Editor View



Three Men in a Boat

(TO SAY NOTHING OF THE DOG).

CHAPTER I.

Three Invalids. - Sufferings of George and Harris. - A victim to one hundred and seven fatal maladies. - Useful prescriptions. - Cure for liver complaint in children. - We agree that we are overworked, and need rest. - A week on the rolling deep? - George suggests the river. - Montmorency lodges an objection. - Original motion carried by majority of three to one.

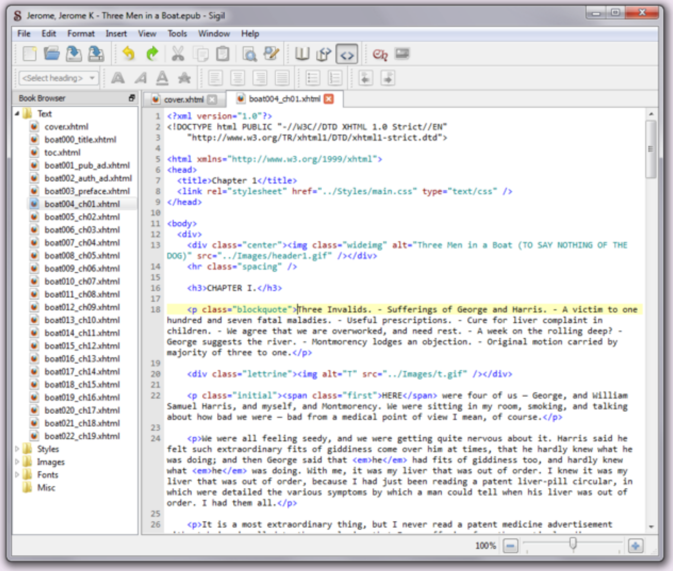
THERE were four of us – George, and William Samuel Harris, and myself, and Montmorency. We were sitting in my room, smoking, and talking about how bad we were – bad from a medical point of view I mean, of course.

We were all feeling seedy, and we were getting quite nervous about it. Harris said he felt such extraordinary fits of giddiness come over him at times, that he hardly knew what he was doing; and then George said that he had fits of giddiness too, and hardly knew what he was doing. With me, it was my liver that was out of order. I knew it was my liver that was out of order, because I had just been reading a patent liver-pill circular, in which were detailed the various symptoms by which a man could tell when his liver was out of order. I had them all.

It is a most extraordinary thing, but I never read a patent medicine advertisement without being impelled to the conclusion that I am suffering from the particular disease therein dealt with in its most virulent form. The diagnosis seems in every case to correspond

SIGIL (5)

Code View



```
<?xml version="1.0"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">
<head>
  <title>Chapter 1</title>
  <link rel="stylesheet" href="../Styles/main.css" type="text/css" />
</head>

<body>
  <div>
    <div class="center"><img class="wideimg" alt="Three Men in a Boat (TO SAY NOTHING OF THE
DOG)" src="../Images/header1.gif" /></div>
    <hr class="spacing" />

    <h3>CHAPTER I.</h3>

    <p class="blockquote">Three Invalids. - Sufferings of George and Harris. - A victim to one
hundred and seven fatal maladies. - Useful prescriptions. - Cure for liver complaint in
children. - We agree that we are overworked, and need rest. - A week on the rolling deep? -
George suggests the river. - Montmorency lodges an objection. - Original motion carried by
majority of three to one.</p>

    <div class="lettrine"><img alt="T" src="../Images/t.gif" /></div>

    <p class="initial"><span class="first">HERE</span> were four of us — George, and William
Samuel Harris, and myself, and Montmorency. We were sitting in my room, smoking, and talking
about how bad we were — bad from a medical point of view I mean, of course.</p>

    <p>We were all feeling seedy, and we were getting quite nervous about it. Harris said he
felt such extraordinary fits of giddiness come over him at times, that he hardly knew what he
was doing; and then George said that <em>he</em> had fits of giddiness too, and hardly knew
what <em>he</em> was doing. With me, it was my liver that was out of order. I knew it was my
liver that was out of order, because I had just been reading a patent liver-pill circular, in
which were detailed the various symptoms by which a man could tell when his liver was out of
order. I had them all.</p>

    <p>It is a most extraordinary thing, but I never read a patent medicine advertisement
```

COMICS

- CBR, CBZ
- Collect all images in order
- zip abc.cbz *.jpg: Done
- 7-Zip, Ark
- CBR, CBZ Formats are Simple.

DJVU

- Scan Tailor Will create larger djvu files: Color Images
- c44, pdf2djvu, Python Scripts
- Djvu photo files have many 'slices'
- Quality can be Controlled by reducing particular slices: c44
- djvm abc.djvu *.djvu: Combine djvu files

OCR

- Tesseract - http://code.google.com/p/tesseract-ocr/
- GNU Ocrad - Feature Extraction Based
- OCRDJVU -wrapper around OCR systems (OCRopus+Tesseract and Cuneiform)
- Engauge-Digitizer: Interactively Converts images to Numeric Data
- Kooka - kdegraphics package supports OCR through GOCR
- WeOCR - Web-enabled OCR

GOCR vs TESSERACT

- Use GOCR if all characters are quite 'Standard'
- GOCR does not attempt a character to character matching
- RTFM - For all the Options to be Passed
- Tesseract is Better Trained, Uses Corpus
- Ocropus: State of the Art Document Analysis and OCR system
- Ocropus: Can recognise Handwritten text

- Simmons, J. et al 'Reading and Sugar' http://flossmanuals.net
- Do It Yourself: http://diybookscanner.org
- Open Hardware: http://diybookscanner.org
- Sigil: http://code.google.com/p/sigil
- Scan Tailor: http://sourceforge.net/apps/mediawiki/scantailor
- Epub Check: http://code.google.com/p/epubcheck/

Open Repositories

- Project Gutenberg: `http://www.gutenberg.org`
- `http://gutenberg.net.au`; `http://gutenberg.net.ca`
- Rural Design Collective: `http://www.ruraldesigncollective.org/lab/ui/`
- Archives.Org: `http://www.archive.org/bookserver`
- Many Books: `http://manybooks.net/`
- FLOSS Manuals: `http://flossmanuals.net`

CHEERS!