Digitization Training Manual

Table of Contents

About

The Digital Library of the Caribbean (dLOC) is a cooperative of partners within the Caribbean and circum-Caribbean that provides users with access to Caribbean cultural, historical and research materials held in archives, libraries, and private collections. dLOC comprises collections that speak to the similarities and differences in histories, cultures, languages and governmental systems. Types of collections include but are not limited to: newspapers, archives of Caribbean leaders and governments, official documents, documentation and numeric data for ecosystems, scientific scholarship, historic and contemporary maps, oral and popular histories, travel accounts, literature and poetry, musical expressions, and artifacts.

dLOC has become a premier international collection of Caribbean research resources. dLOC is composed of content-contributing partners. Established in 2004, dLOC has grown from the initial nine partners to more than seventy (https://dloc.com/dloc1/partners). All materials in dLOC are freely and fully available as open access. dLOC builds capacity in the region to support digitization and preservation to provide access to holdings locally and internationally. As of December 2020, the collection consisted of more than 25,000 titles, over 180,000 items, and over 3.6 million pages of content (https://dloc.com/l/dloc1/itemcount). Also, as of December 2020, dLOC sees over 2 million hits per month (https://dloc.com/l/dloc1/usage).

dLOC's diverse partners serve an international community of scholars, students, and citizens by working together to provide enhanced electronic access to cultural, historical, legal, governmental, and research materials in a common web space with a multilingual interface.

Primary Contacts

dLOC Operations Team: dLOC@uflib.ufl.edu

Digital Scholarship Director:
Laurie Taylor, PhD (UF)
Phone: (352) 273-2902 | Email: laurien@ufl.edu

Technical Director:
Laura Perry (University of Florida/UF Libraries)

Phone: (352) 273-2911 | Email: lauraperry@ufl.edu

Copyright Specialist
Perry Collins, MLS (UF)
Phone: (352)-273-2710 | Email: perrycollins@ufl.edu

Executive Board

The Executive Board is composed of elected representatives from dLOC partner organizations. For more information, see the dLOC website on: https://dloc.domains.uflib.ufl.edu/governance/

Advisory Board

The Advisory Board is composed of scholars of Caribbean Studies. See the 2021 charge and roster for more information: https://dloc.com/UF00093611/00053

Collection Development

dLOC welcomes collections that address the histories, cultures, languages and governmental systems of Caribbean countries. The types of collections appropriate for dLOC include but are not limited to: newspapers, archives of Caribbean leaders and governments, official documents, documentation and numeric data for ecosystems, scientific scholarship, historic and contemporary maps, oral and popular histories, travel accounts, literature and poetry, musical expressions, and artifacts. If you have another collection that you think may be appropriate, please contact us to discuss the collection.

dLOC partners have total control over their collection development and material selection.

Partners often begin digitization by developing digital collection plans. Digital collection plans are a normal part of grant and project planning, and of planning in general, to ensure alignment with anticipated user demand and need for digital preservation. Partners often choose to digitize materials that can be developed into thematic or topical digital collections. These are productive for showcasing materials together for use in online exhibits, lesson plans, and promoting digital preservation of additional materials.

Copyright and Permissions

Copyright and permissions are supported in the *Guide to Requesting Permissions.*

Recommended Minimum Equipment

Computer workstation
PC with Windows 7 or 10; 16 GB RAM ; 500GB drive; CD-RW; 21 inch monitor

Scanner Recommendations
Large format, graphic arts quality scanner that will support digitization of photographs, book pages, maps, etc. Scanner will need to be calibrated periodically.

Recommended Scanners
https://guides.uflib.ufl.edu/digital_support_services/equipment

Flatbed scanner: Epson Expression 12000XL Graphic Arts Scanner or Microtek ScanMaker 9800XL

Sheet-feed scanner: Canon Image FORMULA DR-G1130

Recommended Minimum Software

Image Capture Software
Digital image capture software is supplied by scanner manufacturers. Not all image capture software is the same however. Epson image capture software supports a full range of image settings for flatbed scanning, whereas most consumer market scanner software does not.

Image Processing Software
Used to perform single and batch image editing (crop and deskew).

Recommended Image Processing software:
Adobe Photoshop (CS4, CS5, CS6, or CC)
Or, Adobe Photoshop Elements
GIMP – GNU Image Manipulation Program (Free download)

dLOC Applications and tools
Used to create digital object packages and metadata and to transfer files.

Metadata spreadsheet template: https://guides.uflib.ufl.edu/dloctech

Archive File Transfer Applications
Used to archive locally in addition to archiving remotely with dLOC.

Use "cloud" environments, (eg: Google Drive, OneDrive, Dropbox) to remotely transfer files.

Metadata Creation Overview

Entering metadata into the metadata spreadsheet template

Contact the operations team (dLOC@uflib.ufl.edu) for the template spreadsheet for creating metadata. Partners use the spreadsheet to visualize the available metadata fields in dLOC with formatting examples.

Once the metadata spreadsheet has been completed, partners return the completed spreadsheet with images to the dLOCoperations team via cloud file transfer or harddrive, and email to confirm dLOC@uflib.ufl.edu. The operations team will ingest the images and metadata into dLOC and provide the list of corresponding BIBID's.

Image Theory and Specifications

Requirements for preservation files:

- 8-bit Grayscale or 24-bit RGB Color (depending on whether the original has significant color)
- Minimum of 300 dpi for standard text of 600 dpi for stand-alone images (photographs, maps)
- Save archival files as uncompressed TIFFs

Bits Depth

In digitization, three levels of *Bit Depth* are widely used: 1 Bit, 8 Bit, and 24 Bit images.

A **1 Bit** image is referred to as "bi-tonal" or, less precisely, as "black-and-white". The picture elements of a 1 Bit image are expressed in stings of one bit. That bit may be either one color or an alternate and, frequently either black or white.

An **8 Bit** image is referred to as "grey-scale", though an 8 Bit image may represent a very limited color spectrum as well. Most scanning equipment defaults 8 Bit imaging to grey-scale. The picture elements of an 8 Bit image are expressed in strings of eight bits, for example: 00001111. 8 Bit images allow for as many as 255 shades or colors. (N.B. Technically, 8 Bit images allow for 256 shade/color values, but one of these is reserved as a check-digit and is not used to express a shade/color value.)

An **24 Bit** image is referred to as "true color" or, less precisely, as a "color" image. The picture elements of a 24 Bit image are expressed in strings of twenty-four bits. 24 Bit images allow for as many as 16,777,216 shades or colors. You may hear digitization specialists using the short-hand "sixteen million colors". The 24 bits are divided into three 8 Bit channels, one for each of three composite colors (Red, Green, and Blue.)

Color Space

Color fidelity is fundamental to accurate reproduction of source materials.

Digitization, faithful to original colors, requires a basic understanding of color and how color reproduction differs from printing technology to digital technology. Fundamental to these differences is the media on which a color image is *printed*.

The color space most commonly used by digitization projects and required by dLOC, is a standardized Red/Green/Blue (*sRGB*) color space.

Choosing the Appropriate Bit Depth and Color Space

1 bit   8 bit   24 bit

dLOC recommends that 1 Bit imaging should **not** be used. 1 Bit images, even at very high resolution (see, *Resolution* below), tend to *pixelate* text. Imperfections on the page or artifacts of age may read as *black*, obscuring text in 1 Bit images. In the 1 Bit page image above, bleed through from the text printed on the inverse page as well as artifacts of age obscure the text. Obscured text will introduce imperfections that reduce the accuracy of text conversion by optical character recognition (OCR) software.

The 8 Bit grayscale image above captures the textual information. And, the reader of the page can make sense of the text.

Readers of Latin religious texts, such as that seen above, will recognize the red text as instructions to the faithful, commentary on the spoken text of a religious service, or the narrative of the priest as opposed to that of the congregation's response.

dLOC advocates preserving *meaningful* color. Meaningful color is color required to interpret the text. In the case of a newspaper with colored images, a color image accompanying an article demonstrates meaningful color, while a color advertisement may not.

It is true that "*The greater the Bit Depth the greater the size of the digital image file*". However, **digitization technicians are encouraged to produce images that meet the reader's needs rather than the needs of the digitization technician to conserve space.**
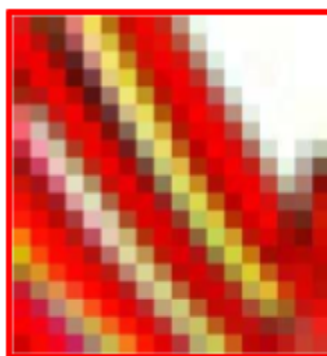
Resolution

The resolution of digital images is expressed in terms of pixels. A pixel is a picture element or, simply, a block of solid shade or color that, together with other picture elements, comprises a digital image.

Trinidad and Tobago's Coat of Arms   (*Zoom area in black box.*)

RESOLUTION	USE FOR
OR 300 pixels per inch (ppi) 118 pixels per centimeter (ppc)	Printed text with *normal* sized fonts Oversized documents and maps Manuscripts with legible script
OR 600 pixels per inch (ppi) 236 pixels per centimeter (ppc)	Photographs and select graphic arts Printed text with *very small* fonts Manuscripts with difficult scripts

The dLOC's minimum digital resolution standard for printed text with normal sized fonts is 300 pixels per inch (ppi) or 118 pixels per centimeter (ppc). This threshold is based on both the characteristics of printed graphics and optical character recognition (OCR) tests.

300 ppi / 118 ppc

The Rationale for Printed Graphics

In general, the resolution of printed graphics does not exceed 300 dots per inch (dpi) or 118 dots per centimeter (dpc). Dots per inch/centimeter are rough equivalents of pixels per inch/centimeter; so comparison is appropriate.

Carifesta '72 logo printed in Guyana's Sunday Post and Weekend Argosy (*Zoom area in red box.*)

Graphics printed in newspapers, for example, often have 80 to 100 dpi (32 to 40 dpc). Most graphics in magazines are printed with 120 dpi (48 dpc) print resolution while graphics in high-end magazines and on post-cards are printed with 300 dpi (118 dpc) print resolution.
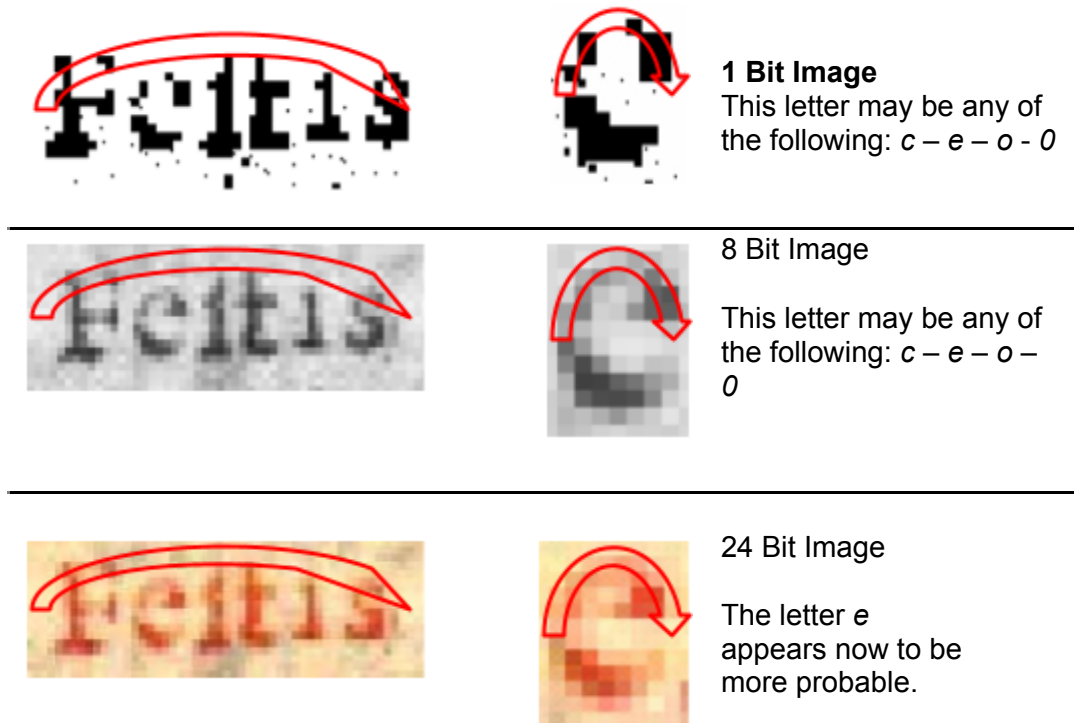
Digitization of printed graphics at resolution greater than 300 ppi (118 ppc) would be excessive.

The Rationale for Optical Character Recognition (*Text Generation*)

When a document page is digitized an image of the page is created. All text page images sent to the dLOC's central servers are subject to Optical Character Recognition (OCR). OCR is a process by which page images are converted to searchable text. Several OCR programs are in common use. Most are optimized for the conversion of images digitized with 200, 300, 400 or 600 ppi (80, 118, 158 or 236 ppc). Images created with other resolutions can be converted to searchable text but, generally, with less accurate results.

Resolution and OCR Accuracy in high contrast images			
75 ppi Image	**150 ppi Image**	**300 ppi Image**	**600 ppi Image**
OCR results			
L ~iC ud L~ddPa	Label C of d Laid Pa	Label C of d Laid Pa	Label C of d Laid Pa

The Importance of Bit-Depth on Text: Example of recognition for the Latin word *Feltis* (Goodness)

1 Bit Image
This letter may be any of the following: *c – e – o - 0*

8 Bit Image

This letter may be any of the following: *c – e – o – 0*

24 Bit Image

The letter *e* appears now to be more probable.

dLOC central servers use enterprise-level OCR software, configured with multiple OCR engines to ensure a high level of accuracy in text generation. For printed texts with normal size fonts, whether plain (*sans serif*) or embellished (*serif*), tests demonstrate that the average modern printed document is accurately recognized at 200 ppi (80 ppc).

dLOC sets a slightly higher standard, 300 ppi (118 ppc), for printed texts with normal size fonts to compensate for occasional uses of small fonts or colored, aged (discolored), or blemished paper.

Digitization of normal printed texts at higher resolution (e.g., 600 ppi/236 ppc), in tests, generally showed no increase in text conversion accuracy. 600 ppi/236 ppc images result in higher conversion accuracy only when the source document is printed with very small fonts.

600 ppi / 236 ppc

dLOC recommends digitizing at 600 ppi (236 ppc) only when working with printed texts with very small fonts; photographs and other continuous-tone graphics, and manuscripts with difficult scripts.
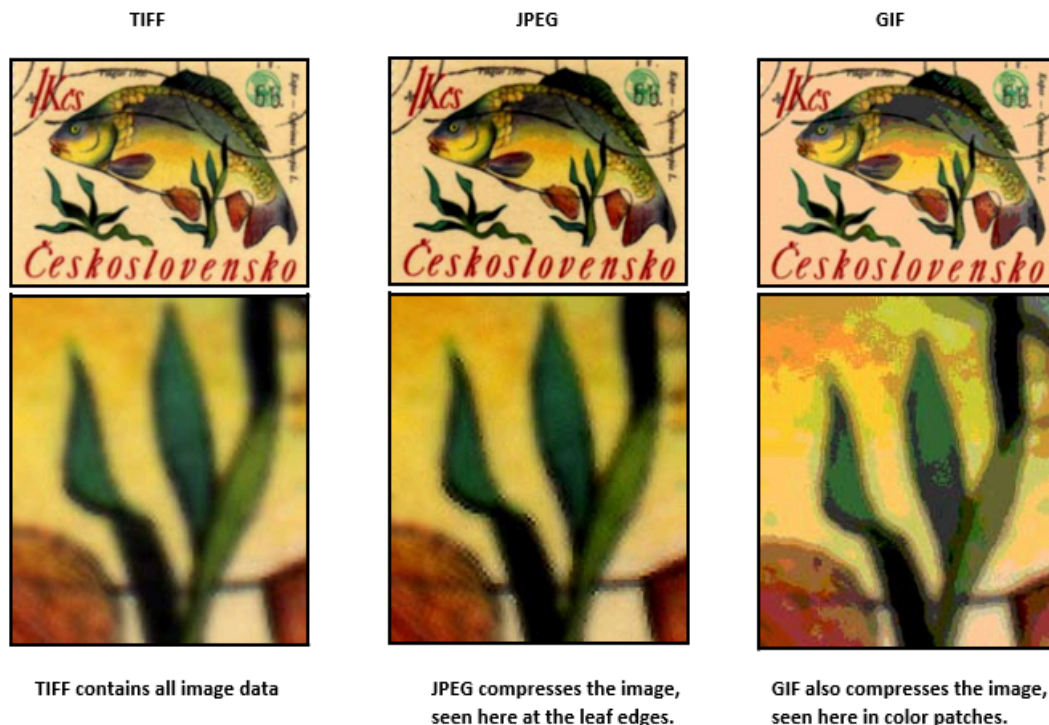
Photographs

Photographs, unlike printed graphics, have continuous-tone. In the source document, one shade or color blends into adjacent shades and colors. Continuous-tone images may be digitized at any resolution. dLOC recommends 600 ppi (236 ppc) resolution to facilitate special uses of images.

Users of digital photographs frequently consult images for their various subjects as for the whole image. A user may want to zoom on the jewelry or hair braids in the photograph of a woman or on shop sizes in the photograph of a street scene. dLOC central servers use JPEG 2000 technology to facilitate zoom. Images digitized at 600 ppi (236 ppc) produce clearer, sharper, and more *readable* images than do 300 ppi (118 ppc) images.

Saving Files and Image Compression

Once created, the issue of saving or archiving the file remains. It is preferable not to lose a quality image to the imperfections of file saving and image compression routines.

TIFF contains all image data   JPEG compresses the image, seen here at the leaf edges.   GIF also compresses the image, seen here in color patches.

Saving Files

Files should always be saved as **uncompressed TIFFs** in order to meet archival standards and ensure the images will display properly within dLOC.

Only the TIFF (sometimes: TIF; *Tagged Image File Format*) is considered archival within the international digital library community. It is the primary digital archival file. There are several reasons for this, primarily: image quality with image compression. The illustration above demonstrates image quality issues as a factor in file choice.

For speed of access online, dLOC creates additional derivative or secondary file formats from the primary archival file. With the digital archival file remains in TIFF, all needs are supported which again demonstrates the importance of saving in the TIFF format.

Image Compression

When sending images to dLOC, provide **uncompressed TIFFs**. Otherwise problems like negative/inverted images result.

Scanning

Color Targets

Creating Directories

Before scanning, create the folder(s) in which you save the scanned images. For each item a separate folder should be created with the appropriate dLOC ID, for example

CA00000001_00001
CA00000002_00001
CA00000003_00001

---

dLOC Requirement for Primary Digital Files

8-bit Grayscale or 24-bit RGB Color (depending on whether original has significant color) 300 dpi for standard text or 600 dpi for stand-alone images (photographs, maps)
Save archival files as <u>uncompressed TIFFs</u>
Folder and file names should not contain spaces of special characters (Ex. @, #, !, *).
Users should use underscores _ in place of spaces.

Flatbed Scanning: Epson Expression 10000 XL

The following screen prints are specific to the Epson Expression scanner, but the same settings apply to any flatbed scanner.

Also, a DLSR setup is good – link to that manual

Scan Settings

Scan documents using ~~Adobe Photoshop~~ rather than using the scanner's stand-alone image capture software and check the scanner settings with each new document. Most affordable software is a lifetime license to ViewScan - $100. The scanner software can also be used….

Each time you scan, turn on the scanner and make sure that the bed is clean and free of any dust, debris, etc. If necessary, clean the glass with a lint free cloth and a very small amount of glass cleaning fluid.

Launch your scanning software

Select: *File* "" *Import* "" *Epson Expression 10000 XL*
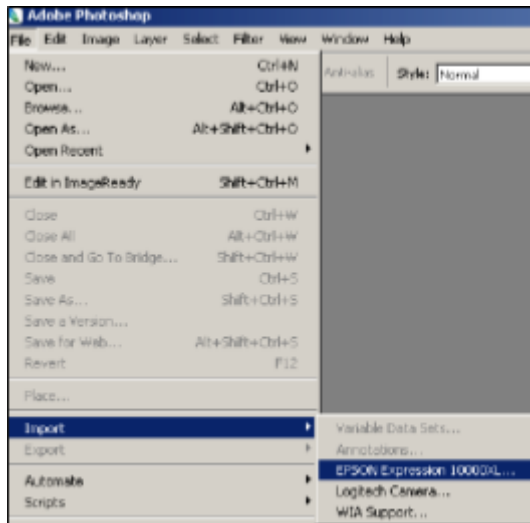
The scanning interface will then open two windows: a scan settings window and a preview window (*as seen here*)
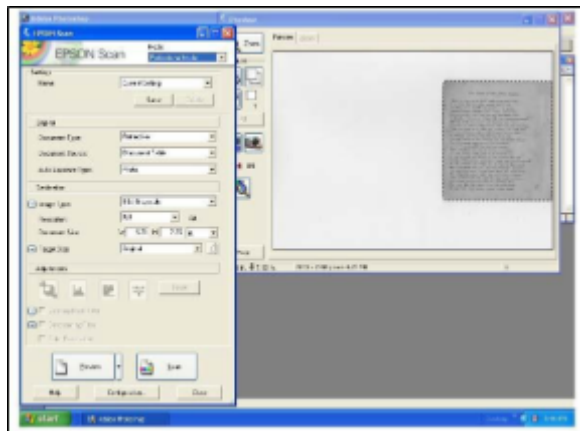
Select the appropriate settings for your document

At the top of the scan settings windows select PROFESSIONAL MODE

Always select the following settings
Documents Type: REFLECTIVE
Document Source: DOCUMENT TABLE
Auto Exposure Type: PHOTO
Document Size: DO NOT ADJUST
Target Size: ORIGINAL

Select the appropriate color space and bit depth:
8-bit grayscale for items without significant color
24-bit RGB color for all other items
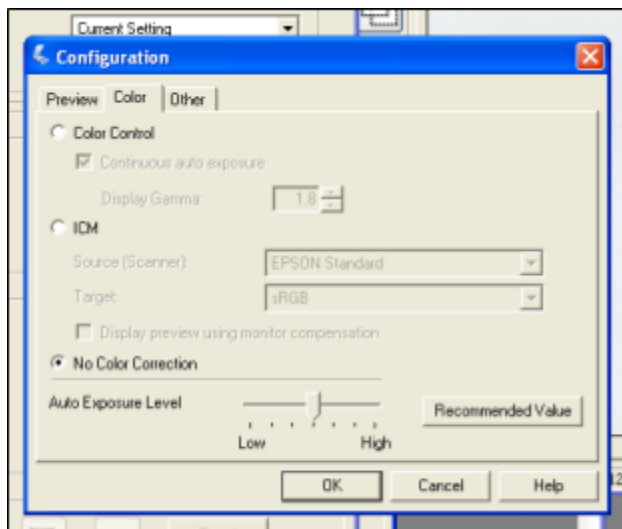
Select the appropriate resolution

300 dpi for mostly textual items
600 dpi for stand-alone image items (photographs, maps, etc..)

Click the CONFIGURATION button (*below the Preview and Scan buttons*)

Click the COLOR tab
Select NO COLOR CORRECTION
Click OK

1. Place item, image down, on scanner glass. Be careful to place item as straight as possible in order to save time later. Close the scanner lid as much as item permits.

2. Click the PREVIEW button in the Scan Settings window. A small preview of your image will
    appear in the preview window. Make sure the entire document is visible, if not reposition on glass and re-preview.

3. Draw a bounding box around your entire image. If your original has 2 pages facing each other, draw a second box by selecting the dual marquee button . Arrange each box to completely

include each side of the item. Once you are satisfied with the boxes positioning click the button. DO NOT move the boxes or change settings after pressing this button!

4. Click the ⬚ button

Saving Files

1. Save your image by selecting: *File → Save*

2. Select the dLOC ID folder that corresponds to which the image being saved belongs. For example: in separate folders with the appropriate dLOC ID:
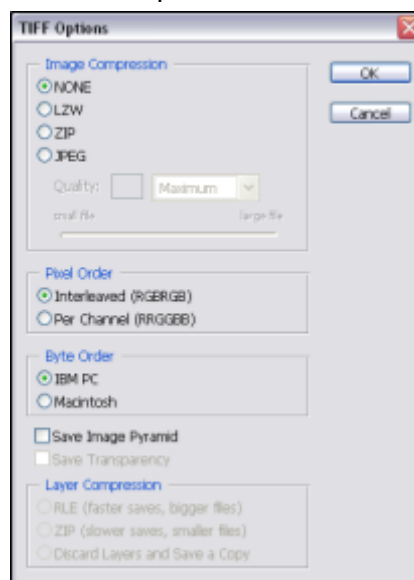   *CA00000001_00001*
    CA00000002_00001
    CA00000003_00001

3. Type in a sequential four digit file name, such as *0001*, *0002*, *0003*, etc.

4. Select *TIFF* from the file format drop down menu

5. Always uncheck the ICC profile box

Color:	☐ Use Proof Setup: Working C⊀
	☐ ICC Profile: Gray Gamma 2.2

6. Click Save
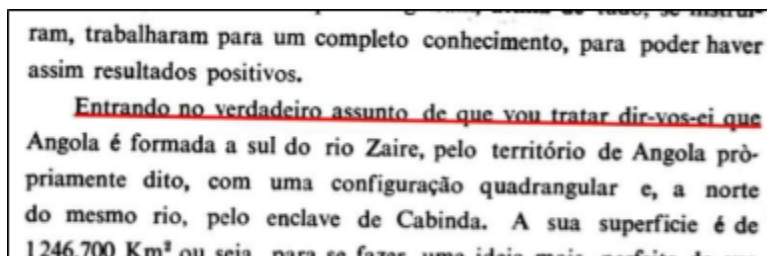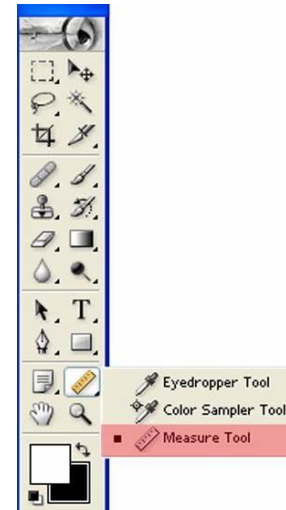
For TIFF Options select same as below

Always review the Other > Type > Format section to ensure images are **NOT** compressed
by default.

Image Correction

The intent of any digitization should be a faithful reproduction of the original document. Toward this goal, images will need to be deskewed and cropped to fit the in-hand original. In addition, it may be desirable to perform color correction either to reproduce the in-hand original, or the original state of the document. Applying these techniques in **Adobe Photoshop** is the topic of this section.

Image Correction in Adobe Photoshop

1. To straighten drastically skewed images:

    a. Click and hold the Eyedropper
       Tool in the Photoshop tool box –
       Select the Measure Tool

    b. Click and draw a line to follow the bottom
       of any printed text, line or image
       (*line is red, here, for purposes of illustration*)

> ram, trabalharam para um completo conhecimento, para poder haver assim resultados positivos.
>
> Entrando no verdadeiro assunto de que vou tratar dir-vos-ei que Angola é formada a sul do rio Zaire, pelo território de Angola pròpriamente dito, com uma configuração quadrangular e, a norte do mesmo rio, pelo enclave de Cabinda. A sua superfície é de 1.246.700 Km² ou seja, para se fazer uma ideia mais perfeita de sua

    c. Select: *Image* """ *Rotate Canvas* """ *Arbitrary* (*DO NOT change the angle*) """ click
       OK

2. Crop the image to remove any excess borders added during
   straightening using the crop tool

3. If necessary (e.g., if the image is muddy), adjust the levels/histogram by selecting
   *Image* """ *Adjustments* """ *Levels*
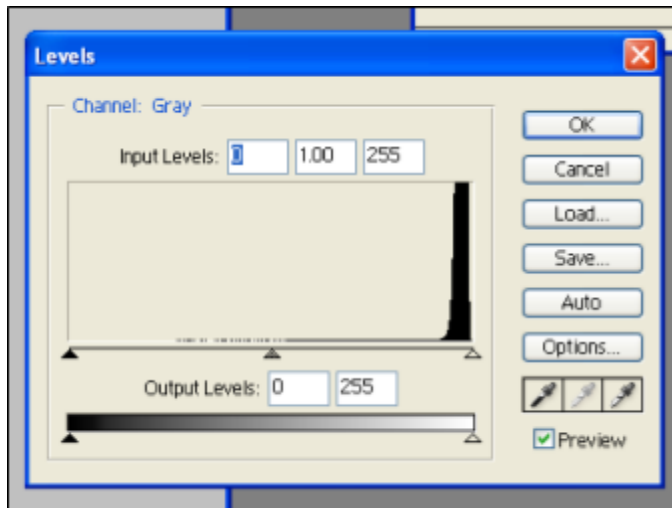
Adjust the black, white and midpoints to improve your image quality and contrast.

If the image is COLOR, you may make histogram adjustments for each RGB channel: Red, Green and Blue.

But, do not over correct and eliminate detail. A histogram shows the distribution of tones over a range. The image characterized by the histogram above is predominantly white. While the image contains shades of gray, deeper tones of black are almost entirely absent.

4. Images with good, thick printed text can also be quickly corrected by selecting the document's white point. This is done by opening the levels/histograms by selecting *Image* """ *Adjustments* """ *Levels*. In the levels window select the eyedropper furthest to the right and then select the point in your image that should be the brightest white. The images below show this effect before and after the white point selection.
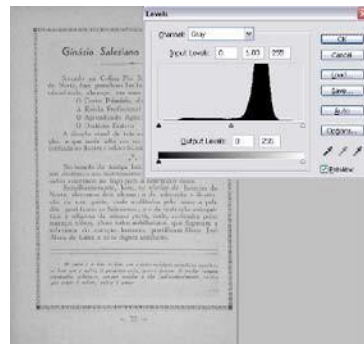

You will notice that the background becomes almost uniformly white, but the text is also lightened. Before selecting OK in the levels/histograms you will need to bring in the black

point in order to improve the text. This is done by moving the arrow furthest to the left, in towards the right. You will notice that the numbers in the Input Levels boxes increase.
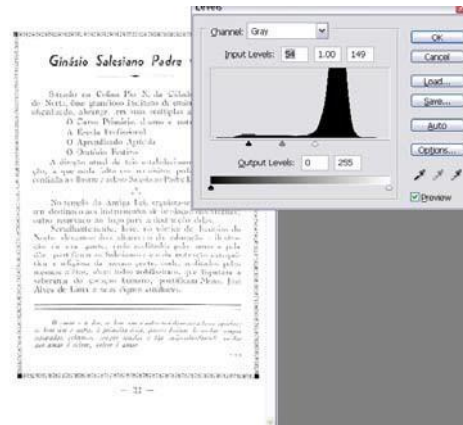
It is helpful to perform this correction while zoomed in to 100% on your image, as shown below.
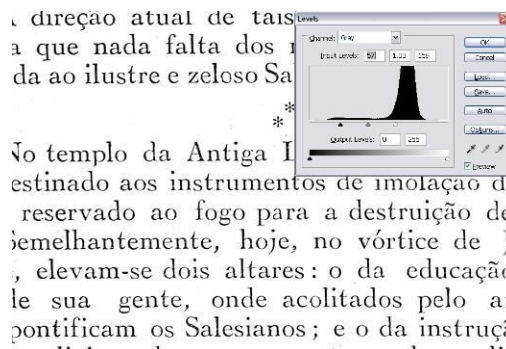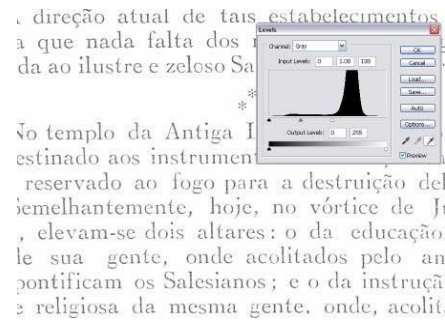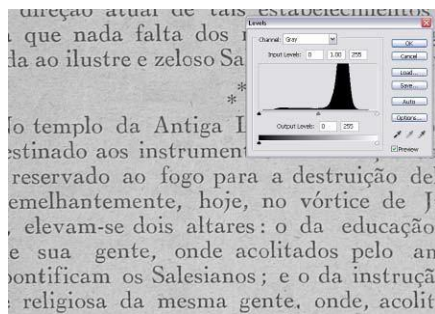
5. If the image is extremely stained the document should be scanned in RGB and if possible, the stains should be lightened using *Image* """ *Adjustments* """ *Replace Color*
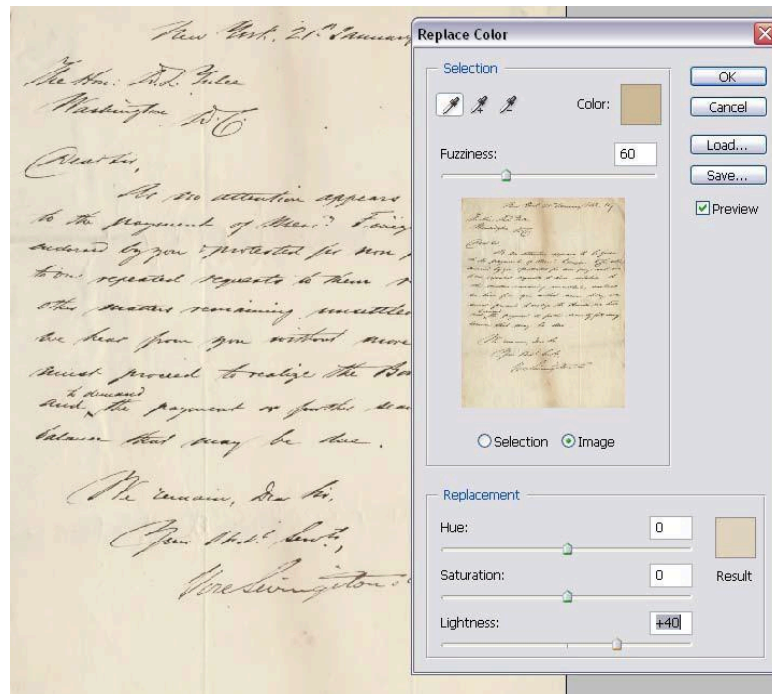
Select "Image" and not "Selection" in the Replace Color Window. Then using the eyedropper tool select the darker color of the stain. Adjust the Lightness, Saturation and Hue slider bars as needed to minimize the stains. The fuzziness meter indicates how closely a color must match the selected color to be replaced. Be aware that stains may be similar in color to text and therefore too much manipulation is undesirable in order to not lose information.

Often it is useful to zoom into one section of text while performing the color replacement. One must be careful not to make the text harder to read for the OCR engine.

6. Remember that any adjustments done to images can be undone as long as the file remains open. Maintain your history window open by selecting *Window* """ *History* in Photoshop, then simply select the previous step done. You can always go back several steps and re-correct your image.

Other Image Editing Resources

The original Adobe Photoshop installation package should include a tutorial of the software you purchased. In addition, Adobe has an on-line resource at the following URL: https://helpx.adobe.com/photoshop/tutorials.html

Adobe, the Adobe Logo, and Photoshop are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries.

Additional Resources

This manual is continually updated, and is always available from this permanent link: http://dloc.com/AA00002865

Resources Guide: https://guides.uflib.ufl.edu/dloctech

Appendices include:
Guide to Reformatting Audio: http://dloc.com/AA00002865/00002
Guide to Reformatting Video: http://dloc.com/AA00002865/00003
Guide to Requesting Permissions: http://dloc.com/AA00002865/00004
Appendix: Photoshop Shortcuts: http://dloc.com/AA00002865/00008

dLOC Metadata Guide: http://dloc.com/AA00002864

DLSR manual

Training Videos