Conversion Best Practices

Provided courtesy of Sean Keegan. The material may be used in connection with SensusAccess solutions provided Sean Keegan is credited.

The quality of a conversion is dependent upon the quality of the original document. Additionally, the resulting output format may include enhancements for navigation if the original file contains the appropriate semantic markup. For instance, a MS Word document containing the heading style markup for chapters (e.g., Heading 1, Heading 2, etc.) will convert into a more usable DAISY or EPUB format with the relevant chapter navigation elements. The following best practices identify simple methods to prepare the file before converting in order to achieve a high-quality output.

PDF & Image-based Files

- PDF and image-based files will be processed using optical character recognition (OCR) to create a text-based version of the 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 any 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 version.
- If you are starting with an image-based format and wish to convert to a text format, you may achieve better results by initially converting to Tagged PDF and then copying/pasting the text into a MS Word document. While you can convert directly from an image file to a text file with SensusAccess, you may find better results for some image documents if converting to Tagged PDF and then to a text file (see "Converting to MS Word and Text Files" section).

Converting to MS Word and Text Files

SensusAccess will convert image-based documents into MS Word, RTF, and text files. You may also find it useful with some image-based documents to 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.

With the MS Word version of the document, you can more accurately "clean" the content for conversion into MP3 audio or for use with assistive technologies. Most conversions will take just a few seconds within MS Word and involve the use of the Find and Replace tools. For more information on using the

Find and Replace tools, see Using the Find and Replace in MS Word 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

- Submit the image-based document to SensusAccess and select Tagged PDF as the output option.
- Open the Tagged PDF and select all the text. Copy and paste this into a MS Word document (Open Office may also be used).
- Using 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.
- Save the document in your preferred text format.

Image-File to MS Word Document

To clean-up a MS Word file for use with assistive technology or for creating MP3 files, perform a "search and replace" to remove optional hyphens and section breaks. Identify the special character you wish to find in the "Find:" box and leave the "Replace with:" box empty. See Using the Find and Replace in MS Word for additional information on removing special characters in a document.

- Submit the image-based document to SensusAccess and select Microsoft Word as the output option.
- Open the converted Microsoft Word document (Open Office may also be used).
- Using 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".
- Save the document in your preferred text format.

Authoring MS Word, RTF, Text Files

- 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. It is best to use only one "Heading 1" to facilitate accurate conversions into other document formats (e.g., DAISY, EPUB, Braille, etc.).
- Provide short descriptions for content-related images in your MS Word document.

- Avoid using text-boxes in your document. If you want to customize the layout, use a Column Tool or a Section Break.
- If converting to DAISY, page numbers will be identified based on the MS Word pagination. To obtain custom pagination, use the PageNumber style from the Save As DAISY plug-in for Microsoft Office for your custom page numbers.

Authoring HTML Files

- Use HTML heading markup (e.g., <h1>, <h2>, etc.) to designate headings in the document. 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.
- Provide short descriptions for content-related images in the HTML document.