

University of Colorado Herbarium Partners in Existing Networks “Lichens and Bryophytes: Sensitive Indicators of Environmental Quality and Change” Workflow.

The goal of the Lichen and Bryophyte PEN project is to capture images of label data for our North American specimens so they can be analyzed using optical character recognition. Once the label data has been captured it is uploaded to the University of Wisconsin so the images can be read using optical character recognition (OCR). The University of Colorado Herbarium (COLO) Lichen and Bryophyte Collection was not sorted geographically and did not have a barcode or global unique identifier (GUID). Specimens housed at COLO are stored in drawers in individual specimen packets (the collection does not use herbarium sheets for any of the lichen and bryophyte collection). Label data is affixed to the front of each packet and annotation labels are generally affixed to the back of each packet (Figure 1).

Workflow

The initial workflow started by separating North American and non-North American specimens for any species that occupied two or more drawers in the collection. Specimens from Colorado were also separated and marked with blue index cards to keep continuity between the vascular and non-vascular collections.

Pre-Imaging

Species from the target area are removed from the collection and moved to a staging area. At this point a barcode (aka GUID) is placed on the bottom right corner of the back of each packet (Figure 1). To minimize wear and reduce the potential of the barcode being removed accidentally, the GUID is not placed on the edge of the packet. While the GUID is being applied to each specimen, specimens are also further separated geographically to help facilitate skeletal data capture at the imaging stage. COLO opted to capture geographic information while imaging to create database that can immediately handle both taxonomic and broad scale geographic queries. Initial results indicate that we can barcode specimens at a rate of approximately 300 specimens per hour without sorting geographically in the process. Sorting does slow down the process of applying GUIDs to approximately 225 specimens per hour, however, sorted specimens are imaged at a faster rate because fewer changes need to be made to collect geographic skeletal data. When batches of specimens are from the same area, e.g. Boulder County, Colorado geographic and taxonomic information pass from one record to the next allowing the imager to capture taxonomic and geographic location for all specimens without the need to type or change settings. This is particularly helpful for areas of the collection that are well represented such as Colorado and Utah where there are often multiple specimens from the same county.

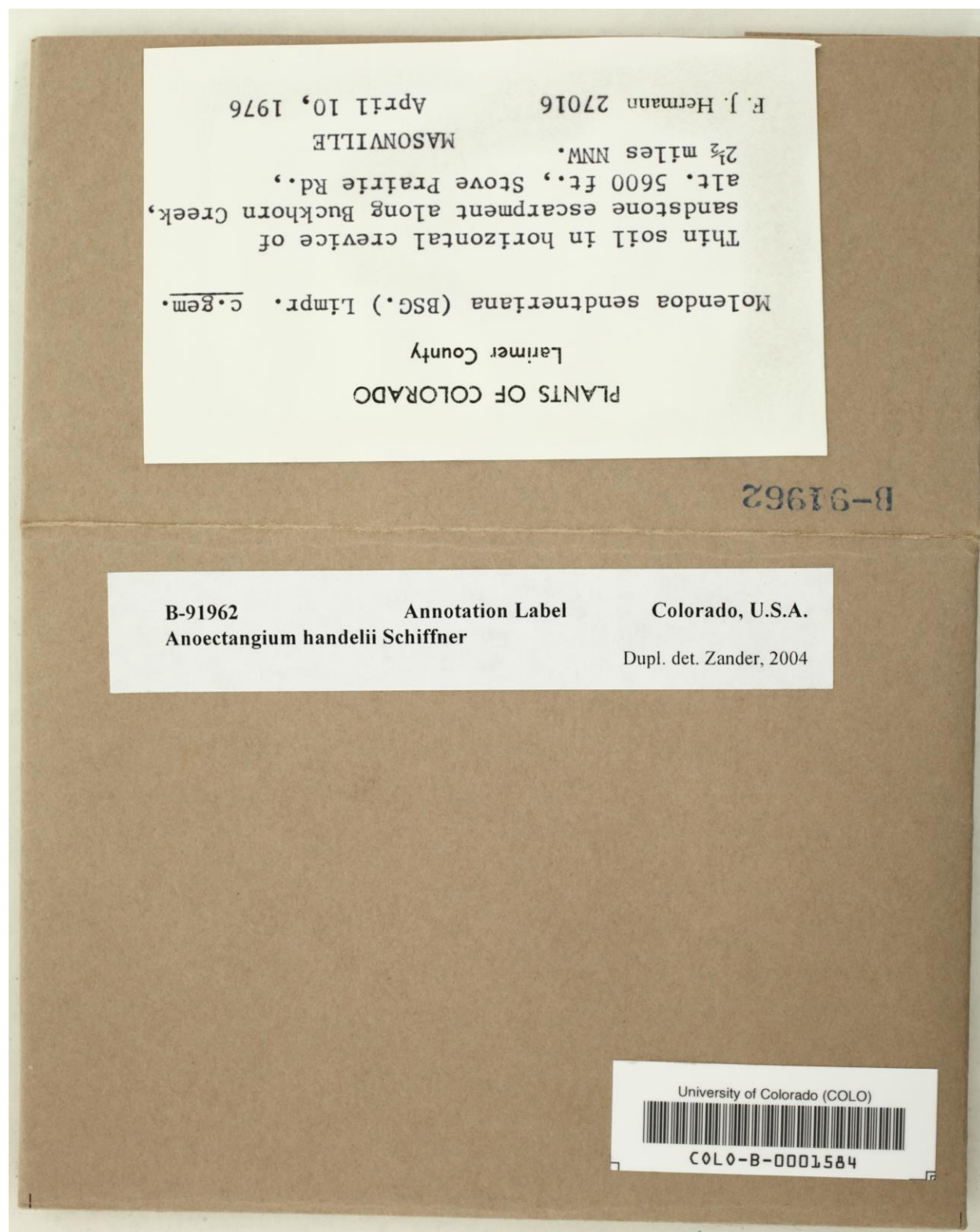


Figure 1. A typical packet from COLO with an annotation label affixed to the back and a GUID/Barcode in the lower right corner set in from the corner.

Imaging

COLO uses a copy stand with a mounted Canon EOS 5D Mark II camera and an Ortech PhotoBox. Images are captured as JPEGs to conserve storage space and to optimize the images for OCR. Images that have too high of a resolution can cause problems in the OCR process. Because the majority of COLO's annotation labels are adhered to the back of packets we are able to capture both the original and annotated specimen information in a single image. Images are captured with the original collection information upside down with the orientation of the text being corrected during post process using Photoshop.

The imaging station is set up using a guide packet to maintain consistency between images. Under the camera live view mode there is the ability to bring up a grid to break the field of view into equal squares (Figure 2). Setting the grid view to 2 X 2 allows the user to determine the middle of the image. The guide packet should be positioned to reflect the middle of the image. This ensures that future batch processing such as cropping will find the same line on the specimen.

Specimens that are sorted geographically during the barcoding process can be imaged at a rate of approximately 100 per hour. This is an increase from approximately 80 images per hour when specimens are not organized. The increase in imaging rate is attributed to a more streamlined skeletal data capture process and the ability to persist information from one record to the next.



Figure 2. Using a custom 2 x 2 allows the imager to quickly find the center of the image and allows the use of the grid to assure the fold of the packet represents the center of each image captured.

COLO has found that images taken with the settings (seen in Figure 3) seem to give good results with optical character recognition. Images that are relatively bright without being washed out seem to run through OCR better than dark images.



Figure 3. Very dark and very bright washed out images do not OCR well. COLO uses the settings above to create a balance between the two extremes and seems to produce images that OCR well.

Capturing Skeletal Data

Skeletal Data is captured while imaging using a Java application written by Robert Anglin at the University of Wisconsin (Figures 4-7). The first step in setting up the tool is to point the application to the folder where images are being saved. This is done in the “Working Folder” tab (Figure 4) of the application. The tool utilizes a processing folder where images are stored by Digital Photo Professional before the tool renames the images and a destination folder where the images are moved after they are renamed and skeletal data are captured. COLO images specimens into the Public folder on the C drive so images can be accessed by any user.

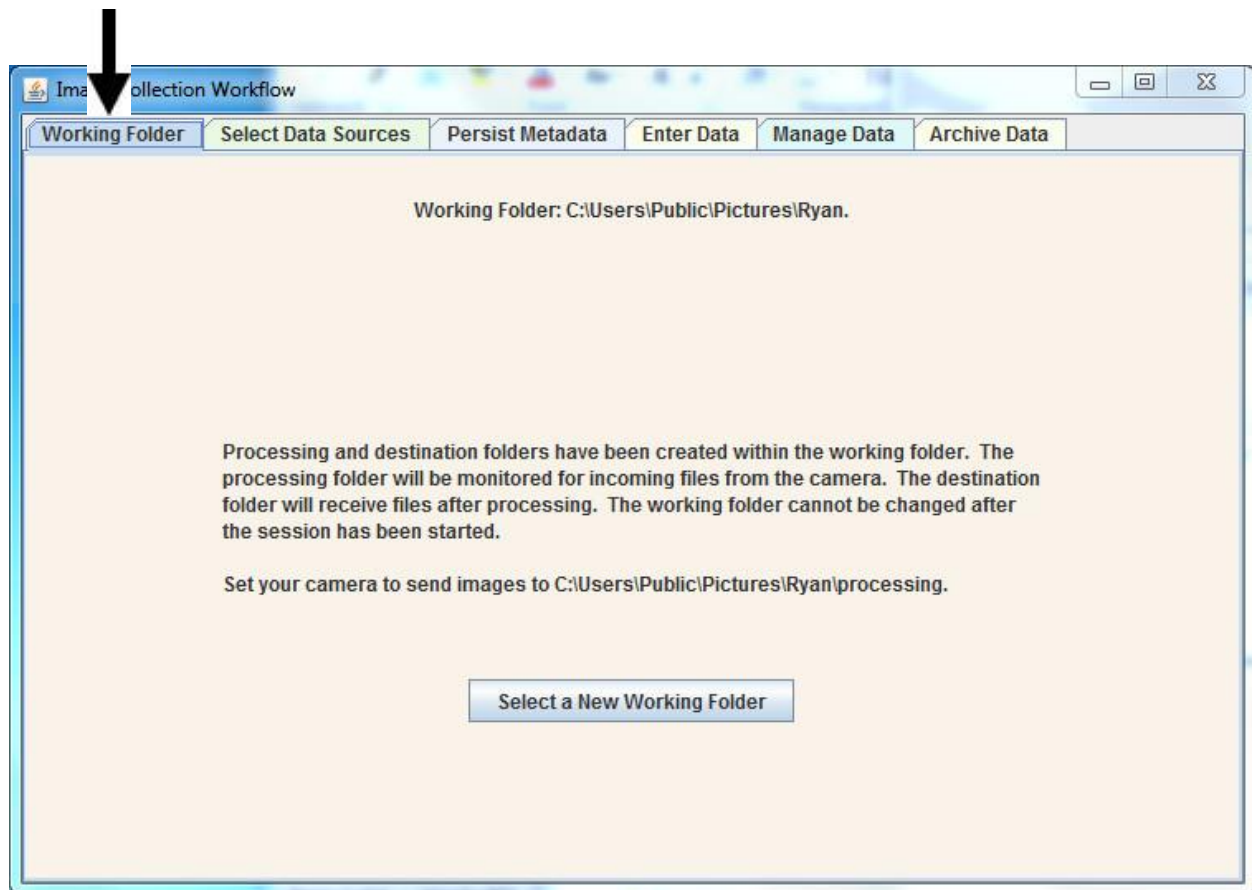


Figure 4. Use this tab to select the folder where images will be saved prior to being renamed using the application.

The Imaging Workflow Application is set up to allow for end users to populate the taxonomy dropdown menu with any .tab file they choose. For this project we are using two aggregated lists, one for lichens and one for bryophytes. Select the appropriate file for the collection you are currently imaging using the "Upload a New Name File" button and selecting either the lichen or bryophyte list (Figure 5). This only needs to be done at the start of each collection. The taxonomy files will automatically load after they are set the first time. The data source section dealing with exsiccate is not applicable for COLO as exsiccate have been separated and are filed taxonomically.

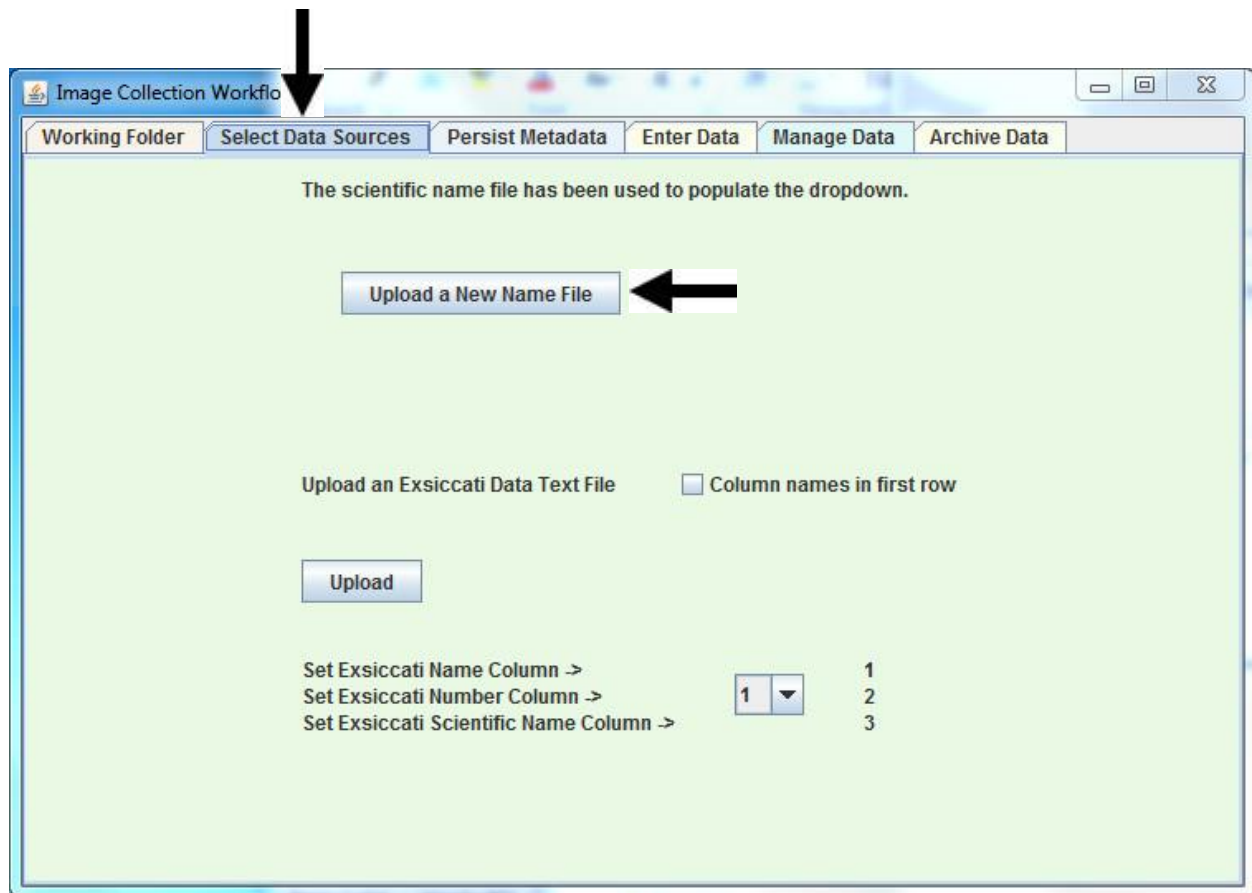


Figure 5. Use this tab to upload the .tab file that contains the list of lichens or bryophyte scientific names.

The persist metadata tab is where the real power of the application becomes apparent especially when specimens have been sorted geographically. The “Persist Metadata” tab (Figure 6) allows for information to be applied across multiple specimens only requiring the user to make changes as the geography of the specimens change. This tab also allows for imager name, institution code and name to be applied to every specimen. At COLO we set the tool to persist scientific name, country, the state or province field and county.

Image Collection Workflow

Working Folder | Select Data Sources | **Persist Metadata** | Enter Data | Manage Data | Archive Data

Collection-specific Metadata (persists between sessions)

Image Recorder Name: Ryan Allen

Institution Code: COLO

Institution Name: University of Colorado

Collection Code:

Collection Name:

Select metadata fields to persist between images

☐ Collection Date ☒ Country ☒ State
☐ County ☐ Collector ☐ Collector Number
☒ Scientific Name ☐ Exsiccati ☐ Exsiccati Number
☒ Label Type ☐ ID Qualifier

Figure 6. This tab allows for the imager to select what metadata will be applied to each image. Collection specific metadata will appear on every image across different sessions unless it is changed. The bottom half of the interface allows the imager to select what field information will persist from image to image.

The “Enter Data” tab (Figure 7) is used to populate the skeletal file. Since scientific name, country, state or province and county were set to persist from record to record the imager will only need to enter this information as it changes. At the end of the day, images are moved from the imaging computer to the server. Post processing takes place on a different computer. Moving the images ensures they are backed up and allows them to be accessible on the post processing computer. We recommend editing the “.tab files” for geography to contain only areas that are applicable to the project. This can be done by opening the .tab files provided with the imaging tool and removing locations that are not applicable to the project. Collecting geographic information does go above the minimum standards for the collection of skeletal data but we have found that imagers tend to be able to maintain focus for longer periods of time and generally find the imaging process less tedious.

Image Collection Workflow

Working Folder | Select Data Sources | Persist Metadata | **Enter Data** | Manage Data | Archive Data

Barcode

Scientific Name

ID Qualifier

Collector

Collector Number

Collection Date

Country

State/Province

County

Exsiccati

Exsiccati Number

Label Type




Click "New Session" to begin monitoring the processing folder

New Session

Figure 7. This is the interface where all of the metadata is captured. Any item checked in the persist metadata tab will carry over from specimen to specimen. Persisted information can be changed at any time by overwriting the information contained in the field. i.e. if the imager changes the county from Boulder to Clear Creek before submitting the image the next image would be prepopulated with Clear Creek for the county field.

Post Imaging

Before images can be uploaded to the server they need to be processed using Photoshop so all of the text is facing the same direction. Photoshop allows for bulk processing of images using a combination of Photoshop Actions and automation tools. Actions allow the user to set up a set of predefined processes to be run on an image. In our case we set Photoshop to select the top half of each image and freeform rotate this portion of the image to create a single image that includes both the label information and the barcode. This process is then associated with a Droplet that can be placed on the computer desktop.

To create an Action, navigate to Window and then Actions (Figure 8) to bring up the Actions options. To create a new Action click the  button near the bottom right of the Actions options (Figure 9). COLO rotates, selects the top half of the image does a freeform rotation and saves the image using an Action (Figure 10). To start recording the Action click the  button, and to complete the Action click on the  button. A summary of what steps are taken in the Action can be seen in Figure 11.

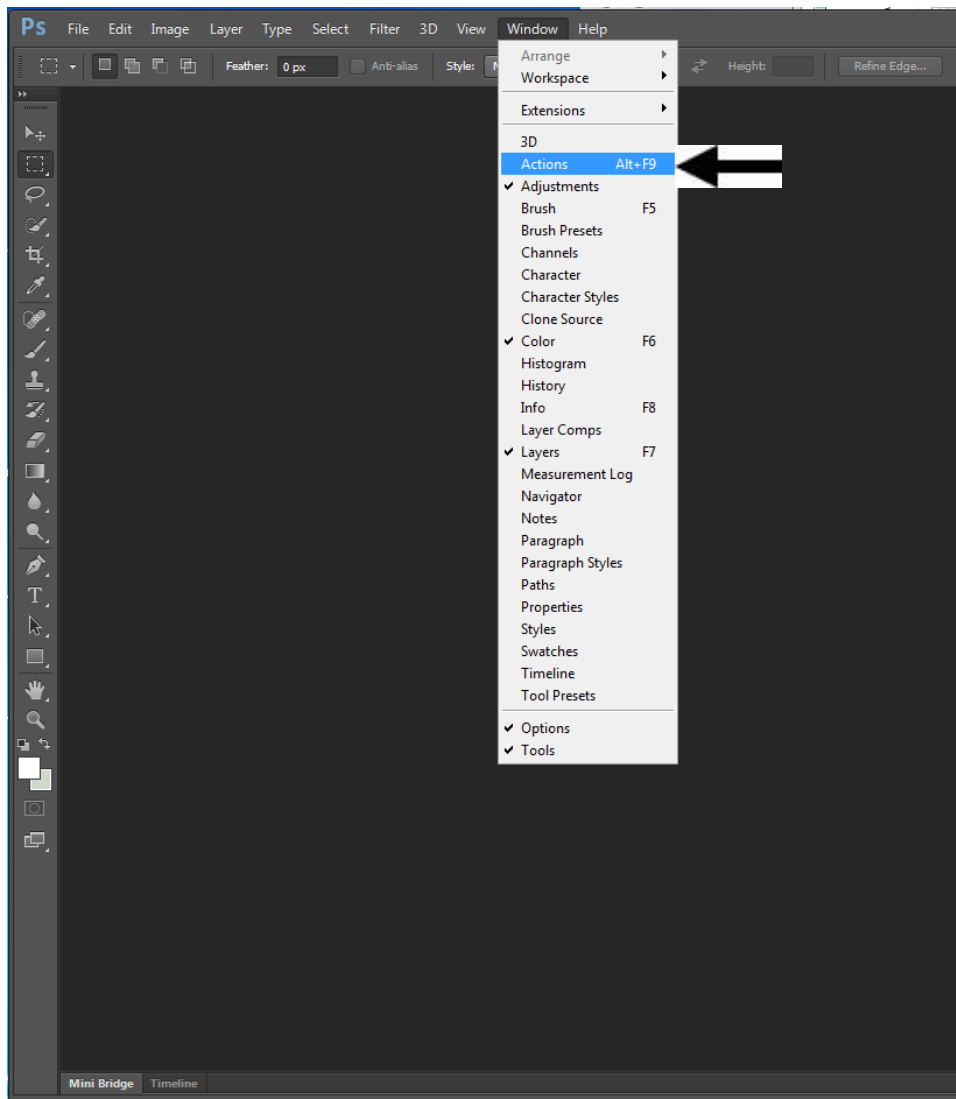


Figure 8. To launch the Actions function navigate to the windows tab and select Actions.

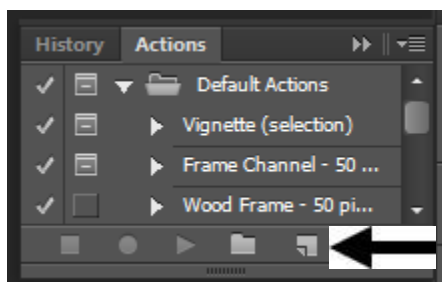


Figure 9. Clicking on the folded paper icon will start the process of creating a new Action.

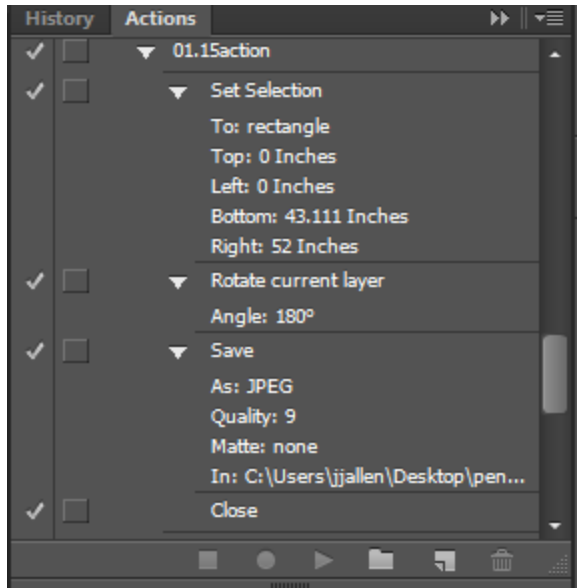


Figure 10. Suggested Action setting to locate the center of an image crop it and rotate the top half of the image. This also directs the image to be saved in a specified processed folder.

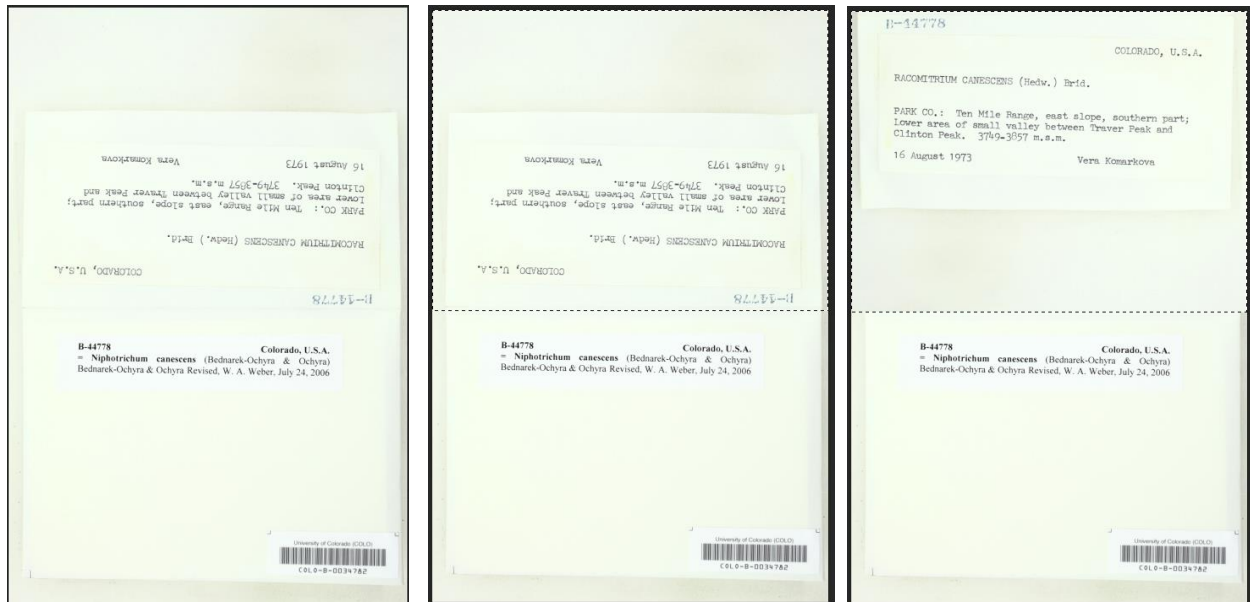


Figure 11. Represents a visual of the steps in the Action process used by COLO. The top half of the image is selected using an Action and the top half of the image is rotated using a free form rotation so all text on the image is facing the same direction.

To facilitate processing images a Droplet can be created on the desktop to run the Action without the user needing to open Photoshop (the program will still open but it will do so as a result of starting the Droplet). To map a set of Actions to a Droplet navigate to the File Tab select Automate and finally select Create Droplet (Figure 12). Folders of images can be “dropped” onto the Droplet icon and will initiate

the series of instruction for each image in the folder. The skeletal file does not have to be removed from the folder before processing. COLO saves its Droplets to the desktop for easy access (Figures 13 and 14).

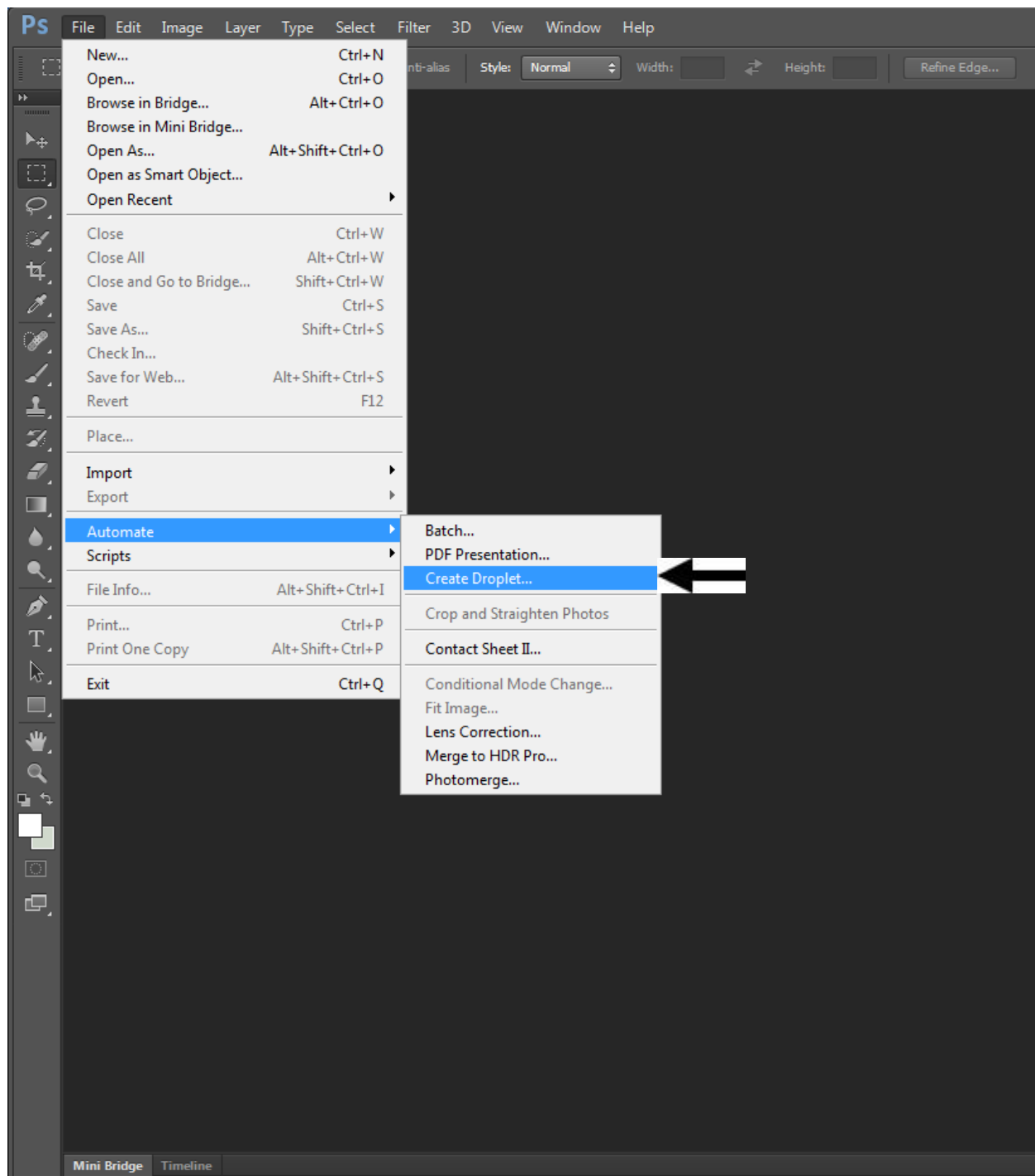


Figure 12. Storing Actions to Droplets allow images to be batch processed by simply dropping a folder on a Droplet icon.

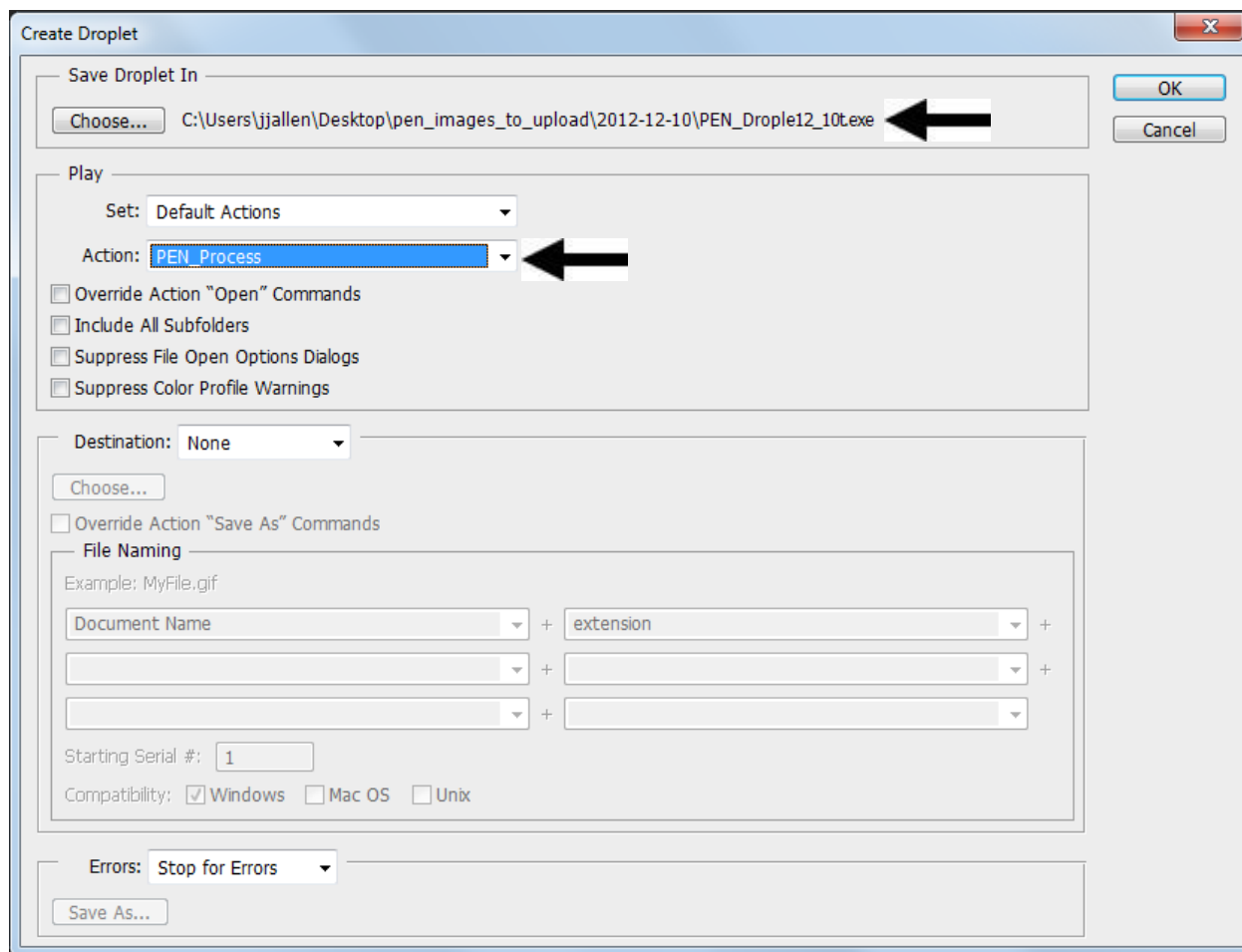


Figure 13. Saving the Droplet to the desktop allows for easy access without the need to open Photoshop and selecting the Action.

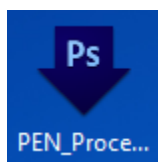


Figure 14. After creating the Droplet it will appear on the desktop and you can drag and drop folders of images onto the Droplet to perform the Actions.

COLO uses a few different Droplets and Actions to facilitate post processing. Periodically COLO Specimens have annotation labels affixed to the inside of packets (Figure 15). Specimens with interior annotation labels are imaged twice; once for the outside of the packet and once for the inside. The inside image should always be the second one captured. The Imaging Tool will automatically assign a “_a” to the end of the barcode name when two pictures are captured for the same specimen; both images need to be captured before hitting the “Enter” button or the second image will just overwrite the main image (Figure 16). COLO also uses a Droplet and Action that selects and rotates the top half of

the image just below the center line. This is helpful for situations when the image is not perfectly aligned or an accession number is stamped right at the packet fold (Figure 17).

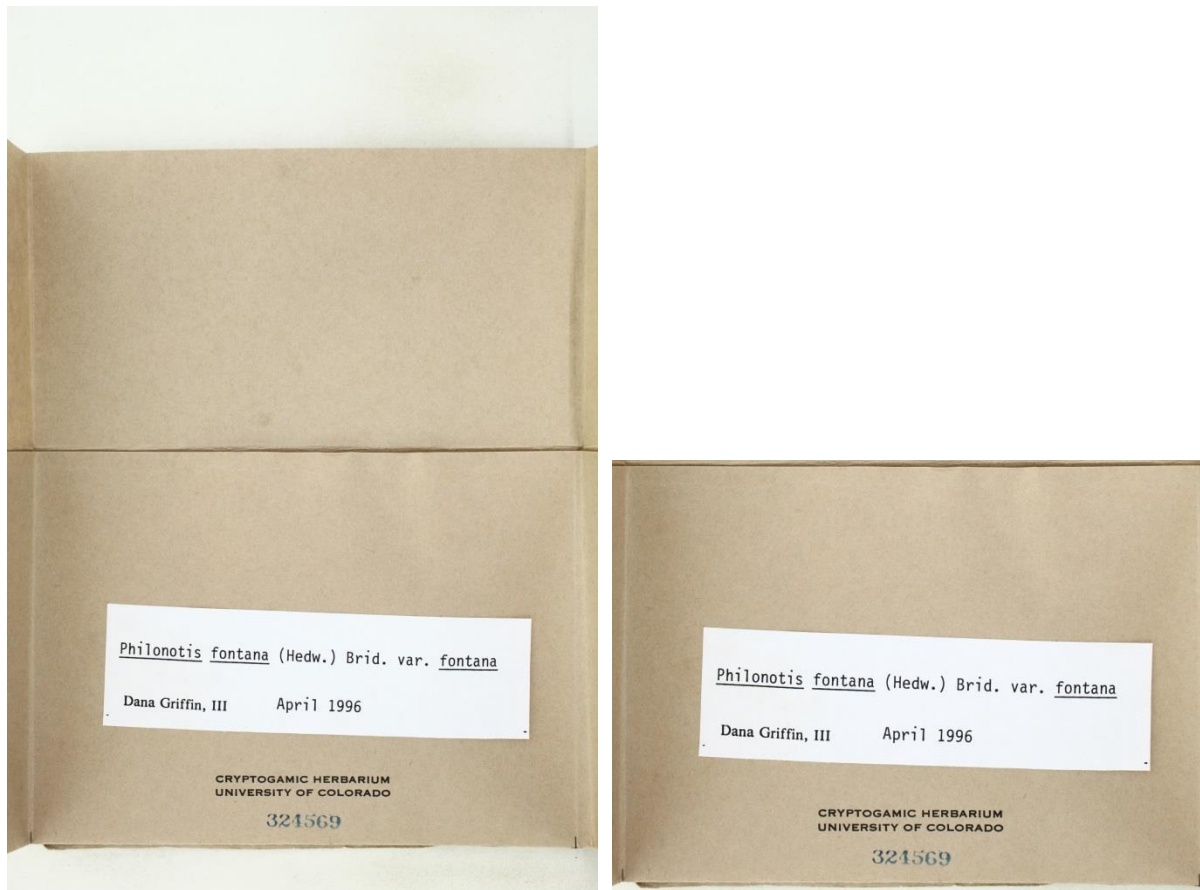


Figure 15. The inside annotation Droplet selects the bottom half of the image and crops the image to include only the area that contains information.

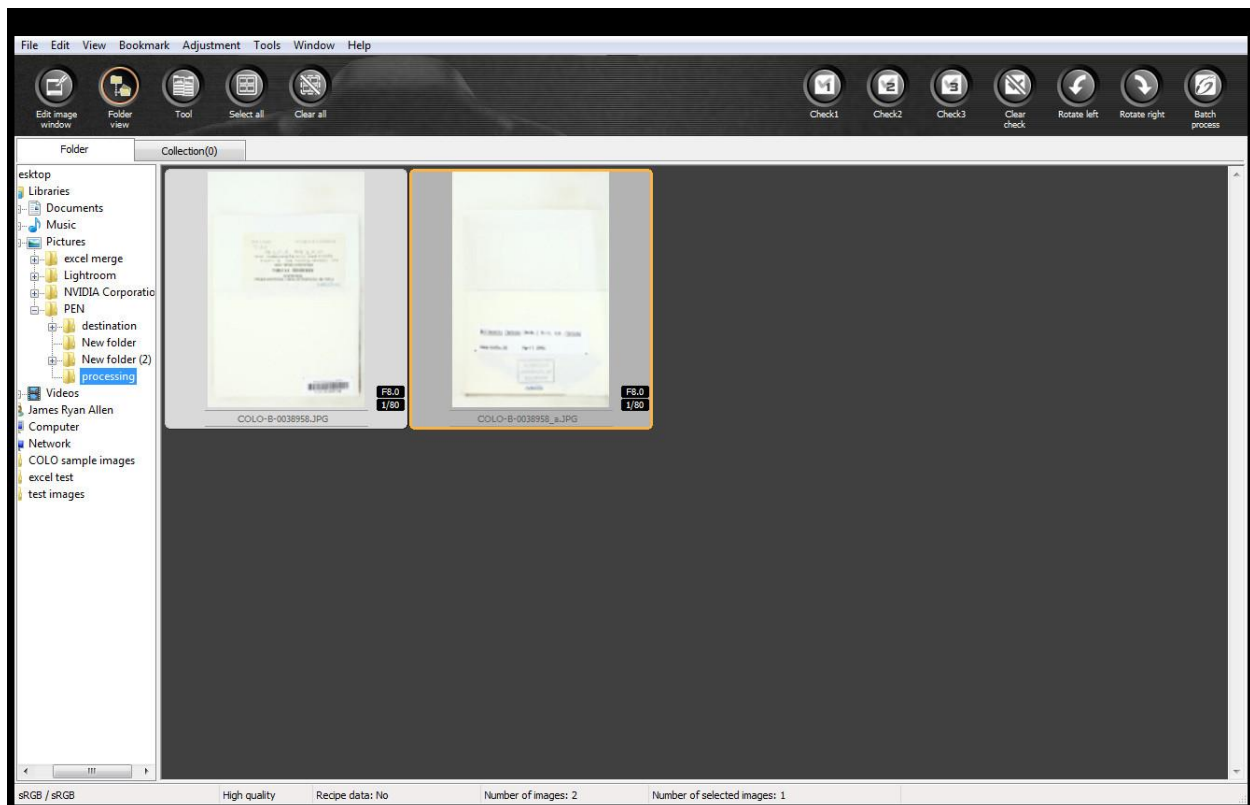


Figure 16. In many cases the specimen cannot be documented in a single image. If there is an interior annotation label or a label that will not fit into the standard field of view two images of the specimen need to be captured.

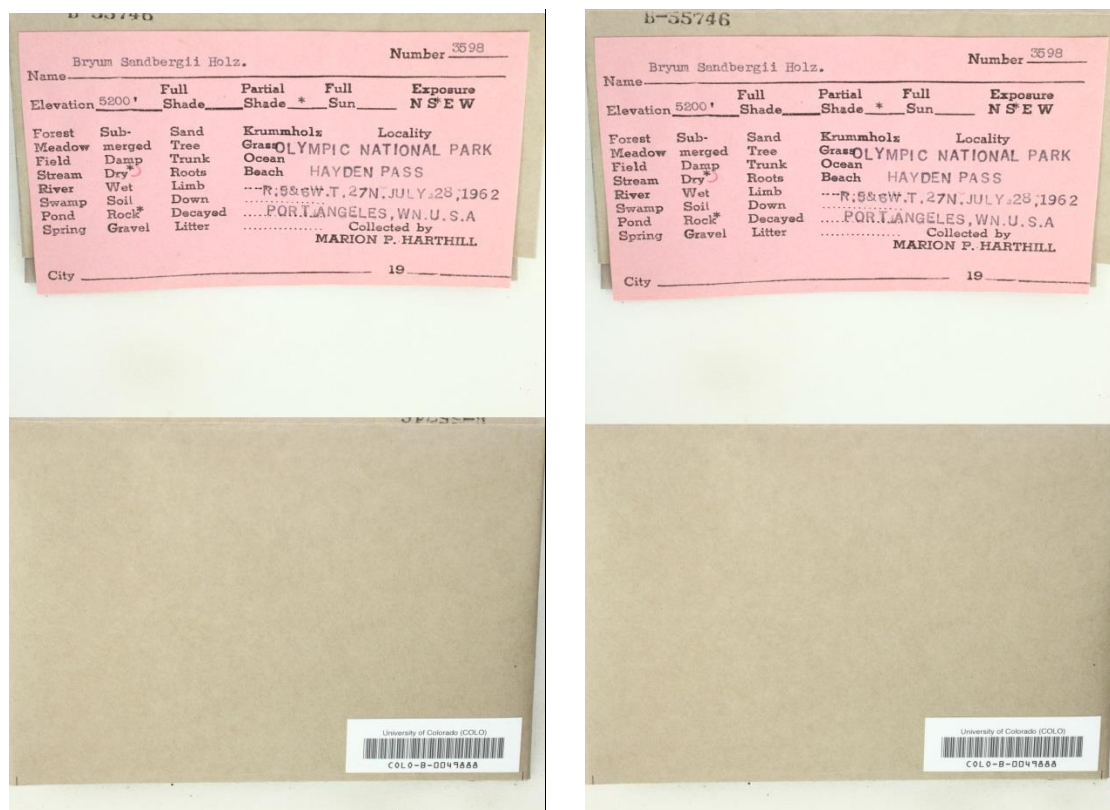


Figure 17. Periodically the Droplet will not select and rotate the image as desired. If the specimen is not aligned properly or there is an accession number along the fold of the packet a different Action should be used to better process the image.

Image submission

Image data are submitted to the Consortium of North American Bryophyte and Lichen Herbaria through FTP transfer. After images have been processed using Droplets they are visually inspected for quality and double checked to make sure non-standard images have been processed correctly (interior annotations, oversized labels etc.). COLO uses FileZilla to create ftp transfers to iDigBio. After images have been transferred they are backed up to a hard drive. COLO keeps a copy of both the original image and the post processed image on a server and a second copy on a dedicated hard drive. Keeping the original image allows for potential future iterations of images without the loss of image fidelity associated with processing JPEG images.