Wolfsonian Public Humanities Lab (WPHL) Digital Archive

3-29-2024

Image Processing & OCR

Rebecca Bakker

Wolfsonian Public Humanities Lab-FIU

Follow this and additional works at: https://digitalcommons.fiu.edu/wphl-da

Image Processing & OCR

Authored by Rebecca Bakker

The main goal of digital preservation is to produce a digital replica that closely mirrors the original material in its current state. It aims to capture not only the image's quality, tonality, and color but also the context and essence of the original item. The ideal approach for digitization is to work with the original material - be it a slide, negative, photo, or document. If the original isn't accessible, always utilize the highest quality substitute. After digitizing, store the materials in a high-resolution uncompressed format (for images, typically a TIFF). For a detailed walkthrough on digitization techniques, please refer to "Chapter 4 - Digitizing Objects."

Ensuring accurate digital representation requires a workstation with a high-quality monitor capable of displaying the appropriate color space. This ensures the master copy's colors and tones closely match the scanned item. Significant alterations should be designated for access versions or other specific projects. In some digitization projects, using color calibration targets during scanning proves valuable (see sample below). However, these are usually cropped from the final image to avoid confusion by the viewer.

Post-scanning image adjustments should be approached with caution. Basic tweaks, such as adjusting brightness or contrast, are permissible. Avoid changes that might misrepresent the item's cultural or historical significance, like adding artificial colors or saturation, over-sharpening or blurring, or erasing inherent blemishes such as dust, stains, writing, or scratches.

When processing images, opt for software like Adobe Photoshop known for lossless processing. Avoid less powerful photo editing software or tools that come with Windows or Mac computers, as they can introduce complications only seen once an image is accessed online via streaming or embedded online viewers. Suitable image processing, especially for the access copy meant for digital repositories, might include:

- Cropping
- Color and Light Balance
- Rotation
- Deskewing
- Stitching

The following sections provide detailed instructions and step-by-step guides for using Adobe Photoshop (version 2022) and methods for batch processing these procedures.

Cropping: Cropping refines the borders of an image to concentrate on the central content, eliminating any superfluous or distracting areas. The goal is to retain the full quality of the original object, with a slight margin from the scan backdrop. This method demarcates the item's edges and sustains the original proportions, avoiding any changes in the aspect ratio. While cropping's purpose can range from eliminating scanner artifacts to accentuating particular sections, such as in a scrapbook, it's ideal to include a "border zone" around the scanned image, generally 1/4-inch of the background surface. Typically, black or white borders are chosen, with a preference towards a black background for textual pieces. Upholding a uniform cropping approach across a collection is important to ensure the digitized version faithfully represents the original material.

Cropping in Photoshop:
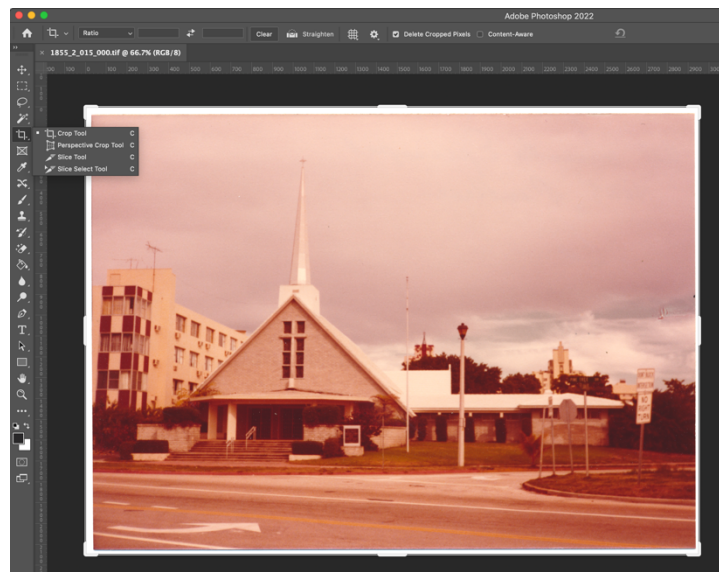
- **Open Image**: Launch Photoshop and open the image you wish to crop.
- **Select Crop Tool**: On the left toolbar, click on the "Crop Tool" (shortcut key: C).
- **Define Crop Area**: Click and drag the corners or sides of the crop box to define the area you want to keep. Remember, you're selecting the area you want to retain, not remove.
- **Maintain Aspect Ratio**: If you want to preserve the original aspect ratio, press and hold the **Shift** key while dragging the corners of the

crop box. Alternatively, you can select a predefined ratio from the top menu bar, such as 1:1 (square) or 4:3.

- **Reposition Image**: Click inside the crop box and drag to reposition the image if necessary.
- **Refine Crop**: Adjust the corners and edges as needed until you're satisfied with the selected area.
- **Apply Crop**: Once you're satisfied with your selection, press the **Enter** key (or **Return** on Mac) to apply the crop. Alternatively, click the checkmark icon on the top menu bar.
- **Save the Image**: Go to **File > Save** or **File > Save As** to save your cropped image. If you want to retain the original, choose "Save As" and provide a new name or save in a different location.

Color and Light Correcting: Adjustments in color, lighting and tone should be based on color calibration targets used during digitization to provide a consistent reference point. While tools like Adobe Photoshop's auto tone, color, and contrast adjustments may seem like quick fixes, they often act globally and can compromise image data inadvertently and thus while tempting, may not always provide the desired results. Many of them are designed with modern color photographic images in mind, assuming a standard tone and color distribution that may not be suitable for all types of older or historic digitized materials. Instead, opt for individualized corrections that offer more control and feedback, like the color balance and curves adjustments. Curves and color adjustments allow users to fine-tune the tonal range and color balance of an image by modifying the shape of individual color channels. Before making color or light corrections, always retain a backup of the original image.
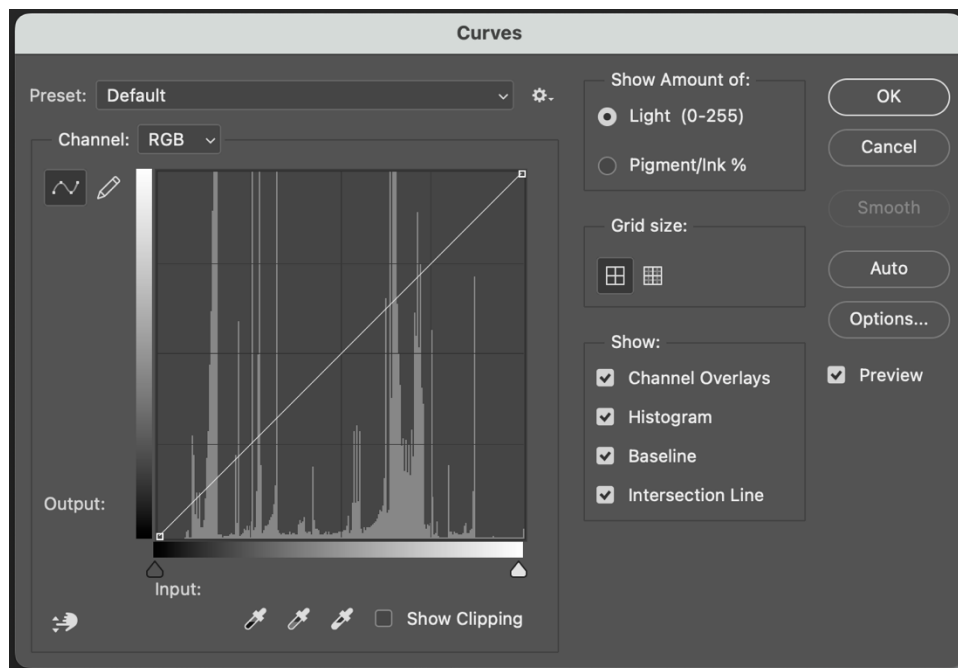
Color Balance in Photoshop:
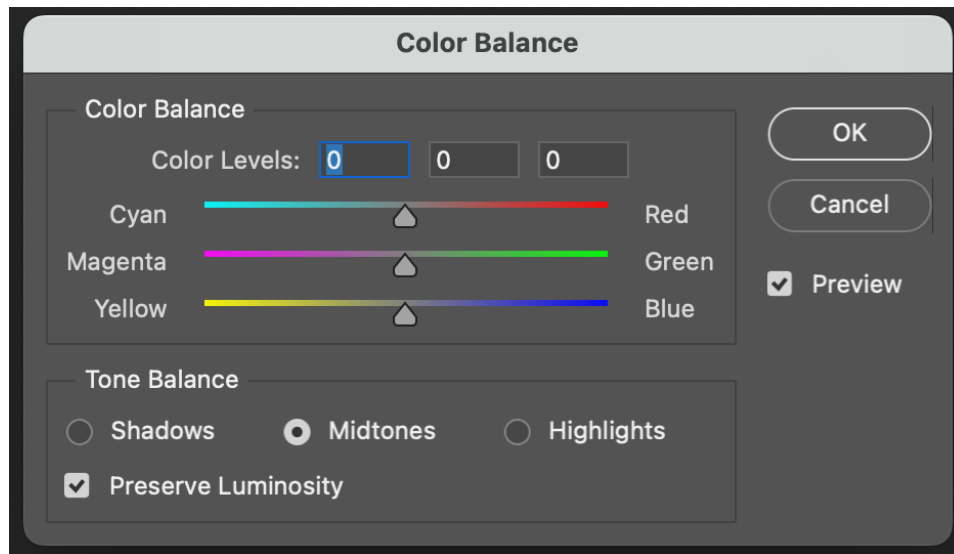
- **Open Your Image**: Launch Adobe Photoshop and open the image you want to adjust.
- **Access the Levels Tool**:
  - Navigate to **Image** in the top menu.
  - Hover over **Adjustments**.
  - Select **Levels**, or use the shortcut **Ctrl** + **L** (Windows) or **Cmd** + **L** (Mac).
- **Adjust Individual Color Channels in Levels**:
  - In the Levels dialog box, click on the dropdown menu where it says **RGB**.
  - Select each color channel (**Red**, **Green**, **Blue**) individually.
  - For each channel, adjust the three sliders (shadows, mid-tones, highlights) until the histogram spikes are spread out and the image looks balanced. You can also adjust the output levels at the bottom if necessary.
- **Use the Color Balance Tool for Finer Adjustments**:
  - Go to **Image** > **Adjustments** > **Color Balance** or use the shortcut **Ctrl** + **B** (Windows) or **Cmd** + **B** (Mac).
  - Adjust the sliders for **Shadows**, **Mid-tones**, and **Highlights** to refine the colors in your image. You can toggle between them using the radio buttons at the bottom of the dialog.
- **Check with the Eyedropper Tool**:

- o Select the Eyedropper Tool from the toolbar.
  - o Click on what should be a neutral area in the image (e.g., a gray or white area).
  - o If it's not neutral, it'll show you which colors are dominant, guiding you on which adjustments to make.
- **Use Curves for Additional Color Correction**:
  - o Go to **Image** > **Adjustments** > **Curves** or use the shortcut **Ctrl** + **M** (Windows) or **Cmd** + **M** (Mac).
  - o Like with Levels, you can select individual color channels and adjust the curve to fine-tune the colors.
- **Save Your Adjusted Image**:
  - o Go to **File** > **Save As** to save the color-corrected image. It's advisable to use a new filename to keep the original unedited version.

Curve Adjustments in Photoshop:

- **Open Your Image**: Launch Adobe Photoshop and open the image you want to adjust.
- **Access the Curves Tool**:
  - o Navigate to the top menu and select **Image**.
  - o Hover over **Adjustments**.

- Select **Curves**, or simply press **Ctrl** + **M** (Windows) or **Cmd** + **M** (Mac) as a shortcut.
- **Understand the Curves Graph**:
  - The graph you see is a representation of the image's tonal range. The bottom left of the graph represents the shadows (darks), the middle represents the mid-tones, and the top right represents the highlights (lights).
  - The horizontal axis of the graph represents the input levels (original values), and the vertical axis represents the output levels (new adjusted values).
- **Adjusting the Curve**:
  - To make an adjustment, click anywhere on the curve to add an anchor point.
  - Drag the point up to lighten that part of the image or down to darken it.
  - For finer adjustments, you can add multiple anchor points along the curve.
  - For example, if you want to brighten the mid-tones without affecting the shadows or highlights, you'd place an anchor point in the middle of the curve and drag it slightly upward.
- **Adjusting Individual Color Channels**:
  - By default, you're adjusting the composite RGB channel (all colors combined).
  - However, you can modify individual color channels (Red, Green, Blue) by selecting the channel from the dropdown menu in the Curves dialog.
  - Adjusting individual channels can help correct color casts in an image.
- **Preview the Changes**:
  - As you adjust the curve, you can toggle the "Preview" checkbox on and off to see a before-and-after comparison.
- **Finalize Your Adjustments**:
  - Once you're satisfied with the adjustments, click **OK** to apply them to your image.
- **Save Your Image**: Go to **File** > **Save As** to save the adjusted image, preferably with a new filename to retain the original unedited version.

Rotation: The digital version should be a faithful representation of the original item, including its orientation. For example, textual documents should be set so the text reads from left-to-right and top-to-bottom. Utilize gridlines or guides for precision rotation when necessary. When rotating, adhere to standard angles like 90, 180, or 270 degrees, unless the original item demands otherwise. Additionally, always retain a backup of the original, unaltered image to safeguard against potential issues in subsequent steps.

Rotating Images in Photoshop:
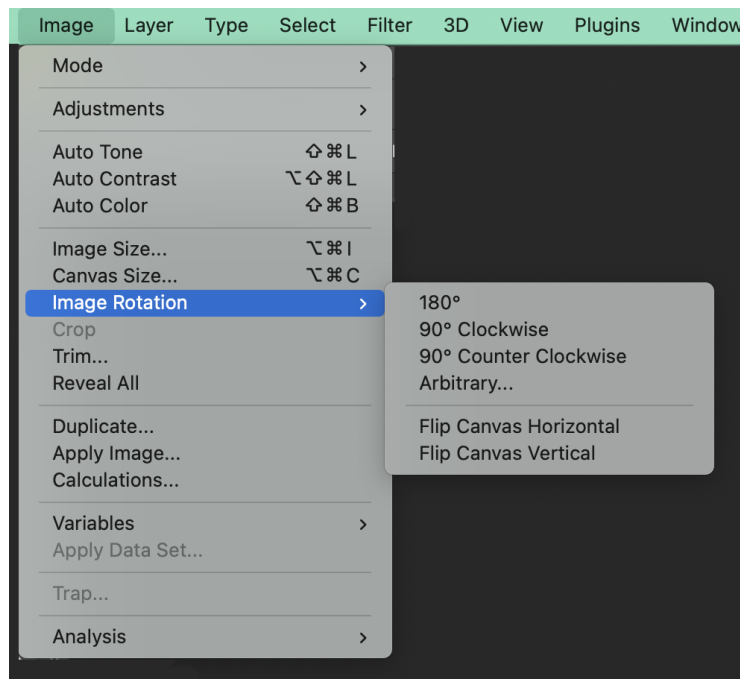
- Open Your Image:
- Go to File > Open and select your image.
- Choose the Appropriate Rotation Method:
  a. Rotate Canvas for Broad Adjustments:
     - Navigate to Image > Image Rotation.
     - You'll find options to rotate the image 180 degrees, 90 degrees clockwise, or 90 degrees counterclockwise.
     - There's also an option for a custom angle if you need a specific degree of rotation.
  b. Free Transform for More Flexible Adjustments:
     - Press Ctrl + T (Windows) or Cmd + T (Mac) to activate the Free Transform tool.

- Hover near a corner until you see a curved arrow. Click and drag to rotate.
- Press Enter (Windows) or Return (Mac) to apply the transformation.

c. Rotate View for Non-Destructive Viewing:
- This tool allows you to rotate your view of the image without actually rotating the image itself—useful for editing at different angles.
- Select the Rotate View tool from the toolbar or press R.
- Click and drag in the image window to rotate your view.

- Use Guides for Precision:
  o If you need to align your image to a specific element or ensure its level, use guides or gridlines.
    - Go to View > New Guide... to place a horizontal or vertical guide.
  o Rotate using the Free Transform tool, using the guide as a reference.

- Save Your Work:
  o Once you're satisfied with the rotation, go to File > Save or File > Save As... to store your changes.
  o Remember to retain a copy of the original if needed.

Deskewing: Deskewing an image may be necessary to ensure that the accuracy and authenticity of the original material are replicated in the digital copy. During the digitization process, pages may not lie flat, for example, when part of a bound book is placed on a flatbed scanner, or an image is captured using an overhead camera. This can cause the content to appear "skewed" or distorted, with the text or image not perfectly horizontal or vertical and the sides of the page being warped. In addition to looking distorted, these irregularities can impact the functionality of the digitized content, particularly when OCR is involved. Before deskewing, retain a backup of the original image.

Deskewing Images in Photoshop:
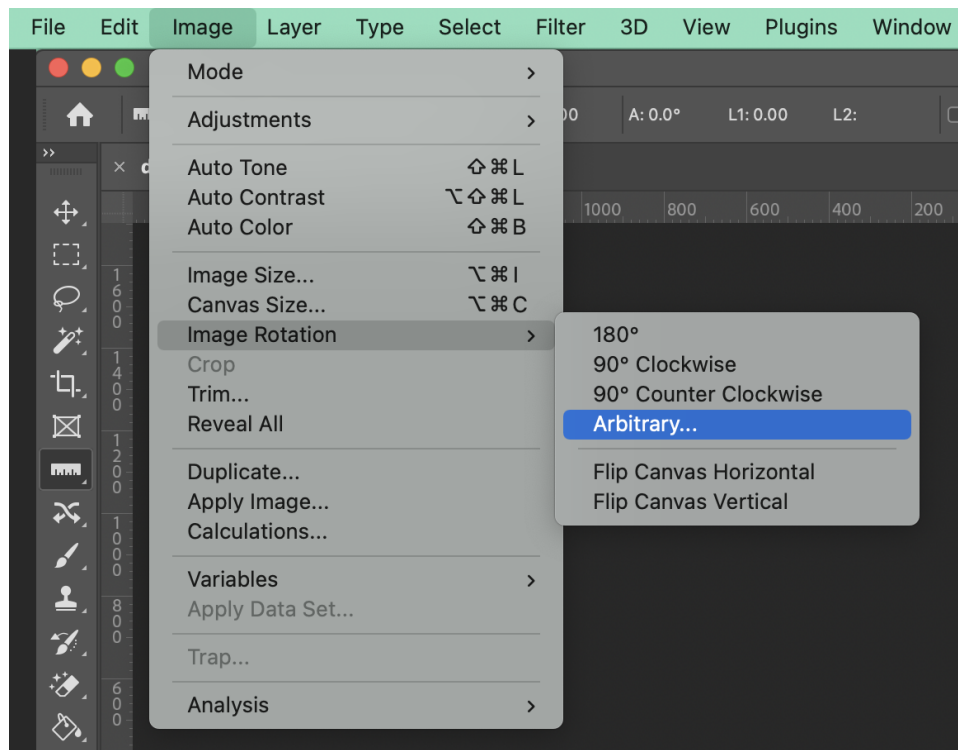
- Open the Image: Start by opening the skewed image in Photoshop. Navigate to File > Open, and select your image.
- Zoom In for Better Visibility: Use the zoom tool or press Ctrl + + (Windows) or Cmd + + (Mac) to magnify the image, allowing you to see the skew clearly.
- Select the Ruler Tool: In the toolbox, find and select the Ruler Tool. It might be hidden under the Eyedropper Tool, so you may need to click and hold the Eyedropper icon to reveal it.

- Draw a Reference Line: With the Ruler Tool active, click and drag along a line in the image that should be straight, such as the edge of a page or a line of text. This line will act as a reference for Photoshop to determine the correct angle to deskew the image.
- Rotate the Image: Once you've drawn your reference line, go to Image > Image Rotation > Arbitrary. Photoshop will automatically fill in the angle of rotation needed to straighten the image based on the reference line you drew. Click OK to apply the rotation.
- Crop the Image: After deskewing, there might be white or transparent areas along the edges. Use the Crop Tool to remove these, ensuring your image retains its focus on the essential content.
- Save the Deskewed Image: Once satisfied with the deskewed result, save your image by navigating to File > Save As, and choose the desired format and location.
- Review and Double-Check: Always review the deskewed image to ensure it aligns correctly and maintains its original integrity. It's a good practice to cross-check with the original item or a reference to ascertain its alignment.

Stitching/Photomerge: In digital preservation, there are instances where objects are too large to be captured in a single scan, especially with oversized documents, maps, or artworks. While utilizing a large format scanner that can accommodate the item's full size is always the best practice and the preferred method, there may be occasions when this isn't feasible. In such cases, the item may be scanned in sections and then digitally stitched back together using tools like "Photomerge" in Photoshop. This method ensures that every detail of the original is captured, though care must be taken to maintain consistency and accuracy across all sections.

Photomerge in Photoshop:
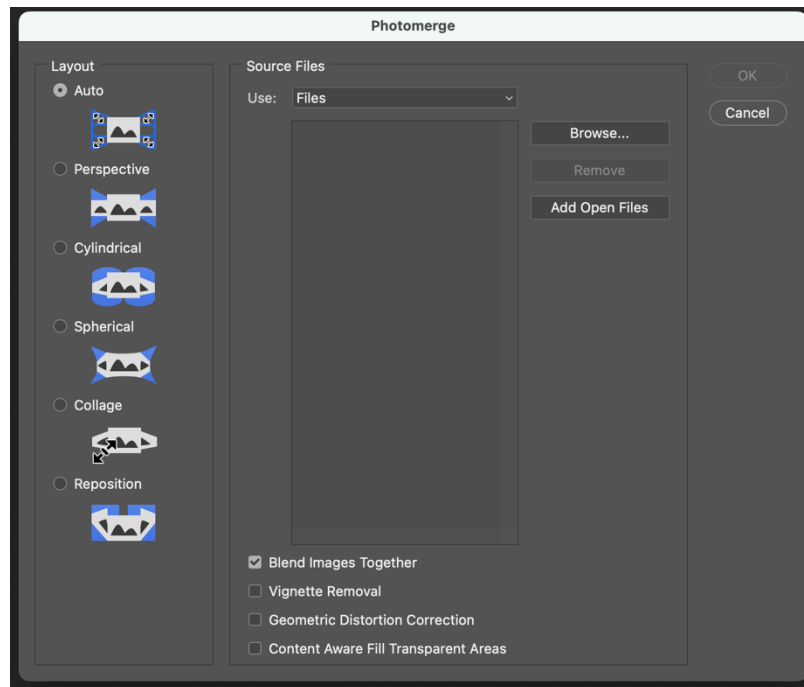
- **Launch Photoshop**: Open Adobe Photoshop on your computer.
- **Access Photomerge**: Navigate to **File** > **Automate** > **Photomerge**.
- **Select Source Files**:
  - Click on the **Browse** button.
  - Navigate to and select the images you want to stitch together.
  - Click **Open** to add them to the Photomerge window.
- **Layout Options**: There are several layout options available:
  - **Auto**: Photoshop analyzes the source images and selects what it thinks is the best layout option.
  - **Perspective**: Designates one of the images (usually the middle one) as the reference, and the other images are transformed accordingly.
  - **Cylindrical**: Reduces the "bow-tie" distortion that can occur with the perspective layout.
  - **Spherical**: Aligns and transforms the images as if they were for mapping the inside of a sphere.
  - **Collage**: Rotates or scales any of the source layers to match the content.
  - **Reposition**: Only aligns the images, but does not transform them.

- o **For most projects, the "Auto" option should suffice**.
- **Advanced Options**:
  - o **Blend Images Together**: This option should typically be left checked, as it blends the images for a seamless transition.
  - o **Vignette Removal** and **Geometric Distortion Correction**: These can be checked if needed based on the characteristics of the source images.
- **Press OK**: Once you've selected your layout and set your options, click **OK** to begin the merging process.
- **Review and Adjust**: Once Photomerge completes the process:
  - o Inspect the resulting image for any inconsistencies or misalignments.
  - o Use Photoshop's tools to make any necessary corrections or enhancements.
- **Save**: After ensuring the final image is satisfactory, navigate to **File** > **Save** to store the merged image in your desired format and location.

Batch Processing/Actions: Batch processing is a method that allows for simultaneous editing of multiple images. This approach is beneficial when dealing with large sets of digitized materials, as it can enhance productivity and maintain the quality of the original content. This technique is most effective when addressing uniform issues across a set of images. If each image has unique problems, batch processing may not provide consistent results. For example, if all scans need to be rotated slightly at the same angle, or all images need the same color correcting features applied, batch processing may be appropriate.

Keeping original copies is an essential safeguard against potential errors during the batch process. This backup ensures that any mistakes can be corrected without data loss. The primary benefit of this method is time efficiency, eliminating the need for manual edits on each image. However, after batch processing, it's crucial to conduct a quality control review of some of the edited images to confirm the accuracy of the applied changes.

Creating Actions in Photoshop:
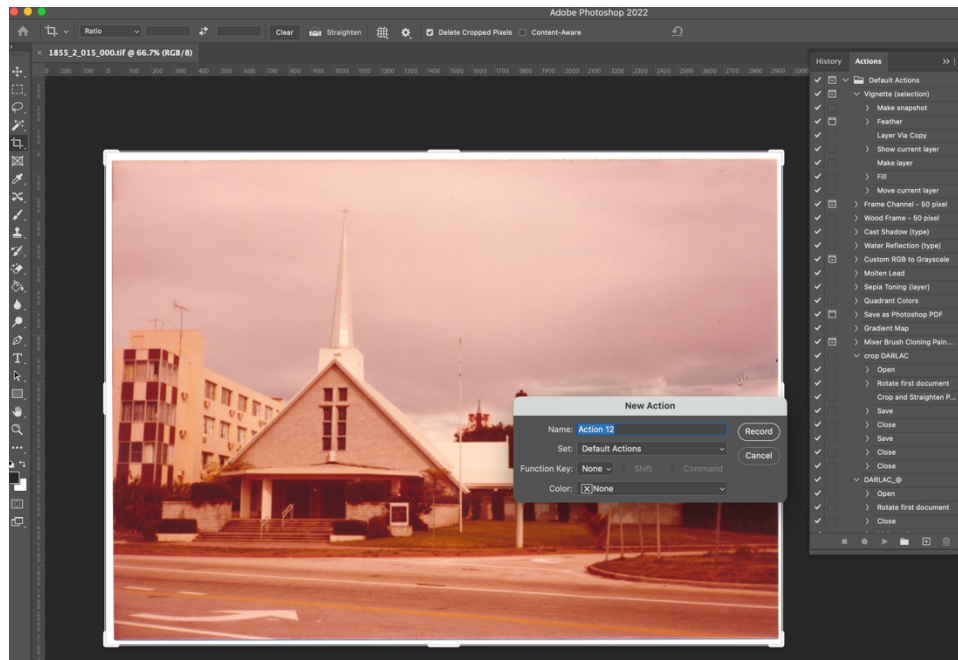
- **Open Photoshop**: Launch the Adobe Photoshop program on your computer.
- **Load Your Image**: Open the image you want to apply the action to by going to **File** > **Open**.
- **Open the Actions Panel**: Navigate to **Window** > **Actions** to open the Actions panel.
- **Create a New Action**: In the Actions panel, click on the **New Action** button (it looks like a paper icon). Give your action a name.
- **Record the Action**:
  - Once you've named your action, click **Record**.
  - Photoshop is now recording everything you do. Perform the edits you want on your image.
- **Stop Recording**: After you've finished making your edits, click the **Stop** button at the bottom of the Actions panel (it looks like a square icon).
- **Applying the Action**:
  - Open another image (or multiple images) to which you'd like to apply the action.
  - With the image open, select the action you've created in the Actions panel and click the **Play** button (it looks like a triangle icon).

- **Saving Your Actions**: If you want to save your actions for future use, click on the menu icon (three lines) in the top right corner of the Actions panel and select **Save Actions**.

Running an Action:
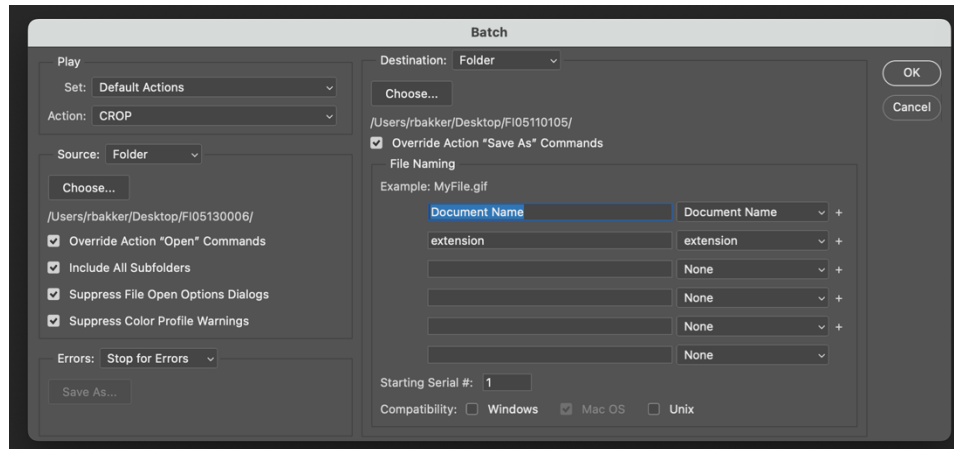
- **Open Photoshop**: Launch the Adobe Photoshop program on your computer.
- **Navigate to Batch Process**: Go to **File** > **Automate** > **Batch**.
- **Select the Action**: In the "Play" section of the Batch dialog box, choose the set where your desired action is saved from the "Set" dropdown. Next, select the specific action you want to run from the "Action" dropdown.
- **Choose Source Folder**: In the "Source" section, select "Folder" from the dropdown. Then click "Choose" and navigate to the folder containing the images you want to process.
- **Destination Settings**:
  - **Destination**: Choose a location to save the processed images. This can be the same as the source folder or a different one.
  - **Folder**: If you select "Folder" as the destination, click "Choose" to select a destination folder.
  - **File Naming**: Optionally, adjust file naming conventions if you want to rename the processed files.
- **Additional Settings**: Depending on your needs:
  - Check "Override Action 'Open' Commands" if your action includes an open command you wish to override.

- o Check "Override Action 'Save As' Commands" if you want to specify a new location or file naming convention different from those in the action.
- **Run the Batch Process**: Click the **OK** button to start the batch process. Photoshop will automatically process each image from the source folder using the selected action and save them to the destination folder.
- **Review the Results**: After the batch process completes, navigate to the destination folder and review the images to ensure the action applied correctly.

Optical Character Recognition

Optical Character Recognition, commonly known as OCR, is a system developed to transform a variety of document formats into text that machines can read. This encompasses both scanned documents and digital images. OCR is frequently characterized as the capability to recognize text. Its role is pivotal in the process of digitization, particularly for retrieving information. The effectiveness of OCR has a direct bearing on the ease with which data can be read and processed in a digital environment. Using OCR, vast quantities of paper-based records can be promptly converted into electronic data, streamlining the availability, searchability, and sharing of information. Beyond just conversion, OCR helps preserve the essence of aging or fragile documents by transitioning them to a digital medium, ensuring they remain accessible over time. It's distinct from "Intelligent Character Recognition" (ICR), which is tailored to identify handwritten entries. OCR doesn't just identify and encode text; it also attempts to discern and code the layout aspects of a document, such as its columns and non-textual graphical components.

To achieve accurate OCR results, it's important to follow best practices, starting with document preparation, when available. Good contrast between text and background is helpful, as it enables OCR software to identify characters more easily. Fonts larger than 6 points also yield better accuracy. Additionally, documents printed post-1950 tend to provide better results due to advancements

in print technology and more standardized fonts. Inconsistencies such as varied font sizes, italics, underlining, and irregular spacing might hamper accuracy.

When converting physical documents to digital format, it is standard to use RGB mode to capture detailed image data. However, in some instances, capturing in grayscale with a slight enhancement in contrast will yield a better result. A resolution of 300 dpi is suitable for a majority of documents. However, for pages with smaller text (below 10 points), aged fonts, or those with significant background "noise," a resolution ranging from 400-600 dpi is recommended.

OCR is typically applied during image-to-PDF conversion, positioning the recognized text beneath the image in the final PDF. However, several factors may pose challenges to accuracy, including symbols that look similar, certain typographical choices, scripts other than the Latin alphabet, unique proper names, and any bleed-through from the other side of a page. Being aware of these potential hurdles can help refine the preparation process. Always keep the original documents as a reference in case there are any discrepancies or errors following the OCR process.

OCR Software Options

Various software solutions are available for OCR, each with its unique features tailored to different needs. Adobe Acrobat is a widely accessible option for many institutions, providing basic OCR capabilities. However, one potential limitation is its restricted functionality when it comes to editing the recognized text post-OCR. ABBYY FineReader offers a notable advantage in this realm, giving users the option to manually zone text, a particularly useful feature for documents with complex layouts. Additionally, its user-friendly interface and multilingual capabilities facilitate easier text editing or rekeying, further enhancing its appeal.

PrimeOCR stands out for its capabilities in batch processing and auto-zoning, simplifying the task when dealing with large volumes of documents. On the other hand, Tesseract, being open-source and AI-driven, is a popular choice for those looking for a low-cost OCR solution. However, potential users should note that

while powerful, Tesseract may require a bit of coding know-how given its limited graphical user interface.

Choosing the right OCR software often hinges on various factors such as the nature and scale of the project, the type of documents being processed, language specifications, and other considerations like software availability, user-friendliness, and budget constraints. Therefore, partner institutions should evaluate each option considering their specific requirements to determine the most suitable choice for their OCR projects.