

# A Look at EPUBs

And how to digitize a physical book using open source software

Davis Claiborne

LUG @ NC State

August 25, 2020



**Linux Users Group**  
at NC State University

# Overview

---

- What is an EPUB?
  - What makes up an EPUB?
  - Important files
- Digitization
  - Software and how to use it
  - Best practices I've learned

- What is an EPUB?
  - What makes up an EPUB?
  - Important files
- Digitization
  - Software and how to use it
  - Best practices I've learned

- This presentation is mainly about two things: epub, and how to turn a physical book of yours into an epub
- As far as EPUBs go, I'll talk about two main things: what an EPUB is and the important files that make it up
- For digitization, I'll talk about the software that I use to digitize books and some best practices

# Why?

---

- Why should I care?
  - Make your own
  - Fix existing ones
  
- Why should I digitize?
  - Convenience
  - Backup

# Making an eBook

└ Introduction

└ Motivation

└ Why?

- Why should I care?
  - Make your own
  - Fix existing ones
- Why should I digitize?
  - Convenience
  - Backup

- You may be wondering: "Why do I care about EPUBs?"
- Understanding the format of EPUBs allows you to easily make your own, as well as fix or modify existing EPUBs
- For instance, several of the digital books I read this summer contained minor issues - from typos to chapters being in the wrong order - that I was able to fix
- There are plenty of reasons why you'd want to convert your own books to EPUBs as well
- For one, EPUBs are much easier to travel with - you probably already take your phone with you virtually everywhere anyways, so why not read a book instead of checking Twitter?
- Additionally, it can be nice just to have a second copy if you want to loan out the book or if you're worried about it being damaged

# Why not?

---

- Hassle
- Time
- Don't want/need digital book

# Making an eBook

└ Introduction

└ Motivation

└ Why not?

## Why not?

- Hassle
- Time
- Don't want/need digital book

- There are also plenty of reasons **not** to digitize as well. It's a lot more work to digitize it than it is to just read a book.
- Additionally, it can be a decently large time-investment. And since the process can be error-prone and involves manual correction, some people may not consider it worth the amount of time and work it takes.
- If you would like to have the book easily on-hand to reference later, or are okay with some minor errors while you're reading, these two factors can be negligible.
- Basically, if you just don't really want or need an ebook, there's very little reason to do this

# Overview

---

- Zipped HTML and CSS files [2]
- Metadata/layout/TOC
- And more!



- Zipped HTML and CSS files [2]
- Metadata/layout/TOC
- And more!

- EPUBs files are just a zipped collection of HTML, XML, and CSS files, plus the other resources (images, fonts, etc.)
- All EPUB adds to HTML and CSS is a few files that make it easier to order and reference certain book-specific things, like the order, table of contents, and some metadata, like the publisher
- Although I just said that they're essentially just zipped HTML files, EPUBs are actually capable of more than just that - they can include sound, videos, and more. For this presentation I'll only be covering the very basics necessary to digitize your average book, but EPUBs are capable of handling more than just text.

# Overview

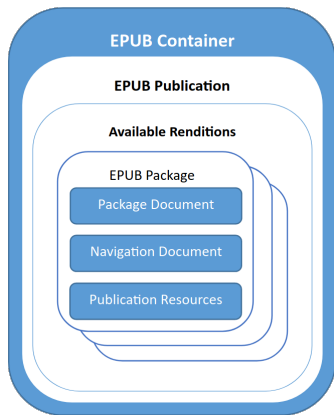


Figure 1: Basic structure of an EPUB [3]



Figure 1: Basic structure of an EPUB [1]

- This figure shows the basic layout of EPUBs. The container is the main zip file.
- Most of the terminology here is unimportant unless you're developing a reader for it, but there are a few that are interesting.
- A rendition is a viewport of some of the EPUB. There can be multiple renditions on the screen at once if you're looking at multiple pages at once.
- The package document has the manifest, which allows you to refer to different resources more easily, and defines the reading order
- The navigation document is an XHTML document which contains navigation information. The navigation information is different from the reading order, since it just indicates jump points.
- Publication resources dictate how the document should be rendered. Things like CSS, images, embedded fonts, etc.

# File structure

Directory:

- `mimetype`
- `META-INF/`
  - `container.xml`
- `OEBPS/`
  - `content.opf`
  - `text.html`
  - `stylesheet.css`
  - `toc.ncx`
  - `images/`
    - `cover.png`

[1]

```
Directory:
  • mimetype
  • META-INF/
    • container.xml
  • OEBPS/
    • content.opf
    • text.html
    • stylesheet.css
    • toc.ncx
    • images/
      • cover.png
[4]
```

- These files are in the root directory of the zipped file, where the bolded files are required to be named that way
- `mimetype` is a simple file that just contains file information.
- `container.xml` points to the file that's named `content.opf` here
- The next directory can be named whatever you want, but it's typically called "OEBPS," which stands for "Open eBook Publication Structure."
- `content.opf` points to all the other content in the book, including text, pictures, and the navigation file, `toc.ncx`.
- `toc.ncx` is the table of contents file. `ncx` stands for Navigation Center eXtended.
- See the citation for the exact content each file needs

# Overview

---

- Photograph book
- ScanTailor
- Tesseract
- OCRFeeder
- Manual revision

- Photograph book
- ScanTailor
- Tesseract
- OCRFeeder
- Manual revision

- If you want to digitize a book, the first step is to get pictures of the book somehow.
- Next, you use the program ScanTailor to perform simple operations to the pictures, like text de-distortion and page separation.
- Next, you can use OCRFeeder and Tesseract to extract the text and export the text to HTMLs.
- Finally, you can perform some manual revisions as needed.

# Photographing the book

---

- Camera
- Flatbed scanner
- Book scanner
- Document feeder



- Camera
- Flatbed scanner
- Book scanner
- Document feeder

- There are plenty of ways to get pictures of the book.
- You can use a camera, but if you do you need to be aware of the lighting, as well as distortion due to the lens. Also, try to avoid lossy compression (including jpg) if possible.
- You can use a regular scanner if the book is small enough.
- Companies produce specialized scanners, like the ones in the library, for scanning books.
- If you don't care about preserving the book you can cut out the pages and send them through a document feeder. This method yields the best results since it eliminates most sources of text distortion.

# ScanTailor

- Load pictures
- Simple tasks:
  - Fix rotation
  - Split pages
- Complex tasks:
  - Deskewing and dewarping text
  - Image detection



# Making an eBook

Digitization

Software

ScanTailor

- Load pictures
- Simple tasks:
  - Fix rotation
  - Split pages
- Complex tasks:
  - Deskewing and dewarping text
  - Image detection



- Load the pictures of the pages into ScanTailor
- ScanTailor takes care of a lot of the heavy lifting involved in this task
- It can do simple tasks, like fixing the rotation on pages or splitting pages (if you take pictures of two pages at a time)
- It can also do complex tasks though, like deskewing and dewarping text and detecting which content is text or not (which is useful for books with pictures)
- This can be run from the command line if you're not picky about how your images look, but in my experience the automated outputs usually require a decent bit of tweaking
- Note that ScanTailor is abandoned - there are active forks currently, but I still use the abandoned version because there's a feature it has that's not implemented in the active versions

# OCRFeeder + Tesseract



- Identify images/text
- Perform OCR
- Export to HTML



- Identify images/text
- Perform OCR
- Export to HTML

- Next, I feed the images output from ScanTailor to another program called OCRFeeder, which helps with OCR processing
- OCRFeeder supports many different OCR tools; Tesseract is the most common choice because it supports many different languages and does a generally pretty good job
- As with ScanTailor, you could also use a command line tool to do this, though the results usually won't be nearly as good
- After loading in the images you can select text and picture regions and have OCR done on the text regions
- Additionally, you can correct the OCR'd text output to catch errors if you'd like, though I tend to mostly just leave the output alone and correct it as I read along later
- Finally, once you've finished added text and image regions, you can export the project as HTML

## Zathura + Vim

- Edit in-place
- Hot-reloads
- Alternatives:
  - Sigil
  - Calibre



- Edit in-place
- Hot-reloads
- Alternatives:
  - Sigil
  - Calibre



- I know what you're probably thinking, and yes, that's right: I can't give a single presentation without talking about Vim
- But seriously, after moving around some of the HTML files and creating some of the necessary files for EPUBs, Vim is perfectly capable of editing the document, since it supports editing zipped files in-place (without unzipping/rezipping)
- Zathura, a PDF viewer with Vim-like keybindings, can view EPUBs using an extension, and will automatically reload the file when it detects changes
- This method does require a decent bit of HTML/CSS knowledge, so if you'd like a simpler method, there are plenty of tools that exist to just edit EPUBs in an easy-to-do format
- I've never used any of these so I can't vouch for them, but they seem to be fairly popular

## Step 1: Get pictures





# Making an eBook

## Demo

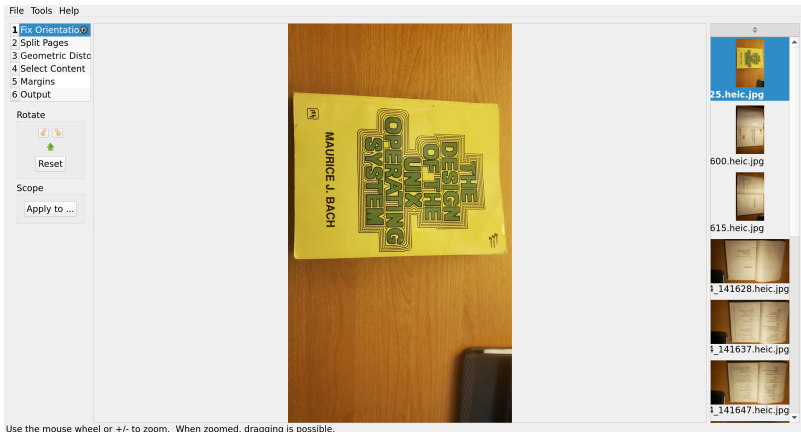
### Pictures

#### Step 1: Get pictures



- The first step to digitizing is to get pictures of your book
- You don't have to be super careful or picky about it if you're just planning to turn this into an EPUB, though the more prep time you spend, the less corrections you'll have to do later
- I generally just try to make sure the lighting is good and the pages are relatively flat
- As you can see, some pictures are upside-down - that's okay, as this can be fixed easily in scan tailor
- Some people get fancy with it - using pieces of glass to help flatten out the pages, creating stands to help the book lay flat
- These are helpful, and probably do improve the end result, but they're not necessary

## Step 2: Load images into ScanTailor



File Tools Help

- 1 Fix Orientation
- 2 Split Pages
- 3 Geometric Distc
- 4 Select Content
- 5 Margins
- 6 Output

Rotate

Reset

Scope

Apply to ...

Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.

2020-08-25

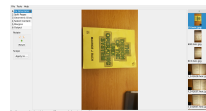
# Making an eBook

Demo

ScanTailor

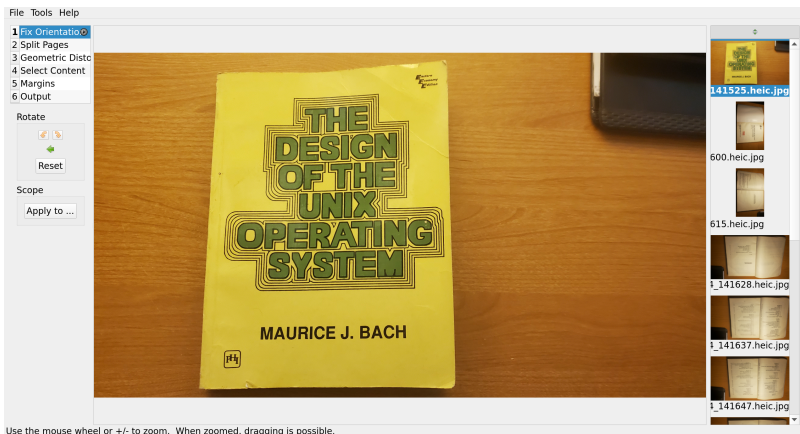
Step 2: Load images into ScanTailor

Step 2: Load images into ScanTailor



- The next step is to load the images into ScanTailor
- The interface here is pretty simple - all the pages are on the right and the actions you can do are on the left
- The actions are listed in the order they're performed - so the first action is fixing orientation, then splitting pages, etc.
- As you can see here, the first three pages don't have the right orientation; so let's try to fix that

## Step 2.1: Fix orientations



The screenshot displays the ScanTailor application window. The menu bar at the top includes 'File', 'Tools', and 'Help'. On the left, a sidebar contains a numbered list of tools: 1 Fix Orientation (highlighted), 2 Split Pages, 3 Geometric Distort, 4 Select Content, 5 Margins, and 6 Output. Below this list are controls for the 'Rotate' tool, including icons for clockwise and counter-clockwise rotation, a 'Reset' button, and a 'Scope' section with an 'Apply to ...' button. The main workspace shows a yellow book cover for 'THE DESIGN OF THE UNIX OPERATING SYSTEM' by MAURICE J. BACH. To the right, a vertical list of image thumbnails shows scanned pages, with the first one labeled '141525.heic.jpg' and others including '600.heic.jpg', '615.heic.jpg', and several '\_141628.heic.jpg' files. At the bottom of the window, a status bar reads: 'Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.'

# Making an eBook

Demo

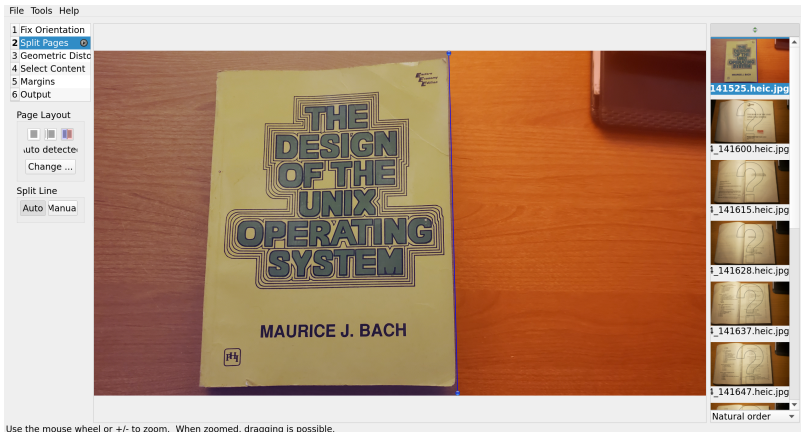
ScanTailor

Step 2.1: Fix orientations



- By clicking the play button next to the Fix Orientations label, ScanTailor will try to fix the orientations for each image
- Generally, this doesn't really do much, so you'll have to manually correct the images using the arrows below the actions list
- In this image, I've already applied the correct rotation to the first image
- Now I can select the pages I want to apply this transformation to using Control and left mouse and clicking the pages on the right
- Next, click "Apply to", then choose the "Selected pages" option to fix the rotation for the first three images

## Step 2.2: Split pages



The screenshot displays the ScanTailor application window. The main area shows a scan of a book cover titled "THE DESIGN OF THE UNIX OPERATING SYSTEM" by MAURICE J. BACH. The cover is yellow with blue and black text. The application's interface includes a menu bar (File, Tools, Help) and a sidebar with the following sections:

- 1 Fix Orientation**
- 2 Split Pages** (highlighted)
- 3 Geometric Distortion
- 4 Select Content
- 5 Margins
- 6 Output

Under "Page Layout", there are icons for page layout and a "Change ..." button. Under "Split Line", there are "Auto" and "Manual" options. On the right side, a vertical list of scanned pages is shown, each with a thumbnail and a filename:

- i\_141525.heic.jpg
- i\_141600.heic.jpg
- i\_141615.heic.jpg
- i\_141628.heic.jpg
- i\_141637.heic.jpg
- i\_141647.heic.jpg

At the bottom right, there is a "Natural order" dropdown menu. Below the main image, a note reads: "Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible."

2020-08-25

# Making an eBook

Demo

ScanTailor

Step 2.2: Split pages

Step 2.2: Split pages

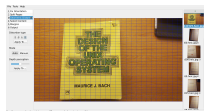


- Next, we need to split the pages - click the play button and wait for it to finish
- The blue line you see is how the page is going to be split - you can click and drag the circles on the end to change how it's split
- For this picture, since it's only a single "page", you can actually click the leftmost box under "Page Layout", which tells ScanTailor that this is only a single page
- This generally does a pretty good job, though I like to do a quick scan to make sure all the splits are good

## Step 2.3: Geometric distortion

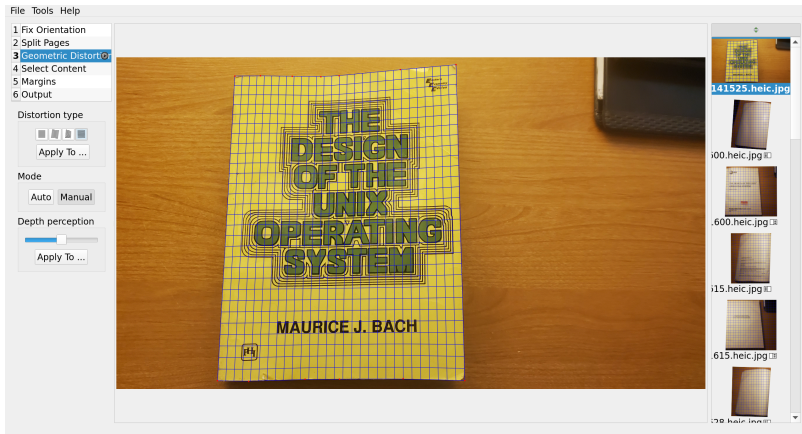
The screenshot shows the ScanTailor application interface. The main window displays a yellow book cover titled "THE DESIGN OF THE UNIX OPERATING SYSTEM" by MAURICE J. BACH, which is overlaid with a blue grid. The left sidebar contains a menu with the following options: 1 Fix Orientation, 2 Split Pages, 3 Geometric Distortion (highlighted), 4 Select Content, 5 Margins, and 6 Output. Below the menu are controls for "Distortion type" (with icons and an "Apply To ..." button), "Mode" (with "Auto" and "Manual" buttons), and "Depth perception" (with a slider and an "Apply To ..." button). The right sidebar shows a vertical list of image thumbnails, with the top one labeled "141525.heic.jpg" and others below it labeled "100.heic.jpg", "600.heic.jpg", "115.heic.jpg", "615.heic.jpg", and "170.heic.jpg". At the bottom of the window, a status bar reads: "Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible."





- Now we need to try and get the pages as straight and flat as possible
- On the left you see four options; from left to right: the first tells ScanTailor to do nothing; the second performs just rotation to deskew, the third tries to fix the “vertical angle” the picture was taken from, and the fourth tries to match the page and transform it to a rectangle
- I usually do the fourth, called “Curved lines,” since it’s pretty necessary in order to get good results later on with the setup I have - this option is also why I still use the abandoned version of ScanTailor
- Click the curved lines option, then choose apply to all and click play, and then save after - ScanTailor has a nasty habit of crashing later if you used the third or fourth options
- The grid you see here is what ST identified as the boundaries of the page (it did a really bad job here)

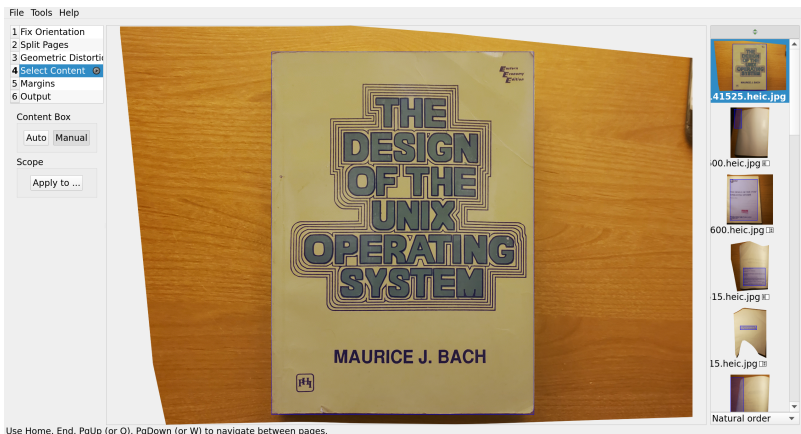
## Step 2.3: Geometric distortion





- Luckily, we can fix that - across the top and bottom there are several red dots you can click and drag to adjust it.
- After a bit of adjustment, here's what I ended up with
- The general idea is to try and get it to match the shape of the page as close as you can
- You can be as precise or inprecise as you want, though the closer it is to matching, the better your results will be
- For blank pages, I use the leftmost distortion option (No distortion) because it's bad at correcting blank pages (and the result won't matter anyways)

## Step 2.4: Select content



File Tools Help

- 1 Fix Orientation
- 2 Split Pages
- 3 Geometric Distortion
- 4 **Select Content**
- 5 Margins
- 6 Output

Content Box

Auto Manual

Scope

Apply to ...

THE DESIGN OF THE UNIX OPERATING SYSTEM

MAURICE J. BACH

41525.heic.jpg

00.heic.jpg

600.heic.jpg

15.heic.jpg

15.heic.jpg

Natural order

Use Home, End, PgUp (or Q), PgDown (or W) to navigate between pages.

## Making an eBook

└ Demo

└ ScanTailor

└ Step 2.4: Select content



- This is the part that causes ST to crash - make sure to save before starting it
- If ST does crash, you can narrow down the file by removing files from the project (right click > Remove from project) and trying again - if it does crash, at least one of the remaining pictures has a problem. If you do this, **do not save**, since you'll want to keep the pictures you removed.
- Assuming ST doesn't crash, you can now select which parts of the page you want by using the blue box
- In general, I crop it so that the header of each page isn't included. Beyond that, you can identify ones that turned out poorly by changing the "Natural order" on the bottom right to either "increasing width" or "increasing height"

## Step 2.5: Tweak geometric distortions

File Tools Help

1 Fix Orientation  
 2 Split Pages  
 3 Geometric Distortions  
 4 Select Content  
 5 Margins  
 6 Output

Content Box

Auto Manual

Scope

Apply to ...

21.heic.jpg  
 30.heic.jpg  
 30.heic.jpg  
 38.heic.jpg  
 8.heic.jpg  
 Natural order

Figure 3.1 Architecture of UNIX System

UNIX system architecture

Other application programs

User application programs

System

Kernel

Use the context menu to enable / disable the content box.

2020-08-25

# Making an eBook

Demo

ScanTailor

Step 2.5: Tweak geometric distortions

Step 2.5: Tweak geometric distortions



- As you're doing the select content step, you'll notice that some of the pages didn't turn out like you thought they would
- For instance, here the diagram is very wiggly, so I decided to go back and redo the curved lines so that they took up the whole page
- You can go back, redo the geometric distortion, then reselect "Select content" to see how the changes look

## Step 2.6: Margins

The screenshot shows the ScanTailor application window. The menu bar includes File, Tools, and Help. A list of steps is on the left, with '5 Margins' selected. The 'Margins' panel shows settings for Top, Bottom, Left, and Right, all set to 0.0%. Below this is an 'Alignment' section with radio buttons for 'Don't match size' (selected), 'Match size by growing margins', and 'Match size by scaling'. There are also icons for zooming and an 'Apply To ...' button. The central area displays a book cover for 'THE DESIGN OF THE UNIX OPERATING SYSTEM' by MAURICE J. BACH. The cover has a yellow background with green text and a circuit-like border. To the right is a thumbnail gallery showing various pages of the book, with '25.heic.jpg' at the top and '28.heic.jpg' at the bottom. The status bar at the bottom left says 'Natural order'.

File Tools Help

- 1 Fix Orientation
- 2 Split Pages
- 3 Geometric Distortions
- 4 Select Content
- 5 Margins**
- 6 Output

Margins

Top 0.0%

Bottom 0.0%

Left 0.0%

Right 0.0%

Apply To ...

Alignment

- Don't match size
- Match size by growing margins
- Match size by scaling

Apply To ...

25.heic.jpg

00.heic.jpg

00.heic.jpg

11615.heic.jpg

141615.heic.jpg

28.heic.jpg

Natural order

Resize margins by dragging any of the solid lines.



2020-08-25

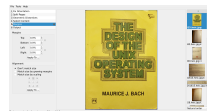
# Making an eBook

Demo

ScanTailor

Step 2.6: Margins

Step 2.6: Margins



- Next, choose your margins
- For creating an EPUB, I always just set the margins to 0, then select “Don’t match size,” since anything that I didn’t select as content is irrelevant to me

## Step 2.7: Output

File Tools Help

- 1 Fix Orientation
- 2 Split Pages
- 3 Geometric Distortions
- 4 Select Content
- 5 Margins
- 6 Output**

Resolution Enhancement

1x 1.5x 2x

This page: 3119 x 4306 px

Mode

Color / Grayscale ▾

White margins

Equalize illumination

Apply To ...

Output

Picture Zones

25.heic.jpg

Fill Zones

600.heic.jpg

Despeckling

00.heic.jpg

11615.heic.jpg

141615.heic.jpg

628.heic.jpg

Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.

2020-08-25

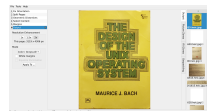
# Making an eBook

Demo

ScanTailor

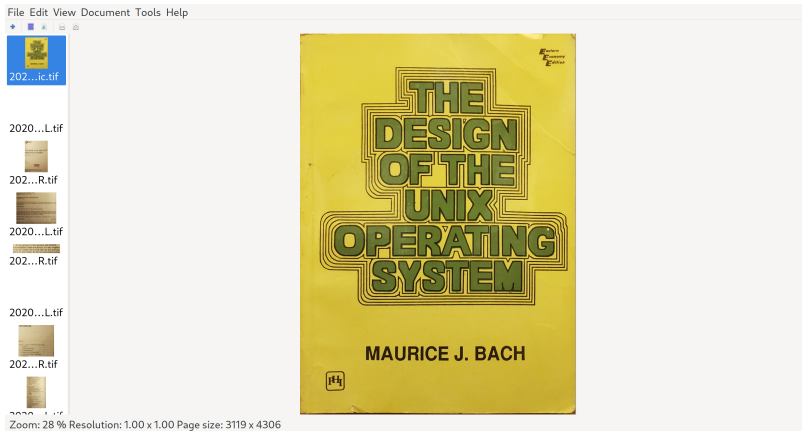
Step 2.7: Output

Step 2.7: Output



- Finally, we get to the output stage
- Here, you can choose to increase or decrease the resolution, as well as choose from three options how you want the pages to appear (color, black and white, and mixed)
- I tend to choose color for EPUBs, but I'll choose mixed if I'm making a PDF since it makes the file size a good bit smaller
- While ST does have pretty good Black and White conversion and image detection, in my experience Tesseract performs better on color images, which is why I choose color
- That's it for ST! If you look in the directory where the original images were, there's now a directory called "out" with the final images as tif files

## Step 3: Load images into OCRFeeder



2020-08-25

# Making an eBook

Demo

OCRFeeder

Step 3: Load images into OCRFeeder

Step 3: Load images into OCRFeeder



- Next, we need to OCR the text in the pages and extract the images
- To do that, open OCRFeeder and then choose “File > Add folder” and select the “out” folder from ST
- I’m not the biggest fan of this program’s interface, but you can get used to it relatively quickly



## Making an eBook

Demo

OCRFeeder

Step 3.1: Select text and images



- Left click and drag over a region to select it
- Then, using the menu on the right, identify it as either text or image
- If it's text, there should be an OCR button you can choose to identify the text
- Using the box in the lower right, you can correct the OCR text if you'd like, though I normally prefer to do that as I read through
- I like to break up the text selection by paragraph, since that makes it easier later for the EPUB process

## Step 3.2: Export

The screenshot shows the OCRFeeder application interface. On the left is a file list with thumbnails. The main window displays a document page with a file tree diagram overlaid. The diagram shows a root node with branches for 'etc', 'usr', 'sbin', and 'dev'. Under 'etc' are 'hostname', 'id', 'date', and 'who'. Under 'usr' are 'passwd', 'irc', and 'bin'. Under 'sbin' is 'mod'. Under 'dev' are 'devic' and 'arbois'. A caption below the diagram reads "Figure 1.2. Sample File System Tree." Below the diagram is a text block with a dialog box titled "Export pages" and "Choose the format" with options "HTML" and "Cancel" and "OK".

File Edit View Document Tools Help

202...R.tif  
2020...L.tif  
202...R.tif  
2020...L.tif  
202...R.tif  
2020...L.tif  
202...R.tif  
2020...L.tif  
202...R.tif  
2020...L.tif  
202...#1.tif

Figure 1.2. Sample File System Tree.

Export pages  
Choose the format  
HTML  
Cancel OK

Type  
Text Image  
Clip  
Directories are like r  
directory as a byte stre  
directory in a predictabl  
Bounds  
Text Properties  
OCR engine to recognize this area:  
Tesseract - OCR  
Text Style Misc  
Directories are like regular files in this  
respect; the system treats the data in a  
directory as a byte stream, but the data  
contains the names of the files in the  
directory in a predictable format so that  
the operating system and programs  
such as

Zoom: 37 % Resolution: 1.00 x 1.00 Page size: 2159 x 3202



2020-08-25

# Making an eBook

Demo

OCRFeeder

Step 3.2: Export

Step 3.2: Export



- Once you've finished OCRing you can export the pages as HTML
- We choose this because, as you may remember, EPUBs are just zipped HTML files
- We are now done with OCRFeeder

## Step 4: Add files and zip

- Create directory/copy
- EPUB files
  - mimetype
  - META-INF/container.xml
  - OEBPS/content.opf
  - OEBPS/toc.ncx
- Create/rename
  - OEBPS/cover.html
  - OEBPS/chapter-01.html
- Zip

- Create directory/copy
- EPUB files
  - mimetype
  - META-INF/container.xml
  - OEBPS/content.opf
  - OEBPS/toc.ncx
- Create/rename
  - OEBPS/cover.html
  - OEBPS/chapter-01.html
- Zip

- All that's left before revision is to add the files EPUBs need
- I like to make a new directory and copy all the files over so I can have the originals in case I make a mistake, then move all OCRFeeder files to it (in a new OEBPS director)
- I also like to add a few blank files that I know I'll need, since I like to give my files meaningful names
- Since you can't add new files while working on zipped files in Vim, these files need to be added now as well
- I look ahead a bit and add files that I know I'll need, like cover.html, chapter-01.html, etc.
- I also like to rename the images in there to give them somewhat logical names; just make sure to rename the references in the HTML files as well
- Now, you can zip the files and edit them with vim

## Step 5: Edit with Vim

- Copy/edit files [1]
- Edit HTML
- Vim tips
  - Spelling: `:set spell`, CTRL-X + s
  - Modifying HTML: `cit`
  - Fancy characters: CTRL-K, `:help digraph-table`
    - C-K + "6: “
    - C-K + "9: ”
    - C-K + M-: —
  - Recording: `q<letter>`, `:help recording`

- Copy/edit files [1]
- Edit HTML
- Vim tips
  - Spelling: :set spell, CTRL-X + s
  - Modifying HTML: cit
  - Fancy characters: CTRL-K :help digraph-table
    - C-K + "Q" -
    - C-K + "B" -
    - C-K + "C" -
  - Recording: q{letter}, :help recording

- Follow the link referenced here to edit the files - I won't really talk about the process for editing these files since it's not actually that interesting
- Now you can format the HTML files however you like. I have a few tips that I like to follow that help speed up the process some:
- One of the main tips is for easily finding misspellings: set spell on, which will help identify misspelled words. You can also use control x + s to get suggestions.
- Next, with the cit key combo, which you can remember with "Change inner tag," you can modify inside HTML tags more easily
- You can insert fancy characters you want using digraphs: control K to start the digraph, followed by two characters you want; some that I use frequently are for open and close quotes and em dash
- Finally, you can record macros for tasks you repeat a lot

# Tips

---

- Whole page
- Smaller groups
- Select content
- Footnotes

- Whole page
- Smaller groups
- Select content
- Footnotes

- Be sure the whole page is in the picture
- I like to split my pictures into smaller groups to make the wait times shorter (it still has the same total wait time, but it feels faster, plus if ScanTailor crashes it's easier to figure out which page caused the problem)
- When on the “Select Content” step, be sure you're not clipping too close to the top or bottom of the text; Tesseract works much better when you give it a bit of extra space
- Footnotes aren't extremely well-supported in EPUBs - the best, most widely-compatible way is to make a new HTML file for each footnote and to put those at the end of the chapter or file

# References

---

- [1] Build a digital book with EPUB <https://www.ibm.com/developerworks/xml/tutorials/x-epubtut/index.html>
- [2] EPUB 3.2 <https://www.w3.org/publishing/epub3/epub-spec.html>
- [3] EPUB 3.2 Overview <https://www.w3.org/publishing/epub3/epub-overview.html>