BHL Digital Imaging Specifications

This document outlines the digitization best practices and minimum digital imaging standards of the Biodiversity Heritage Library (BHL) for libraries in its consortium of partners. Individual institutions may employ higher standards according to available digitization equipment and resources.

Page images available via the BHL website (http://biodiversitylibrary.org) are served directly through the Internet Archive. If Internet Archive experiences service interruptions, page images will not be available on the BHL website and some download features may be temporarily unavailable.

Doc link: http://s.si.edu/BHLImagingSpecs (copy/paste recommended)

Updated 4/26/18

Table of Contents

Digital Imaging Best Practices

Digitization through the Internet Archive

BHL digital imaging standards meet or exceed Internet Archive's imaging standards. Many institutions within BHL's consortium work with their local IA scanning center or IA produced "table-top" machines to digitize materials. IA outputs are as follows:

PPI and Size

i. For Mark 1's the chart is to the right and shows both the book and the PPI settings. These are chosen to optimally capture a given size book.

PPI	Height (inch)	Width (inch)	Height (cm)	Width (cm)
300	16	9.25	40	23
400	10.75	6.75	27	17
500	8.75	5.5	22	14

ii. For Mark ii's the chart is to the right.

PPI	Height (inch)	Width (inch)	Height (cm)	Width (cm)
650	8	5.5	20.3	14.0
500	10.5	7.5	26.7	19.1
350	15	10.5	38.1	19.1

Image Capture

i. The Scribe machine currently captures page images with two digital single lens-reflex (DSLR) cameras, specifically the Canon model 5D, Mark I - 12.8 mega-pixel camera (http://is.gd/IuQc) and the Canon EF 100mm f 2.8-macro lens (http://is.gd/IuRG). In some locations a Canon model 5d, Mark ii – 21.1-mega pixel camera is being used. IA may in the future evaluate and test newer camera models as they come onto the market to determine if they will provide similar or better performance. Note: there is a possibility that some dust particles may be captured from the glass or lens during photographic process. While IA will try to reduce and eliminate this, this might occur and IA cannot be responsible for this. There is also a possibility that on some photographs, particularly in yearbooks, a Moiré effect may occur. IA can't control this and this not part of our standard QA or process control.

ii. The lighting system used for book illumination consists of eight (8) 5000 or 3500 Kelvin, 36 degree, and 35-watt museum-grade Solux bulbs and provides a smooth daylight spectrum with a high color-rendering index. If future alternative lighting methods are found to provide similar or improved results, changes to the lighting system may result.

iii. Please note that since there are two independent cameras in use, there may be a detectable difference in lighting between the recto and verso images.

iv. Reference targets: a color target (such as a ColorChecker 24) are shot at the end of each book as reference tools and may be used for ICC-based color management.

v. Image transfer: images are downloaded in real time to a Scribe management/image-processing computer. This computer is also responsible for running the camera management software that operates the camera shutters.

Equipment Calibration

i. Scribe station frames are calibrated and aligned before being put into use.

ii. Cameras are calibrated per manufacturer's specifications. Cameras that no longer perform within specifications are immediately sent to the manufacturer or repaired in-house.

iii. Kelvin light bulbs used in the digitization process are replaced as necessary. Lights are allowed to stabilize for up to 15 minutes before image capture.

Digitization using “in-house” equipment or scanning vendor

BHL recommends the Federal Agencies Digitization Guidelines Initiative (FADGI) Technical Guidelines for Digitizing Cultural Heritage Materials for text-based materials. BHL partners digitizing materials using “in-house” equipment or contracting digitization services through a vendor are required to upload the digitized materials to the Internet Archive via Smithsonian Libraries and Archives “Macaw” software. All materials must be deposited in the Internet Archive in order to be served through the BHL website.

BHL strives to provide a “faithful rendering of the underlying source document” including completeness, image quality (tonality and color), and with the ability to reproduce pages in their correct (original) order such that a legible printed facsimile could be produced in the same size as the original [FADGI Still Image Guidelines p.51].

FADGI still image guidelines recommend the *Benchmark for Faithful Digital Reproductions of Monographs and Serials. Version 1. December 2002* as a minimum standard.

Minimum Benchmarks for Page Image Masters		
Black and white For text, and may also be used for line drawings, de-screened halftones.	**Grayscale** For covers and illustrations printed in black and white. Recommended, but not required.	**Color** For covers, and meaningful text or illustrations printed in color. Recommended, but not required.
600 dpi, 1-bit or bitonal TIFF images 3. Images must be sized and saved at 1:1 scale to the dimensions of the original page. Images must be saved uncompressed or with lossless compression. Where	**300 dpi, 8-bit grayscale uncompressed TIFF, or lossless compressed image (e.g. LZW, JPEG2000).** Images must be sized and saved at 1:1 scale to the dimensions of the original page. The dpi specification will relate directly to the font-size and page dimensions of the	**300 dpi, 24-bit color uncompressed TIFF, or lossless compressed images (e.g. LZW, JPEG2000).** Images must be sized and saved at 1:1 scale to the dimensions of the original page. RGB and YCC are the recommended color spaces for masters, particularly when only one master version is produced. The dpi specification will relate directly to the font-size and page

images are compressed they must be made available in the Group 4 (ITU-T6) format. The images may be interpolated from 400 optical dpi 8-bit images.	original source document, and to local definitions of legibility and fidelity. In many cases, 400 dpi will be preferred. Where larger pages are concerned, the lower dpi specification may be required.	dimensions of the original source document, and to local definitions of legibility and fidelity. It may also relate to the perceived artifactual value of the source object or the extent to which its physical characteristics such as foxing, etc., are perceived of as conveying some important information or meaning.

http://old.diglib.org/standards/bmarkfin.htm

What to Scan

BHL digitizes books or volumes cover-to-cover meaning that every page image, including covers and blank pages, are digitized. If there are more than 10 blank pages in a row (e.g. "filler") you may stop scanning after the 10th page and resume scanning at the next page with content, or the back end papers, whichever comes first.

The first image scanned can be of the background with colorbar or other calibration instrument - this is useful to keep the page "hand" (recto/verso) correct. Alternately, a colorbar can be placed as the last image in the sequence.

Create images with **one book or volume page per image**, unless the content on the page spans the gutter, as in a notebook or scrapbook, a two-page spread illustration, or in a foldout. Some primary source materials such as field notebooks can be imaged as two-page spreads regardless of whether the text spans the gutter or not.

Foldouts, tissue, and inserts

Foldouts and two page spreads should be handled in the following way to preserve the page order (right, left, right,left) : spreads should have a blank page inserted either before or after, unless there is another two page spread immediately following that will enable preservation of the right/left order. For foldouts, which are typically found on the recto (right side) of the page, convention is to show the folded-up foldout, then a blank “filler” page, then the unfolded foldout, then the verso of the folded (or unfolded - up to you) foldout. The goal is to show all the information contained on the pages and preserve the page order.

Tissue should not be scanned unless it contains information, e.g., overlayed text. In this case, the page should be scanned with the tissue over the underlying image, then scanned again with the tissue rolled back, then (to preserve page order) a blank "filler" page should be scanned.

Inserts - tipped in (attached) inserts should be scanned as if they were a standard page. For inserts that are not tipped in, it is at the discretion of the scanning institution whether or not to scan them. Obviously, inserts relevant to the text should be scanned.

Cropping, De-skewing

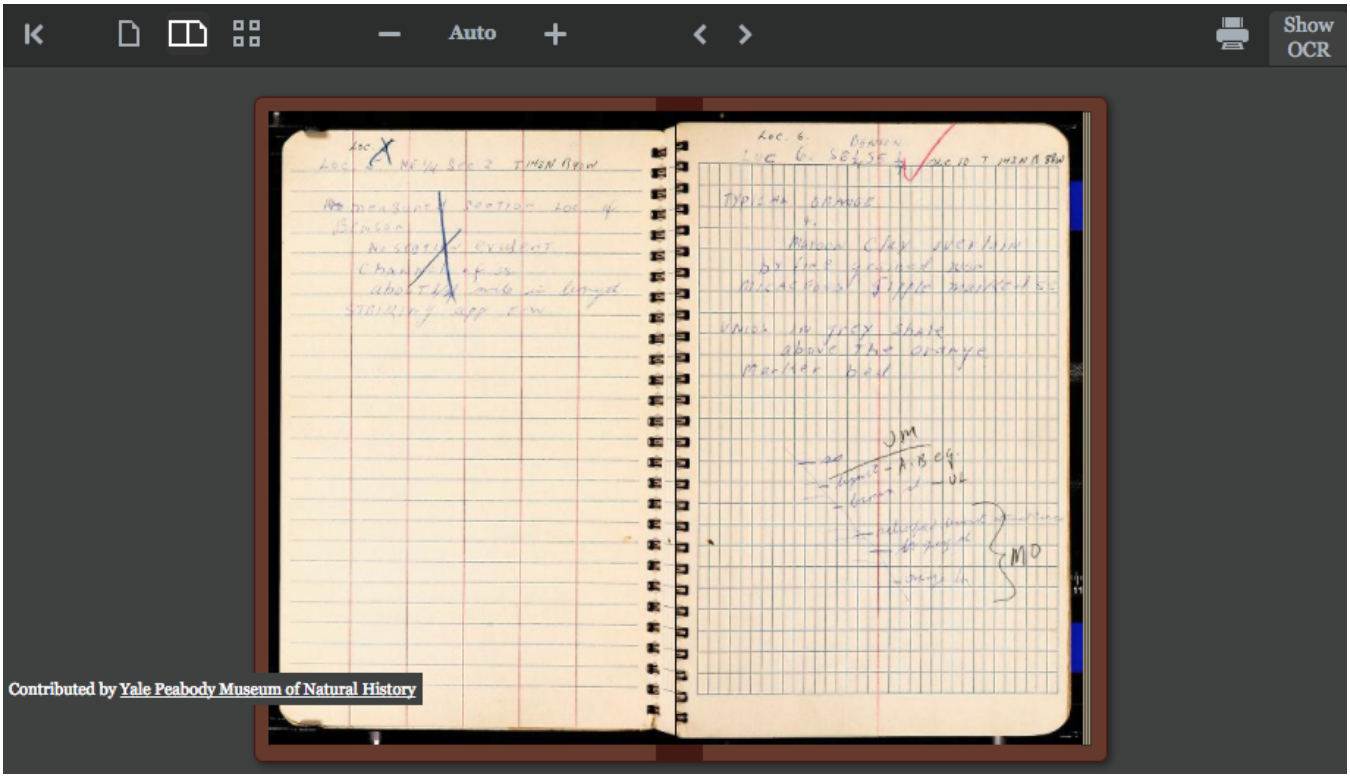
General Collection Items

The images can either be cropped just inside the edges of the page (as Internet Archive does) or just outside the edges of the page to show the entire page has been digitized. General collection materials are *almost always cropped just inside the edges of the page* as there is little need to preserve the artifactual value of the object being digitized. The preference to crop just inside the edges of the page image helps the digital book appear seamlessly integrated within the BHL Book Viewer.

On occasion, there may be a need to pursue uncropped pages for general collection items. In cases where text, images, or annotations run off the edges of the pages, cropping should be just outside the page. In addition, there may be cases where third-party vendors or “boutique” digitization operations produce images with cropping outside the page image. Use your best judgment in determining if further cropping of the page image is required as the labor involved to further crop images may significantly increase turnaround times for submitting content to the collection. Please see our Cropping Outside section for details.

Archival Items

Regarding archival materials, it is a best practice to photograph page images on a black or dark gray background, and *crop outside the edges of the page* as a way of demonstrating the artifactual value of the content within context of the object itself, such as a piece of correspondence or field notebook. Furthermore for collections of archival materials within a folder, Smithsonian Libraries recommends that, “each manuscript page in a folder be photographed using the same focal length, in order to provide the correct sense of scale for all items in that folder, unless a very small item would be unreadable if scanned at a longer focal length.”



Cropping Outside

When cropping outside the edges of the page, do ***not*** include excess background, color-bars or other calibration devices as part of the page image. Harvard Library Imaging Services, for example, recommends no more than a 10mm band outside the page. University of Illinois at Urbana-Champaign's Digitization Services recommends an ⅛ to a ¼ inch border.

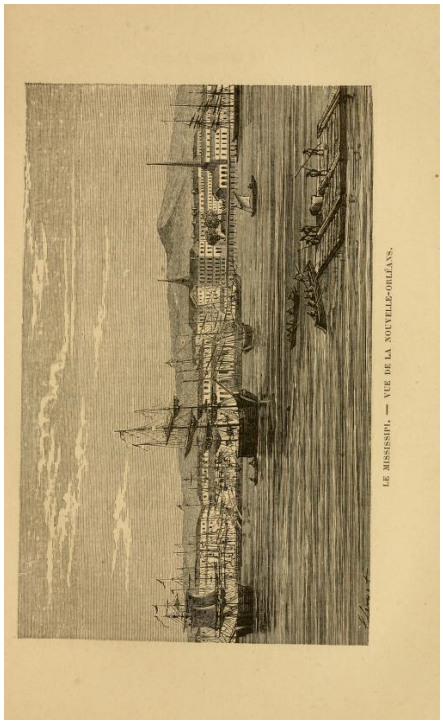
Deskewing

All images should be de-skewed (rotated) to align the text on the page perpendicular to the length of the page, such that OCR can be done efficiently. (Optional for manuscript material, which should only be de-skewed to maximize legibility.)

For incunabula and other unique texts where the text-block and page shape *really* do not line up, exercise your judgment as to how much or if the image should be de-skewed.

Page Orientation

Page orientation should be maintained as the book or volume was published where possible. Keep vertical orientation for horizontal images on a single page such as:



Except in cases where fold-outs or multi-page spreads are *primarily text based* and horizontal

orientation is necessary for the Optical Character Recognition (ORC) software to “read” the text.

Other file processing

This is at the discretion of the scanning institution. In general, BHL is interested in a faithful rendering of the original, and legibility such that the text can be OCR'd with as much accuracy as possible. There should be no need to do color-correction in your software or other touch-ups if your cameras are regularly calibrated.

Formats & Technical Specs

TIFF (preferred file format)

- uncompressed
- 300ppi or better, proportional (1:1) to the size of the original, e.g., a book 28cm on the long edge at 400ppi will be 4409px on the long side of the image
- 24-bit color
- if you *must* scan in grayscale, you can do so, though it is strongly discouraged. Use 8-bit grayscale and the highest ppi you can achieve.

JPEG2000

- uncompressed lossless JPEG2000 is preferred, though a minimally (< 15%) compressed lossless JP2 will work if that is all that is available
- 300 ppi or better, proportional to the original.
- 24-bit color

PDF (not preferred)

- while it is possible to upload PDFs directly to Internet Archive, BHL cannot ingest and display PDFs. It is possible to use Macaw to *process* PDFs to then upload to Internet Archive and ingest into BHL, but those PDFs must be:

- Color (preferably) or grayscale - bitonal PDFs will not work
- of good quality (e.g., originally scanned and saved at 450ppi or better)

Filenaming Conventions

All files should be named using a unique identifier and a "counter" number to keep the images in the same order as they were in the original. Most BHL consortium libraries use the unique identifier of the book (e.g., barcode or catalog record number) followed by an underscore and a four-digit counter then the file extension. If your material does not have unique identifiers, use a portion of the title or author with the year of publication to create a unique identifier (example below uses Internet Archive - generated filenames.)

```
ResearchOnMollu1972Fi_0001.tif

ResearchOnMollu1972Fi_0002.tif

ResearchOnMollu1972Fi_0003.tif
```