PUBLIC LIBRARY

&

MULTIMEDIA INSTITUTE

BOOK SCANNING & POST-PROCESSING MANUAL

BASED ON PUBLIC LIBRARY OVERHEAD SCANNER

Written by:

Tomislav Medak

Dubravka Sekulić

With help of:

An Mertens

TABLE OF CONTENTS

INTRODUCTION:
BOOK SCANNING - FROM PAPER BOOK TO E-BOOK

Initial considerations when deciding on a scanning setup

Book scanning tends to be a fragile and demanding process. Many factors can go wrong or produce results of varying quality from book to book or page to page, requiring experience or technical skill to resolve issues that occur. Cameras can fail to trigger, components to communicate, files can get corrupted in the transfer, storage card doesn't get purged, focus fails to lock, lighting conditions change. There are trade-offs between the automation that is prone to instability and the robustness that is prone to become time consuming.

Your initial choice of book scanning setup will have to take these trade-offs into consideration. If your scanning community is confined to your hacklab, you won't be risking much if technological sophistication and integration fails to function smoothly. But if you're aiming at a broad community of users, with varying levels of technological skill and patience, you want to create as much time-saving automation as possible on the condition of keeping maximum stability. Furthermore, if the time of individual members of your scanning community can contribute is limited, you might also want to divide some of the tasks between users and their different skill levels.

This manual breaks down the process of digitization into a general description of steps in the workflow leading from the printed book to a digital e-book, each of which can be in a concrete situation addressed in various manners depending on the scanning equipment, software, hacking skills and user skill level that are available to your book scanning project. Several of those steps can be handled by a single piece of equipment or software, or you might need to use a number of them - your mileage will vary. Therefore, the manual will try to indicate the design choices you have in the process of planning your workflow and should help you make decisions on what design is best for you situation.

Introducing book scanner designs

The book scanning starts with the capturing of digital image files on the scanning equipment. There are three principle types of book scanner designs:

- flatbed scanner
- single camera overhead scanner
- dual camera overhead scanner

Conventional flatbed scanners are widely available. However, given that they require the book to be spread wide open and pressed down with the platen in order to break the resistance of the book binding and expose sufficiently the inner margin of the text, it is the most destructive approach for the book, imprecise and slow.

Therefore, book scanning projects across the globe have taken to custom designing improvised setups or scanner rigs that are less destructive and better suited for fast turning and capturing of pages. Designs abound. Most include:

- one or two digital photo cameras of lesser or higher quality to capture the pages,
- transparent V-shaped glass or Plexiglas platen to press the open book against a V-shape cradle, and
- a light source.

The go-to web resource to help you make an informed decision is the DIY book scanning community at http://diybookscanner.org. A good place to start is their intro (http://wiki.diybookscanner.org/) and scanner build list (http://wiki.diybookscanner.org/scanner-build-list).

The book scanners with a single camera are substantially cheaper, but come with an added difficulty of de-warping the distorted page images due to the angle that pages are photographed at, which can sometimes be difficult to correct in the post-processing. Hence, in this introductory chapter we'll focus on two camera designs where the camera lens stands relatively parallel to the page. However, with a bit of adaptation these instructions can be used to work with any other setup.

The Public Library scanner

In the focus of this manual is the scanner built for the Public Library project, designed by Voja Antonić (see Illustration 1). The Public Library scanner was built with the immediate use by a wide community of users in mind. Hence, the principle consideration in designing the Public Library scanner was less sophistication and more robustness, facility of use and distributed process of editing.

The board designs can be found here: http://www.memoryoftheworld.org/blog/2012/10/28/our-beloved-bookscanner. The current iterations are using two Canon 1100 D cameras with the kit lens Canon EF-S 18-55mm 1:3.5-5.6 IS. Cameras are auto-charging.

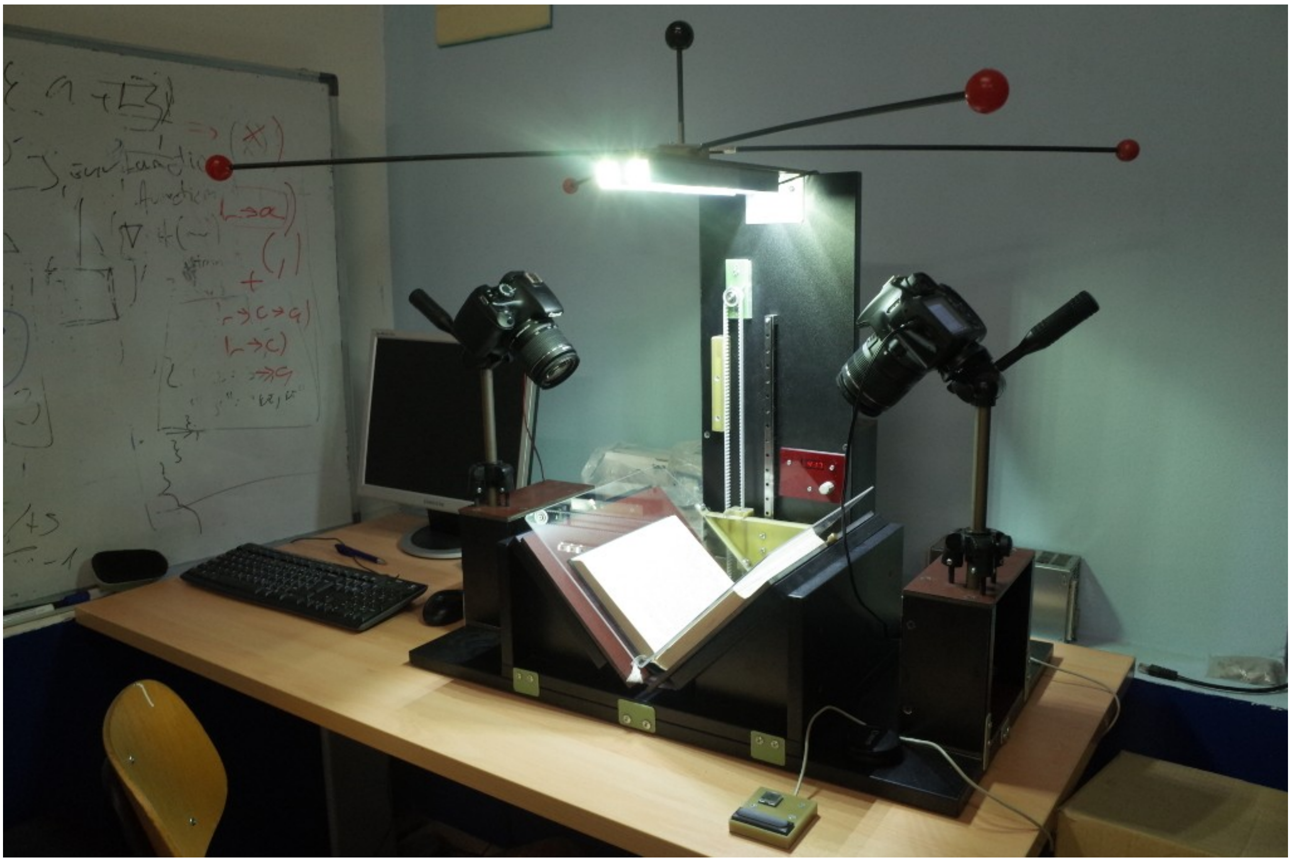

Illustration 1: Public Library Scanner

The scanner operates by automatically lowering the Plexiglas platen, illuminating the page and then triggering camera shutters. The turning of pages and the adjustments of the V-shaped cradle holding the book are manual.

The scanner is operated by a two-button controller (see Illustration 2). The upper, smaller button breaks the capture process in two steps: the first click lowers the platen, increases the light level and allows you to adjust the book or the cradle, the second click triggers the cameras and lifts the platen.

The lower button has two modes. A quick click will execute the whole capture process in one go. But if you hold it pressed longer, it will lower the platen, allowing you to adjust the book and the cradle, and lift it without triggering cameras when you press again.

Illustration 2: A two-button controller

More on this manual: steps in the book scanning process

The book scanning process in general can be broken down in six steps, each of which will be dealt in a separate chapter in this manual:

I. Photographing a printed book
I. Getting the image files ready for post-processing
III. Transformation of source images into .tiffs
IV. Optical character recognition
V. Creating a finalized e-book file
VI. Cataloging and sharing the e-book

A step by step manual for Public Library scanner

This manual is primarily meant to provide a detailed description and step-by-step instructions for an actual book scanning setup -- based on the Voja Antonić's scanner design described above. This is a two-camera overhead scanner, currently equipped with two Canon 1100 D cameras with EF-S 18-55mm 1:3.5-5.6 IS kit lens. It can scan books of up to A4 page size.

The post-processing in this setup is based on a semi-automated transfer of files to a GNU/Linux personal computer and on the use of free software for image editing, optical character recognition and finalization of an e-book file. It was initially developed for the HAIP festival in Ljubljana in 2011 and perfected later at MaMa in Zagreb and Leuphana University in Lüneburg.

Public Library scanner is characterized by a somewhat less automated yet distributed scanning process than highly automated and sophisticated scanner hacks developed at various hacklabs. A brief overview of one such scanner, developed at the Hacker Space Bruxelles, is also included in this manual.

The Public Library scanning process proceeds thus in following discrete steps:

1. creating digital images of pages of a book,
2. manual transfer of image files to the computer for post-processing,
3. automated renaming of files, ordering of even and odd pages, rotation of images and upload to a cloud storage,
4. manual transformation of source images into .tiff files in ScanTailor
5. manual optical character recognition and creation of PDF files in gscan2pdf

The detailed description of the Public Library scanning process follows below.

The Bruxelles hacklab scanning process
For purposes of comparison, here we'll briefly reference the scanner built by the Bruxelles hacklab (http://hackerspace.be/ScanBot). It is a dual camera design too. With some differences in hardware functionality (Bruxelles scanner has automatic turning of pages, whereas Public Library scanner has manual turning of pages), the fundamental difference between the two is in the post-processing - the level of automation in the transfer of images from the cameras and their transformation into PDF or DjVu e-book format.

The Bruxelles scanning process is different in so far as the cameras are operated by a computer and the images are automatically transferred, ordered and made ready for further post-processing. The scanner is home-brew, but the process is for advanced DIY'ers. If you want to know more on the design of the scanner, contact Michael Korntheuer at contact@hackerspace.be.

The scanning and post-processing is automated by a single Python script that does all the work http://git.constantvzw.org/?p=algolit.git;a=tree;f=scanbot_brussel;h=81facf5cb106a8e4c2a76c048694a3043b158d62;hb=HEAD

The scanner uses two Canon point and shoot cameras. Both cameras are connected to the PC with USB. They both run PTP/CHDK (Canon Hack Development Kit). The scanning sequence is the following:

1. Script sends CHDK command line instructions to the cameras
2. Script sorts out the incoming files. This part is tricky. There is no reliable way to make a distinction between the left and right camera, only between which camera was recognized by USB first. So the protocol is to always power up the left camera first. See the instructions with the source code.
3. Collect images in a PDF file
4. Run script to OCR a .PDF file to plain .TXT file: http://git.constantvzw.org/?p=algolit.git;a=blob;f=scanbot_brussel/ocr_pdf.sh;h=2c1f24f9afcce03520304215951c65f58c0b880c;hb=HEAD

I. PHOTOGRAPHING A PRINTED BOOK

Technologically the most demanding part of the scanning process is creating digital images of the pages of a printed book. It's a process that is very different form scanner design to scanner design, from camera to camera. Therefore, here we will focus strictly on the process with the Public Library scanner.

Operating the Public Library scanner

0. Before you start:

Better and more consistent photographs lead to a more optimized and faster post-processing and a higher quality of the resulting digital e-book. In order to guarantee the quality of images, before you start it is necessary to set up the cameras properly and prepare the printed book for scanning.

a) Loosening the book

Depending on the type and quality of binding, some books tend to be too resistant to opening fully to reveal the inner margin under the pressure of the scanner platen. It is thus necessary to "break in" the book before starting in order to loosen the binding. The best way is to open it as wide as possible in multiple places in the book. This can be done against the table edge if the book is more rigid than usual. (Warning – "breaking in" might create irreversible creasing of the spine or lead to some pages breaking loose.)

b) Switch on the scanner

You start the scanner by pressing the main switch or plugging the power cable into the the scanner. This will also turn on the overhead LED lights.

c) Setting up the cameras

Place the cameras onto tripods. You need to move the lever on the tripod's head to allow the tripod plate screwed to the bottom of the camera to slide into its place. Secure the lock by turning the lever all the way back.

If the automatic chargers for the camera are provided, open the battery lid on the bottom of the camera and plug the automatic charger. Close the lid.

Switch on the cameras using the lever on the top right side of the camera's body and place it into the aperture priority (Av) mode on the mode dial above the lever (see Illustration 3). Use the main dial just above the shutter button on the front side of the camera to set the aperture value to F8.0.

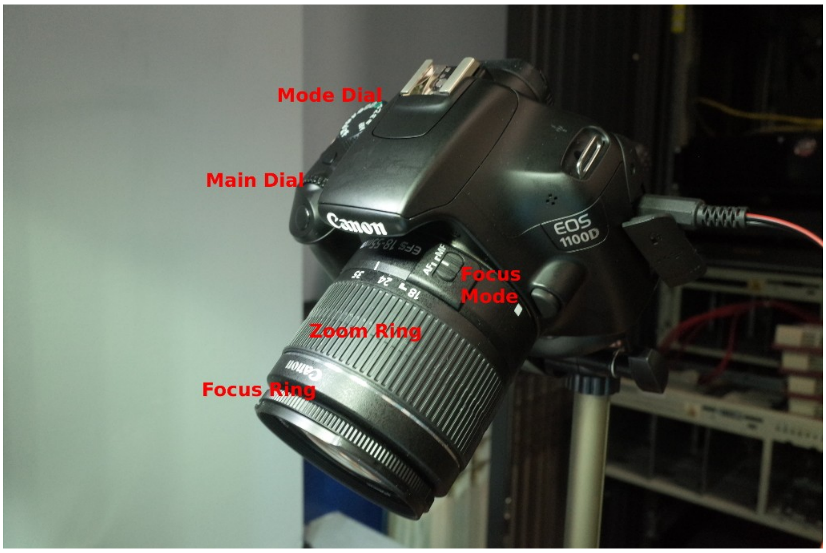

Illustration 3: Mode and main dial, focus mode switch, zoom and focus ring

On the lens, turn the focus mode switch to manual (MF), turn the large zoom ring to set the value exactly midway between 24 and 35 mm (see Illustration 3). Try to set both cameras the same.

To focus each camera, open a book on the cradle, lower the platen by holding the big button on the controller, and turn on the live view on camera LCD by pressing the live view switch (see Illustration 4). Now press the magnification button twice and use the focus ring on the front of the lens to get a clear image view.

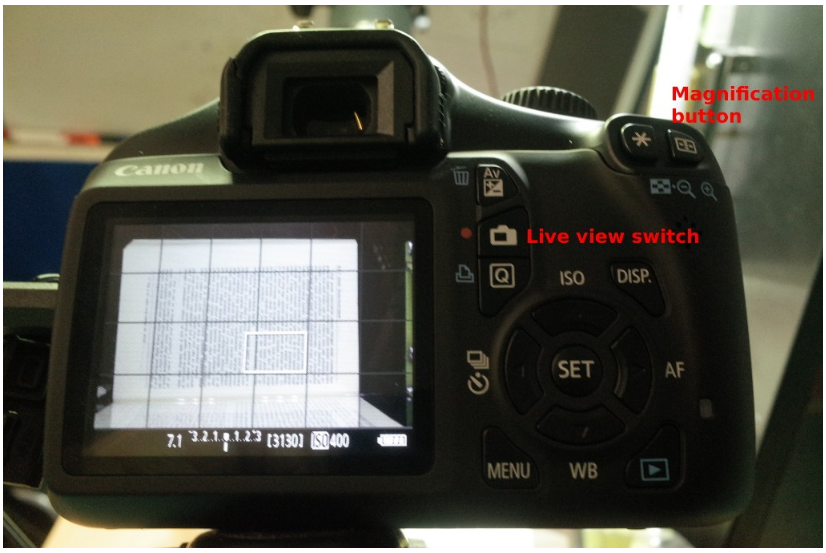

Illustration 4: Live view switch and magnification button

d) Connecting the cameras

Now connect the cameras to the remote shutter trigger cables that can be found lying on each side of the scanner. They need to be plugged into a small round port hidden behind a protective rubber cover on the left side of the cameras.

e) Placing the book into the cradle and double-checking the cameras

Open the book in the middle and place it on the cradle. Hold pressed the large button on the controller to lower the Plexiglas platen without triggering the cameras. Move the cradle so that the the platen fits into with the middle of the book.

Turn on the live view on the cameras' LED to see if the the pages fit into the image and if the cameras are positioned parallel to the page.

f) Double-check storage cards and batteries

It is important that both storage cards on cameras are empty before starting the scanning in order not to mess up the page sequence when merging photos from the left and the right camera in the post-processing. To double-check, press play button on cameras and erase if there are some photos left from the previous scan -- this you do by pressing the menu button, selecting the fifth menu from the left and then select 'Erase Images' -> 'All images on card' -> 'OK'.

If no automatic chargers are provided, double-check on the information screen that batteries are charged. They should be fully charged before starting with the scanning of a new book.

g) Turn off the light in the room

Lighting conditions during scanning should be as constant as possible, to reduce glare and achieve maximum quality remove any source of light that might reflect off the Plexiglas platen. Preferably turn off the light in the room or isolate the scanner with the black cloth provided.

1. Photographing a book

Now you are ready to start scanning. Place the book closed in the cradle and lower the platen by holding the large button on the controller pressed (see Illustration 2). Adjust the position of the cradle and lift the platen by pressing the large button again.

To scan you can now either use the small button on the controller to lower the platen, adjust and then press it again to trigger the cameras and lift the platen. Or, you can just make a short press on the large button to do it in one go.

> ATTENTION: When the cameras are triggered, the shutter sound has to be heard coming from both cameras. If one camera is not working, it's best to reconnect both cameras (see Section 0), make sure the batteries are charged or adapters are connected, erase all images and restart.
>
> A mistake made in the photographing requires a lot of work in the post-processing, so it's much quicker to repeat the photographing process.

If you make a mistake while flipping pages, or any other mistake, go back and scan from the page you missed or incorrectly scanned. Note down the page where the error occurred and in the post-processing the redundant images will be removed.

> ADVICE: The scanner has a digital counter. By turning the dial forward and backward, you can set it to tell you what page you should be scanning next. This should help you avoid missing a page due to a distraction.

While scanning, move the cradle a bit to the left from time to time, making sure that the tip of V-shaped platen is aligned with the center of the book and the inner margin is exposed enough.

II. GETTING THE IMAGE FILES READY FOR POST-PROCESSING

Once the book pages have been photographed, they have to be transfered to the computer and prepared for post-processing. With two-camera scanners, the capturing process will result in two separate sets of images -- odd and even pages -- coming from the left and right cameras respectively -- and you will need to rename and reorder them accordingly, rotate them into a vertical position and collate them into a single sequence of files.

a) Transferring image files

For the transfer of files your principle process design choices are either to copy the files by removing the memory cards from the cameras and copying them to the computer via a card reader or to transfer them via a USB cable. The latter process can be automated by remote operating your cameras from a computer, however this can be done only with a certain number of Canon cameras (http://bit.ly/16xhJ6b) that can be hacked to run the open **C**anon **H**ack **D**evelopment **K**it firmware (http://chdk.wikia.com).

After transferring the files, you want to erase all the image files on the camera memory card, so that they would not end up messing up the scan of the next book.

b) Renaming image files

As the left and right camera are typically operated in sync, the photographing process results in two separate sets of images, with even and odd pages respectively, that have completely different file names and potentially same time stamps. So before you collate the page images in the order how they appear in the book, you want to rename the files so that the first image comes from the right camera, the second from the left camera, the third comes again from the right camera and so on. You probably want to do a batch renaming, where your right camera files start with *n* and are offset by an increment of 2 (e.g. page_0000.jpg, page_0002.jpg,...) and your left camera files start with *n*+1 and are also offset by an increment of 2 (e.g. page_0001.jpg, page_0003.jpg,...).

Batch renaming can be completed either from your file manager, in command line or with a number of GUI applications (e.g. GPrename, rename, cuteRenamer on GNU/Linux).

c) Rotating image files

Before you collate the renamed files, you might want to rotate them. This is a step that can be done also later in the post-processing (see below), but if you are automating or scripting your steps this is a practical place to do it. The images leaving your cameras will be positioned horizontally. In order to position them vertically, the images from the camera on the right will have to be rotated by 90 degrees counter-clockwise, the images from the camera on the left will have to be rotated by 90 degrees clockwise.

Batch rotating can be completed in a number of photo-processing tools, in command line or dedicated applications (e.g. Fstop, ImageMagick, Nautilust Image Converter on GNU/Linux).

d) Collating images into a single batch

Once you're done with the renaming and rotating of the files, you want to collate them into the same folder for easier manipulation later.

Getting the image files ready for post-processing on the Public Library scanner

In the case of Public Library scanner, a custom C++ script was written by Mislav Stublić to facilitate the transfer, renaming, rotating and collating of the images from the two cameras.

The script prompts the user to place into the card reader the memory card from the right camera first, gives a preview of the first and last four images and provides an entry field to create a sub-folder in a local cloud storage folder (path: /home/user/Copy).

It transfers, renames, rotates the files, deletes them from the card and prompts the user to replace the card with the one from the left camera in order to the transfer the files from there and place them in the same folder. The script was created for GNU/Linux system and it can be downloaded, together with its source code, from: https://copy.com/nLSzflBnjoEB

If you have other cameras than Canon, you can edit the line 387 of the source file to change to the naming convention of your cameras, and recompile by running the following command in your terminal: "gcc scanflow.c -o scanflow -ludev `pkg-config --cflags --libs gtk+-2.0`"

In the case of Hacker Space Bruxelles scanner, this is handled by the same script that operates the cameras that can be downloaded from: http://git.constantvzw.org/?p=algolit.git;a=tree;f=scanbot_brussel;h=81facf5cb106a8e4c2a76c048694a3043b158d62;hb=HEAD

III. TRANSFORMATION OF SOURCE IMAGES INTO .TIFFS

Images transferred from the cameras are high definition full color images. You want your cameras to shoot at the largest possible .jpg resolution in order for resulting files to have at least 300 dpi (A4 at 300 dpi requires a 9.5 megapixel image). In the post-processing the size of the image files needs to be reduced down radically, so that several hundred images can be merged into an e-book file of a tolerable size.

Hence, the first step in the post-processing is to crop the images from cameras only to the content of the pages. The surroundings around the book that were captured in the photograph and the white margins of the page will be cropped away, while the printed text will be transformed into black letters on white background. The illustrations, however, will need to be preserved in their color or grayscale form, and mixed with the black and white text. What were initially large .jpg files will now become relatively small .tiff files that are ready for optical character recognition process (OCR).

These tasks can be completed by a number of software applications. Our manual will focus on one that can be used across all major operating systems -- ScanTailor. ScanTailor can be downloaded from: http://scantailor.sourceforge.net/. A more detailed video tutorial of ScanTailor can be found here: http://vimeo.com/12524529.

ScanTailor: from a photograph of a page to a graphic file ready for OCR

Once you have transferred all the photos from cameras to the computer, renamed and rotated them, they are ready to be processed in the ScanTailor.

1) Importing photographs to ScanTailor

- start ScanTailor and open '**new project**'
- for '**input directory**' chose the folder where you stored the transferred and renamed photo images
- you can leave '**output directory**' as it is, it will place your resulting .tiffs in an 'out' folder inside the folder where your .jpg images are
- **select all files (if you followed the naming convention above, they will be named 'page_xxxx.jpg')** in the folder where you stored the transferred photo images, and click 'OK'
- in the dialog box 'Fix DPI' click on All Pages, and for DPI choose preferably '600x600', click 'Apply', and then 'OK'

2) Editing pages

2.1 Rotating photos/pages

If you've rotated the photo images in the previous step using the scanflow script, skip this step.
- Rotate the first photo counter-clockwise, click Apply and for scope select 'Every other page' followed by 'OK'
- Rotate the following photo clockwise, applying the same procedure like in the previous step

2.2 Deleting redundant photographs/pages

- Remove redundant pages (photographs of the empty cradle at the beginning and the end of the book scanning sequence; book cover pages if you don't want them in the final scan; duplicate pages etc.) by right-clicking on a thumbnail of that page in the preview column on the right side, selecting 'Remove from project' and confirming by clicking on 'Remove'.

# If you by accident remove a wrong page, you can re-insert it by right-clicking on a page before/after the missing page in the sequence, selecting 'insert after/before' (depending on which page you selected) and choosing the file from the list. Before you finish adding, it is necessary to again go through the procedure of fixing DPI and Rotating.

2.3 Adding missing pages
- If you notice that some pages are missing, you can recapture them with the camera and insert them manually at this point using the procedure described above under 2.2.

3) Split pages and deskew
Steps 'Split pages' and 'Deskew' should work automatically. Run them by clicking the 'Play' button under the 'Select content' function. This will do the three steps automatically: splitting of pages, deskewing and selection of content. After this you can manually re-adjust splitting of pages and de-skewing.

4) Selecting content
Step 'Select content' works automatically as well, but it is important to revise the resulting selection manually page by page to make sure the entire content is selected on each page (including the header and page number). Where necessary, use your pointer device to adjust the content selection.

If the inner margin is cut, go back to 'Split pages' view and manually adjust the selected split area. If the page is skewed, go back to 'Deskew' and adjust the skew of the page. After this go back to 'Select content' and readjust the selection if necessary.

This is the step where you do visual control of each page. Make sure all pages are there and selections are as equal in size as possible.

At the bottom of thumbnail column there is a sort option that can automatically arrange pages by the height and width of the selected content, making the process of manual selection easier. The extreme differences in height should be avoided, try to make selected areas as much as possible equal, particularly in height, across all pages. The exception should be cover and back pages where we advise to select the full page.

5) Adjusting margins
For best results select in the previous step content of the full cover and back page. Now go to the 'Margins' step and set under Margins section both Top, Bottom, Left and Right to 0.0 and do 'Apply to...' → 'All pages'.

In Alignment section leave 'Match size with other pages' ticked, choose the central positioning of the page and do 'Apply to...' → 'All pages'.

6) Outputting the .tiffs
Now go to the 'Output' step. Ignore the 'Output Resolution' section.

Next review two consecutive pages from the middle of the book to see if the scanned text is too faint or too dark. If the text seems too faint or too dark, use slider Thinner – Thicker to adjust. Do 'Apply to' → 'All pages'.

Next go to the cover page and select under Mode 'Color / Grayscale' and tick on 'White Margins'. Do the same for the back page.

If there are any pages with illustrations, you can choose the 'Mixed' mode for those pages and then under the thumb 'Picture Zones' adjust the zones of the illustrations.

Now you are ready to output the files. Just press 'Play' button under 'Output'. Once the computer is finished processing the images, just do 'File' → 'Save as' and save the project.

IV. OPTICAL CHARACTER RECOGNITION

Before the edited-down graphic files are finalized as an e-book, we want to transform the image of the text into an actual text that can be searched, highlighted, copied and transformed. That functionality is provided by **O**ptical **C**haracter **R**ecognition. This a technically difficult task - dependent on language, script, typeface and quality of print - and there aren't that many OCR tools that are good at it. There is, however, a relatively good free software solution - Tesseract (http://code.google.com/p/tesseract-ocr/) - that has solid performance, good language data and can be trained for an even better performance, although it has its problems. Proprietary solutions (e.g. Abby FineReader) sometimes provide superior results.

Tesseract supports as input format primarily .tiff files. It produces a plain text file that can be, with the help of other tools, embedded as a separate layer under the original graphic image of the text in a PDF file.

With the help of other tools, OCR can be performed also against other input files, such as graphic-only PDF files. This produces inferior results, depending again on the quality of graphic files and the reproduction of text in them. One such tool is a bashscript to OCR a ODF file that can be found here: https://github.com/andrecastro0o/ocr/blob/master/ocr.sh

As mentioned in the 'before scanning' section, the quality of the original book will influence the quality of the scan and thus the quality of the OCR. For a comparison, have a look here: http://www.paramoulipist.be/?p=1303

Once you have your .txt file, there is still some work to be done. Because OCR has difficulties to interpret particular elements in the lay-out and fonts, the TXT file comes with a lot of errors.

Recurrent problems are:
- combinations of specific letters in some fonts (it can mistake 'm' for 'n' or 'I' for 'i' etc.);
- headers become part of body text;
- footnotes are placed inside the body text;
- page numbers are not recognized as such.

V. CREATING A FINALIZED E-BOOK FILE

After the optical character recognition has been completed, the resulting text can be merged with the images of pages and output into an e-book format. While increasingly the proper e-book file formats such as ePub have been gaining ground, PDFs still remain popular because many people tend to read on their computers, and they retain the original layout of the book on paper including the absolute pagination needed for referencing in citations. DjVu is also an option, as an alternative to PDF, used because of its purported superiority, but it is far less popular.

The export to PDF can be done again with a number of tools. In our case we'll complete the optical character recognition and PDF export in gscan2pdf. Again, the proprietary Abbyy FineReader will produce a bit smaller PDFs.

If you prefer to use an e-book format that works better with e-book readers, obviously you will have to remove some of the elements that appear in the book - headers, footers, footnotes and pagination.

This can be done earlier in the process of cropping down the original .jpg image files (see under III) or later by transforming the PDF files. This can be done in Calibre (http://calibre-ebook.com) by converting the PDF into an ePub, where it can be further tweaked to better accommodate or remove the headers, footers, footnotes and pagination.

Optical character recognition and PDF export in Public Library workflow

Optical character recognition with the Tesseract engine can be performed on GNU/Linux by a number of command line and GUI tools. Much of those tools exist also for other operating systems. For the users of the Public Library workflow, we recommend using gscan2pdf application both for the optical character recognition and the PDF or DjVu export.

To do so, start gscan2pdf and open your .tiff files. To OCR them, go to 'Tools' and select 'OCR'. In the dialog box select the Tesseract engine and your language. 'Start OCR'. Once the OCR is finished, export the graphic files and the OCR text to PDF by selecting 'Save as'.

However, given that sometimes the proprietary solutions produce better results, these tasks can also be done, for instance, on the Abbyy FineReader running on a Windows operating system running inside the Virtual Box. The prerequisites are that you have both Windows and Abbyy FineReader you can install in the Virtual Box. If using Virtual Box, once you've got both installed, you need to designate a shared folder in your Virtual Box and place the .tiff files there. You can now open them from the Abbyy FineReader running in the Virtual Box, OCR them and export them into a PDF.

To use Abbyy FineReader transfer the output files in your 'out' out folder to the shared folder of the VirtualBox. Then start the VirtualBox, start Windows image and in Windows start Abbyy FineReader. Open the files and let the Abbyy FineReader read the files. Once it's done, output the result into PDF.

VI. CATALOGING AND SHARING THE E-BOOK

Your road from a book on paper to an e-book is complete. If you want to maintain your library you can use Calibre, a free software tool for e-book library management. You can add the metadata to your book using the existing catalogues or you can enter metadata manually.

Now you may want to distribute your book. If the work you've digitized is in the public domain (https://en.wikipedia.org/wiki/Public_domain), you might consider contributing it to the Gutenberg project (http://www.gutenberg.org/wiki/Gutenberg:Volunteers'_FAQ#V.1._How_do_I_get_started_as_a_Project_Gutenberg_volunteer.3F), Wikibooks (https://en.wikibooks.org/wiki/Help:Contributing) or Arhive.org.

If the work is still under copyright, you might explore a number of different options for sharing.

QUICK WORKFLOW REFERENCE FOR SCANNING AND POST-PROCESSING ON PUBLIC LIBRARY SCANNER

I. PHOTOGRAPHING A PRINTED BOOK

0. Before you start:

- **loosen the book binding** by opening it wide on several places

- **switch on the scanner**

- **set up the cameras**:
    - **place cameras on tripods** and fit them tigthly
    - **plug in the automatic chargers** into the battery slot and close the battery lid
    - **switch on the cameras**
    - **switch the lens to Manual Focus mode**
    - **switch the cameras to Av mode** and set the aperture to 8.0
    - turn the zoom ring to **set the focal length exactly midway between 24mm and 35mm**
    - **focus** by turning on the live view, pressing magnification button twice and adjusting the focus to get a clear view of the text

- **connect the cameras to the scanner** by plugging the remote trigger cable to a port behind a protective rubber cover on the left side of the cameras

- place the book into the crade

- **double-check storage cards and batteries**
    - press the play button on the back of the camera to double-check if there are images on the camera - if there are, delete all the images from the camera menu
    - if using batteries, double-check that batteries are fully charged

- switch off the light in the room that could reflect off the platen and cover the scanner with the black cloth

1. Photographing

- **now you can start scanning** either by pressing **the smaller button** on the controller once to lower the platen and adjust the book, and then press again to increase the light intensity, trigger the cameras and lift the platen; or by pressing **the large button** completing the entire sequence in one go;

    - ATTENTION: Shutter sound should be coming from both cameras - if one camera is not working, it's best to reconnect both cameras, make sure the batteries are charged or adapters are connected, erase all images and restart.

    - ADVICE: The scanner has a digital counter. By turning the dial forward and backward, you can set it to tell you what page you should be scanning next. This should help you to avoid missing a page due to a distraction.

II. Getting the image files ready for post-processing

- after finishing with scanning a book, **transfer the files to the post-processing computer and purge the memory cards**

- if transferring the files **manually**:
    - **create two separate folders**,
    - **transfer the files** from the folders with image files on cards, using a batch renaming software **rename the files** from the right camera following the **convention** page_0001.jpg, page_0003.jpg, page_0005.jpg... -- and the files from the left camera following the convention page_0002.jpg, page_0004.jpg, page_0006.jpg...
    - **collate image files into a single folder**
    - before ejecting each card, **delete all the photo files on the card**

- if using the **scanflow script**:
    - **start the script** on the computer
    - **place the card** from the right camera into the card reader
    - enter the name of the destination folder following the convention "Name_Surname_Title_of_the_Book" and **transfer the files**
    - **repeat with the other card**
    - script will automatically transfer the files, rename, rotate, collate them in proper order and delete them from the card

III. Transformation of source images into .tiffs

ScanTailor: from a photograph of page to a graphic file ready for OCR

1) Importing photographs to ScanTailor

- start ScanTailor and open '**new project'**
- for '**input directory'** chose the folder where you stored the transferred photo images
- you can leave '**output directory'** as it is, it will place your resulting .tiffs in an 'out' folder inside the folder where your .jpg images are
- **select all files (if you followed the naming convention above, they will be named 'page_xxxx.jpg')** in the folder where you stored the transferred photo images, and click 'OK'
- in the dialog box 'Fix DPI' click on All Pages, and for DPI choose preferably '600x600', click 'Apply', and then 'OK'

2) Editing pages

2.1 Rotating photos/pages

If you've rotated the photo images in the previous step using the scanflow script, skip this step.

- **rotate the first photo counter-clockwise**, click Apply and for scope select 'Every other page' followed by 'OK'
- **rotate the following photo clockwise**, applying the same procedure like in the previous step

2.2 Deleting redundant photographs/pages

- **remove redundant pages** (photographs of the empty cradle at the beginning and the end; book cover pages if you don't want them in the final scan; duplicate pages etc.) by right-clicking on a thumbnail of that page in the preview column on the right, selecting 'Remove from project' and confirming by clicking on 'Remove'.

# If you by accident remove a wrong page, you can re-insert it by right-clicking on a page before/after the missing page in the sequence, selecting 'insert after/before' and choosing the file from the list. Before you finish adding, it is necessary to again go the procedure of fixing DPI and rotating.

2.3 Adding missing pages

- If you notice that some pages are missing, you can recapture them with the camera and insert them manually at this point using the procedure described above under 2.2.

3) Split pages and deskew

- **Functions 'Split Pages' and 'Deskew' should work automatically.** Run them **by clicking the 'Play'** button under the **'Select content' step**. This will do the three steps automatically: splitting of pages, deskewing and selection of content. After this you can manually re-adjust splitting of pages and de-skewing.

4) Selecting content and adjusting margins

- **Step 'Select content'** works automatically as well, but it is important to **revise the resulting selection manually page by page** to make sure the entire content is selected on each page (including the header and page number). Where necessary use your pointer device to adjust the content selection.

- **If the inner margin is cut,** go back to 'Split pages' view and manually adjust the selected split area. **If the page is skewed,** go back to 'Deskew' and adjust the skew of the page. After this go back to 'Select content' and readjust the selection if necessary.

- This is the step where you do **visual control of each page.** Make sure all pages are there and selections are as equal in size as possible.

- At the bottom of thumbnail column there is a sort option that can automatically arrange pages by the height and width of the selected content, making the process of manual selection easier. **The extreme differences in height should be avoided,** try to make selected areas as much as possible equal, particularly in height, across all pages. **The exception should be cover and back pages** where we advise to select the full page.

5) Adjusting margins

- Now go to the 'Margins' step and set under **Margins section both Top, Bottom, Left and Right to 0.0 and do 'Apply to...' → 'All pages'.**

- In Alignment section leave 'Match size with other pages' ticked, **choose the central**

positioning of the page and do 'Apply to...' → 'All pages'.

6) Outputting the .tiffs

- Now **go to the 'Output' step**.

- **Review two consecutive pages** from the middle of the book to see if the scanned text is too faint or too dark. If the text seems too faint or too dark, **use slider Thinner – Thicker to adjust. Do 'Apply to' → 'All pages'.**

- Next go to **the cover page** and select under **Mode 'Color / Grayscale'** and tick on 'White Margins'. Do the same for **the back page.**

- If there are any **pages with illustrations**, you can **choose the 'Mixed' mode** for those pages and then under the thumb 'Picture Zones' adjust the zones of the illustrations.

- To **output the files** press 'Play' button under 'Output'. **Save the project.**

IV. Optical character recognition & V. Creating a finalized e-book file

If using all free software:

1) open gscan2pdf (if not already installed on your machine, install gscan2pdf from the repositories, Tesseract and data for your language from https://code.google.com/p/tesseract-ocr/)

- point gscan2pdf to **open your .tiff files**
- for Optical Character Recognition, **select 'OCR' under the drop down menu 'Tools'**, select the Tesseract engine and your language, start the process
- once OCR is finished and to output to a PDF, **go under 'File' and select 'Save'**, edit the metadata and select the format, save

If using non-free software:

2) open Abbyy FineReader in VirtualBox (note: only Abby FineReader 10 installs and works - with some limitations - under GNU/Linux)

- transfer files in the 'out' folder to the folder shared with the VirtualBox
- point it to the readied .tiff files and it will complete the OCR
- save the file

REFERENCES

For more information on the book scanning process in general and making your own book scanner please visit:

DIY Book Scanner: http://diybookscannnner.org

Hacker Space Bruxelles scanner: http://hackerspace.be/ScanBot

Public Library scanner: http://www.memoryoftheworld.org/blog/2012/10/28/our-beloved-bookscanner/

Other scanner builds: http://wiki.diybookscanner.org/scanner-build-list

For more information on automation:

Konrad Voeckel's post-processing script (From Scan to PDF/A): http://blog.konradvoelkel.de/2013/03/scan-to-pdfa/

Johannes Baiter's automation of scanning to PDF process: http://spreads.readthedocs.org

For more information on applications and tools:

Calibre e-book library management application: http://calibre-ebook.com/

ScanTailor: http://scantailor.sourceforge.net/

gscan2pdf: http://sourceforge.net/projects/gscan2pdf/

Canon Hack Development Kit firmware: http://chdk.wikia.com

Tesseract: http://code.google.com/p/tesseract-ocr/

Python script of Hacker Space Bruxelles scanner: http://git.constantvzw.org/?p=algolit.git;a=tree;f=scanbot_brussel;h=81facf5cb106a8e4c2a76c048694a3043b158d62;hb=HEAD