HUMBOLDT STATE UNIVERSITY

HSU Document Converter

Guide for Faculty and Staff to Convert Image Files to Accessible (Text-Based) Files

Conversion Best Practices written by Sean Keegan, Stanford University

All documents and instructional materials created and distributed must be accessible, at the time of distribution, to all individuals including students with print disabilities. The HSU Document Converter is a web-based service that will convert documents into a range of accessible formats. The service can be used to convert inaccessible documents such as image-only PDF files and JPG pictures into accessible (text-based) formats. The service has been provided as a do-it-yourself solution for faculty and staff who prefer to independently prepare their instructional or departmental materials in an accessible manner. If you are a faculty member and have inaccessible (image-based files) that need converting but are not interested in a do-it-yourself approach, the ONCORES[1] office in the Library is available to assist faculty in converting paper-based handouts and supplemental readings into an electronic and accessible format. The following best practices identify simple methods to prepare the file before converting in order to achieve a high-quality output.

Step 1 : Scan Your Document to Create Image-only PDF file

PDF and image-based files will be processed using optical character recognition (OCR) to create a text-based version of the document. The quality of a conversion is dependent upon the quality of the original document.

- If scanning the document:
  - Ensure the scanned image is free from smudges, dark marks, highlighted text, or artifacts in the image. These will affect the accuracy of the OCR process.
  - Minimize the effects from skewing. If the image is presented at an "off-angle", the accuracy of the OCR process will be lower resulting in a lower quality text conversion.
  - If scanning from a book, scan each page separately rather than scanning facing pages from a book or dual pages. Also, ensure the pages are oriented correctly when you scan them.
  - Scan using either 300 dpi or 400 dpi. Low image resolutions (below 150 dpi) may have a negative impact on OCR quality while excessively high image resolutions (over 600 dpi) do not yield any significant improvements in OCR quality and take a long time to process.

Step 2 : Upload Your Document to the HSU Document Converter

1. HSU Document Converter[2]
2. Select "File" as your source
3. Select "Browse…" to search for the file you would like converted
4. Find the file and select "Open". The file name appears to the right of the "Browse…" button
5. Select "Upload". This may take a few seconds
6. Select the output format. If converting to text, select "Accessibility conversion"
7. Select the format desired for the conversion

---

[1] ONCORES: http://library.humboldt.edu/search/reserves/index.html

[2] HSU Document Converter: http://www.humboldt.edu/disability/servicesavailable-sensus

- The HSU Document Converter will convert image-based document into MS Word, RTF, and text files. [With some image-based documents, you may achieve better results if you convert initially to Tagged PDF and then copy and paste the text from the Tagged PDF into MS Word. This may result in a better reading experience and may remove non-essential content.]

8. Enter your HSU email address (must use HSU email for this service)
9. Select "Submit"

Step 3 : Receive Converted file in your HSU email In-Box

The converted file will be sent to your HSU email in-box. Conversions that are between text-based formats will generally be completed in less than an hour. Conversions that involve converting image-based documents into text-based documents (e.g., scanned PDF to tagged PDF) or the creation of audio files may take longer.

Step 4 : Edit the Document

*** Do not skip this step! *** This step is very important given the OCR process is not 100 percent accurate. Make any corrections necessary to ensure the conversion matches the original **exactly**. Skipping this step will result in a file that is not identical to the original and misinformation may be communicated to students. In most cases, editing will take only a short time within MS Word and involves the use of the Spell Checker and the Find and Replace tools. See Using the Find and Replace in MS Word[3] for additional information on removing special characters in a document.

Please note: in the Find and Replace examples below, replace the <space> value with one spacebar and do not include the quotes.

Image-File to Tagged PDF to MS Word Document Editing Techniques

1. Open the Tagged PDF file and select all the text
2. Select Copy
3. Open MS Word
4. Select Paste
5. Review the text and correct any misspellings etc. that occurred during the OCR conversion process
6. Use Word styles to specify document headings. For example, the style "Heading 1" could be used to identify the title of the document and the style "Heading 2" could be used to identify chapter information
7. Provide short descriptions for content-related images in your MS Word document
8. You may need to clean up spacing issues. If needed, use Find and Replace:
   - Search for ".<space>^p" and replace with ".^p^p".
   - Search for "<space>^p" and replace with "<space>".
   - Search for "<space><space>" and replace with "^p<space>".
   - Search for "-<space>" and replace with no value.
9. Save the document in your preferred text format

[3] Using the Find and Replace in MS Word: http://tinyurl.com/ot42qub

Image-File to MS Word Document Editing Techniques

To clean-up an image file that was converted to a MS Word file, perform a “search and replace” to remove optional hyphens and section breaks.

1. Open the converted Microsoft Word document
2. Review the text and correct any misspellings etc. that occurred during the OCR conversion process
3. Use Word styles to specify document headings. For example, the style “Heading 1” could be used to identify the title of the document and the style “Heading 2” could be used to identify chapter information
4. Provide short descriptions for content-related images in your MS Word document
5. You may need to clean up spacing issues. If needed, use Find and Replace:
    - Search for “Optional Hyphen” under Special Formatting and replace with no value.
    - Search for “Section Breaks” under Special Formatting and replace with “^p^p”.
    - Search for “Manual Page Breaks” and replace with “^p^p”.
6. Save the document in your preferred text format

Step 5 : Distribute the New (Accessible) Document

Upload the new document to Moodle, post to your instruction-related website, send to staff, etc.

Step 6 : Repeat for Any Other Image Files You May Require