**OCR: The What, Why, and How**



● Adobe Acrobat Pro DC version

○ Instructions to run OCR on a **single** file:

* Open the PDF
* On the **Tools** menu, select **Enhance Scans**.
* Select **Recognize Text.**
* Select **In This File.**
* A new tool barwill appear. In the **Pages** dropdown, select **All Pages**.
* In the **Language** dropdown, select the language of the text in your PDF document.
* In the **Settings** section, select **Searchable Image** as the desired **Output.** Select **OK.**
* Click the **Recognize Text** button.
* Save the resulting file by overwriting the existing file.

○ Instructions to run OCR on **multiple** files:

* Open one of the PDFs
* On the **Tools** menu, select **Enhance Scans**.
* Select **Recognize Text.**
* Select **In Multiple Files.**
* In the **Recognize Text** pop-up box, click **Add Files** in the top left corner.  This gives you the option to add multiple files from a single folder or multiple files from multiple folders.  Either choice gives you a pop-up box where you can select the files you want to OCR.  You can use Ctrl + Alt and Shift keys to help in selecting multiple files.
* After they have been selected, click the **OK** button in the bottom right corner of the pop-up box.
* An **Output Options** pop-up box will appear.
  + In the **Target Folder** section, you will have the option to save the file to **The Same Folder Selected** **at Start** or **A Folder on My Computer**.
  + In the **File Naming** section, you will have the option to **Keep original file names** or choose to **Add to original file names** either at their beginnings and/or ends.
  + You can also choose to overwrite the previous files as opposed to creating an OCRed copy of the original file.
  + Click the **OK** button.
* A **Recognize Text – General Settings** pop-up box appears.
  + In **Document Language**, select the language of the text in your PDF document.
  + Select **Searchable Image** as the desired **Output**.
  + Then click the **OK** button.

● Google Drive

* Open Google Drive and upload, if not done previously, the image or PDF file you would like to OCR.
* Right-click on the file.
* Select **Open with** and **Google Docs**.
* The image file is OCRed and then the resulting text appears as a Google Doc.
* Size limit: 2MB per file

● Tesseract

○ Command line or through third party frontend software

○ https://code.google.com/p/tesseract­ocr/

○ Command line instructions (Mac):

* The easiest way to install Tesseract is with MacPorts. Once it is installed, you can install Tesseract by running the command sudo port install tesseract, and any language with **sudo port install tesseract­<langcode>**. List of available langcodes can be found on MacPorts tesseract page.
* Command use: tesseract imagename outputbase [­l lang] [­psm pagesegmode] [configfile...]
* Basic command: tesseract myscan.png out
* Results in out.txt.

○ Frontend software options

* FreeOCR (W[indows) ­ http://www.paperfile.net](http://www.paperfile.net/)
* https://github.com/tesseract-ocr

**Questions?**



Mackenzie Brooks, Digital Humanities Librarian, [brooksm@wlu.edu,](mailto:brooksm@wlu.edu) x8659

Alston Cobourn, Digital Scholarship Librarian, [cobourna@wlu.edu,](mailto:cobourna@wlu.edu) x8657

Digital Humanities Action Team, [DHAT@wlu.edu](mailto:DHAT@wlu.edu)