PDF ACCESSIBILITY

Use OCR and Word to Turn Page Scans into Clean Text

It’s common for instructors to post scanned pages from books and other print sources as PDFs. However, those pages are rendered as images instead of actual text, and therefore need to be made accessible. How big of a job this is depends on the quality of the scans and how well the Optical Character Recognition, or OCR, performed on the document.

BEFORE YOU BEGIN

Before you begin working on your PDF, it must have:

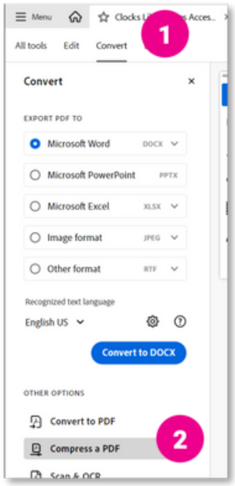
- 1 scanned page per PDF page.
- Text that is upright and does not have to be rotated to be read.

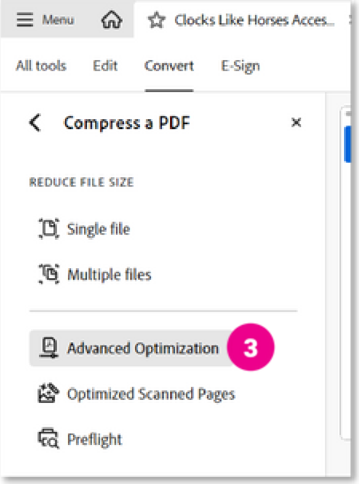
CONVERT PDF TO A TEXT WORD DOCUMENT

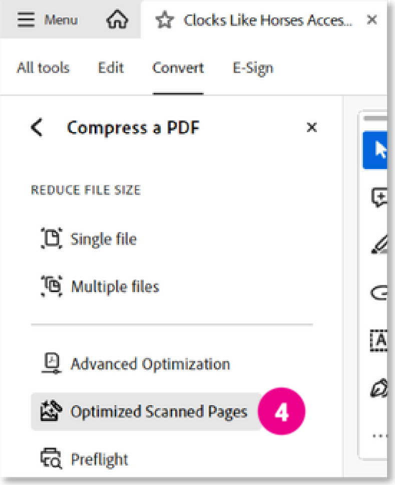
1 Open your PDF in Adobe Acrobat.

2 In the navigation across the top, choose **Convert** (1), click on **Compress a PDF** (2) and choose **Advanced Optimization** (3). You will be prompted to save as a copy.

3 After that completes, you will be working in the copy and still in the same location. Choose **Optimize Scanned Pages** (4).

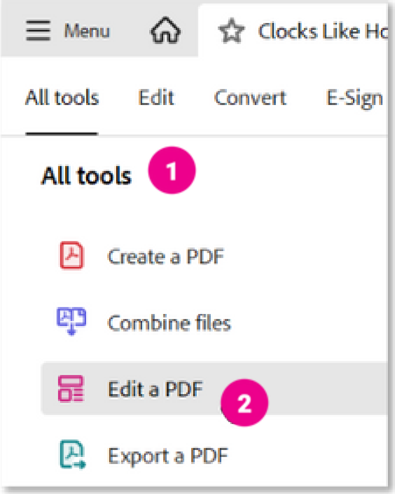






4

In the navigation across the top, click on **All tools** (1), then in the left-hand navigation, click on **Edit PDF** (2), and remove as much of the image garbage as you can, such as images of the book fold. **SAVE** when you are done.



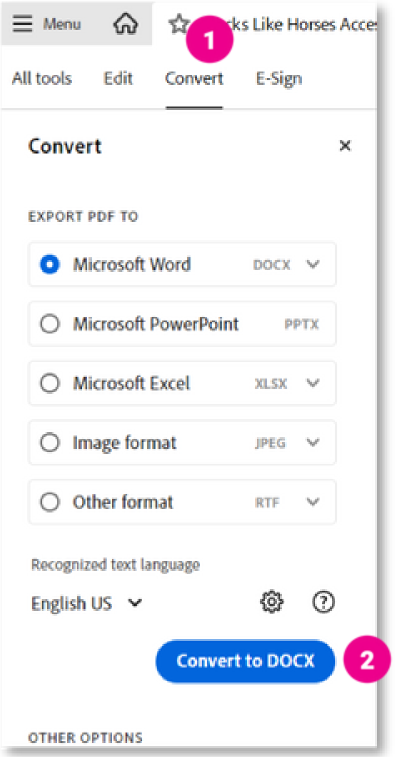
What to watch for

As you delete, watch that significant amounts of type are not deleted as well. **Undo** (Ctrl Z (PC) or Cmd Z (Mac)) to restore the type. Any image garbage you can't remove now will be removed later.

Don't worry about fixing errors in the text for now. Those will be fixed later.

5

Click on **Convert** (1) again, make sure the radio button beside Microsoft Word docx is selected, and click the blue button **Convert to DOCX** (2).Close the PDF after it makes the Word document.



CLEAN UP THE TEXT: 2 OPTIONS

You will need to decide whether or not it's important to retain the look and page breaks of the original scan.

Go with Option 1 if the text is all you need. Cleaning up the text is much easier if you are free to discard the look and page breaks of the original scan.

Go with Option 2 if you need to retain the look (fonts and margins) and page breaks of the original scan. Be aware you will need to do more work and take more care as you go along to retain those features.

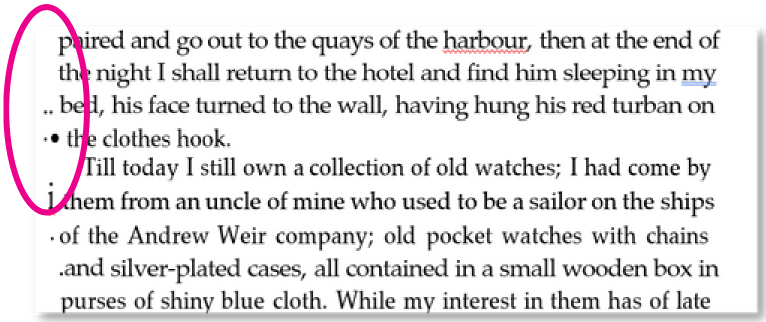
OPTION 1: TEXT-ONLY CLEANUP

1. Open the Word document in your desktop Word app, NOT Word Online.

2. Read through the text to fix OCR errors, doing a comparison with the original PDF scan as you go. Remove hyphens that are there for line breaks. **SAVE** when you are done.

What to watch for

- Common errors are missing letters or words, garbage characters, and words running together, as shown below.
- Errors will tend to be in the words along the margins, so pay close attention there.
- Don't worry about problems in the text or margins caused by the formatting that came over from the PDF. Step 5 will fix all that



paired and go out to the quays of the harbour, then at the end of
the night I shall return to the hotel and find him sleeping in my
.. bed, his face turned to the wall, having hung his red turban on
·• the clothes hook.
Till today I still own a collection of old watches; I had come by
i them from an uncle of mine who used to be a sailor on the ships
·of the Andrew Weir company; old pocket watches with chains
.and silver-plated cases, all contained in a small wooden box in
purses of shiny blue cloth. While my interest in them has of late

3 Delete page numbers and other header and/or footer information throughout the document. **SAVE** when you are done.

4 Copy the text of the entire document.

5 Open Notepad (PC) or TextEdit (Mac) and paste the text into the app.
If using TextEdit, go to **Format** in the menu and choose **Make Plain Text**.
This "scrubs" the text of any graphics, formatting, etc., reducing it to plain text.

6 Create a new Word file, copy the plain text in Notepad or TextEdit and paste it into the new Word document.

7 Select all the text and make it all the same font and size. For example, style all the text in Times New Roman font, 12-point type.

8 Apply appropriate header styles to the document per generally accepted practices on Word document accessibility, as detailed in the USF SAS Accessibility Guide.

9 Run spellcheck and grammar check to help you find errors you might have missed. Be careful not to make a correction that changes the text from what was there originally. **SAVE** when you are done.

10 Copy any images from the old Word document and paste them into this one, positioning them where appropriate. Apply alt-text to them per generally accepted practices on making images accessible. To apply alt-text, **right-click** on the image and choose **View Alt Text** in the menu that appears.

11 Add full citation information at the end of the document.

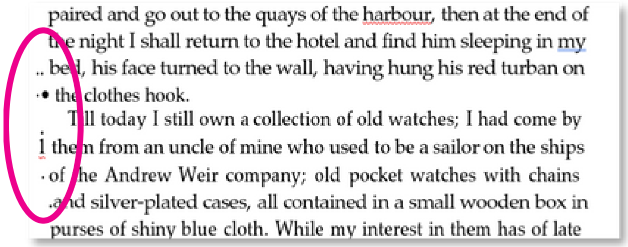
12 **SAVE** when you are done.

OPTION 2: TEXT CLEANUP, KEEP LOOK OF ORIGINAL SCAN

Depending on the quality of the original scan, you might need to make a few passes through the document to get everything.

1 Open the Word document in your desktop Word app, NOT Word Online.

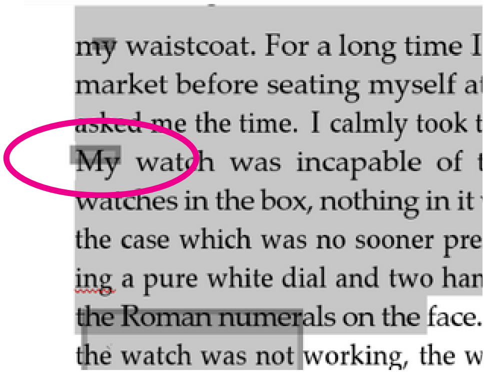
2 In the first pass, remove random characters (usually in the margins) and remaining visual garbage. **SAVE** when you are done.



paired and go out to the quays of the harbour, then at the end of
the night I shall return to the hotel and find him sleeping in my
.. bed, his face turned to the wall, having hung his red turban on
• the clothes hook.
Till today I still own a collection of old watches; I had come by
i them from an uncle of mine who used to be a sailor on the ships
· of the Andrew Weir company; old pocket watches with chains
and silver-plated cases, all contained in a small wooden box in
purses of shiny blue cloth. While my interest in them has of late

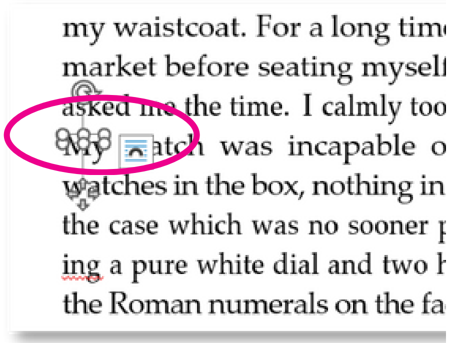
What to watch for

- If you deleted garbage characters and it left a hole, leave it for now. There likely is text that goes here, and leaving the hole helps you find the spot again.
- Select all the text on each page. You might see tiny, empty image boxes appear, as shown by the dark shaded boxes in Figure 1. These all need removed by clicking on each one as shown in Figure 2, and deleting. You might have to zoom in to get the ones that are hard to catch. You'll have to highlight the page again after deleting each box to find the next one.



my waistcoat. For a long time I
market before seating myself at
asked me the time. I calmly took t
My watch was incapable of t
watches in the box, nothing in it
the case which was no sooner pre
ing a pure white dial and two han
the Roman numerals on the face.
the watch was not working, the w

Figure 1



my waistcoat. For a long tim
market before seating myself
asked me the time. I calmly too
My watch was incapable o
watches in the box, nothing in
the case which was no sooner p
ing a pure white dial and two h
the Roman numerals on the fa

Figure 2

3

In the second pass, read through the text to fix OCR errors, doing a comparison with the original PDF scan as you go. Hyphens in words for line breaks will need to be replaced with a "soft hyphen." The keystroke is **CTRL Hyphen** (PC) or **Command Shift Hyphen** (Mac) . Should things shift, the hyphen will disappear if it's no longer needed.

What to watch for

- Errors will tend to be in the words along the margins, so pay close attention there. The page margins might get messed up as you make corrections, but leave the margins alone for now.
- Common errors are missing letters or words, garbage characters, misspellings, words running together, and misplaced hyphens.

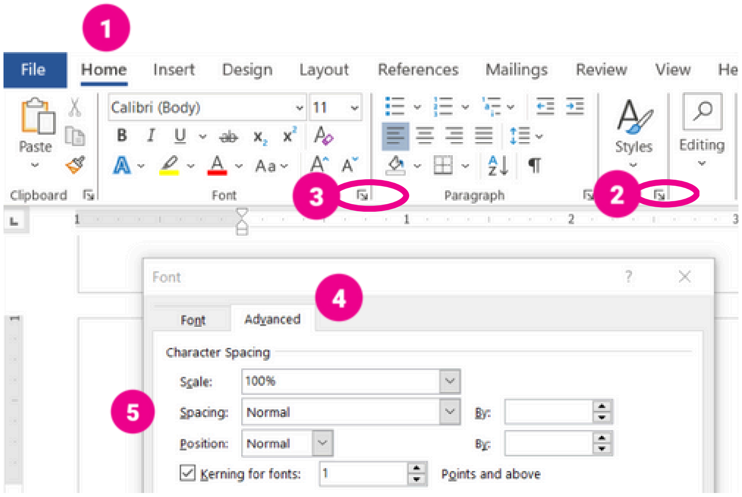
4

Run spellcheck and grammar check to catch errors you might have missed, then **SAVE**.

5

Fix the varying fonts, type sizes, line spacing, margins, etc. and make them the same. Do this one page at a time, and **SAVE** after completing each page. Ensure that each page ends in the same place as the original PDF if that's important to you.

- Adjust line spacing on the **Home** tab (1) in the Paragraph settings pop-out (2).
- Adjust spacing, scale and position of letters in the Font settings pop-out (3), on the **Advanced** tab (4). Set Scale to 100%, and Spacing and Position to Normal (5).



Character Spacing

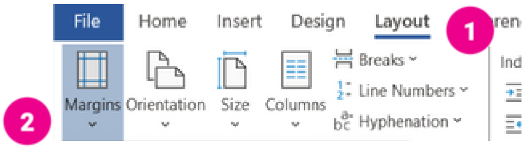
Scale: 100%

Spacing: Normal By:

Position: Normal By:

☑ Kerning for fonts: 1 Points and above

- Adjust margins on the **Layout** tab (1) in the **Margins** drop-down (2).



6 Check and/or apply appropriate header styles to the document per generally accepted practices on Word document accessibility, as detailed in the USF SAS Accessibility Guide.

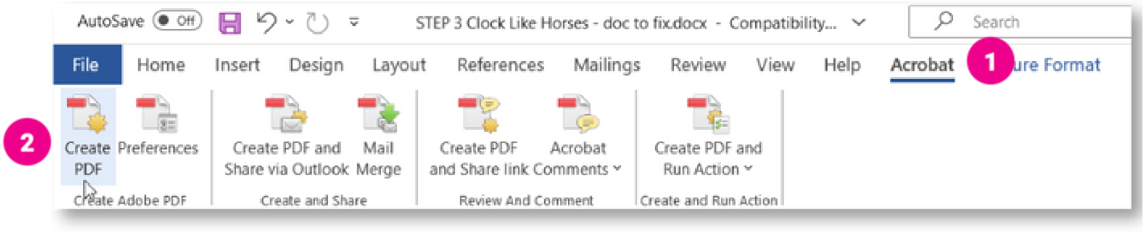
7 If there are images, add alt-text to them by **right-clicking** on an image and choosing **View Alt Text**. Add alt-text per generally accepted practices on making images accessible, as detailed in the USF SAS Accessibility Guide.

8 Add full citation information at the end of the document.

9 Run spellcheck and grammar check one more time, then **SAVE**.

CONVERT WORD DOCUMENT BACK TO PDF, FINAL FIXES

1 With your Word document open, click on **Acrobat** (1) in the ribbon across the top and choose **Create PDF** (2).



2 Close Word and open the new PDF in Acrobat.

3 Check its accessibility and make any fixes per generally accepted practices on PDF accessibility, as detailed in the USF SAS Accessibility Guide.

4 **SAVE** when you are done.