OCR: The What, Why, and How

- Adobe Acrobat Pro
  - Available through the Stable, Adobe applications profile

  - Instructions to run OCR on a **single** file:
    - Open the PDF
    - On the **Tools** menu, select **Recognize Text**
    - Select **In This File**
    - A **Recognize Text** pop-up box will appear. In the **Pages section**, select **All** pages. In the **Settings** section, select **PDF Output Style: Searchable Image.** Select **OK.**
    - Save the file by overwriting the existing file.

  - Instructions to run OCR on **multiple** files:
    - Open one of the PDFs
    - On the **Tools** menu, select **Recognize Text**
    - Select **In Multiple Files**
    - In the Recognize Text pop-up box, click **Add Files** in the top right 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 the beginning and/or end of your file names.
    - You can also choose to overwrite the previous files as opposed to creating an OCRed copy of the original file.
    - A **Recognize Text – General Settings** pop-up box appears. Choose the **Primary OCR Language**. For the **PDF Output Style**, always select **Searchable Image**. For **Downsample To** choose 600 dpi. Then click the **OK** button.

- Google Drive
  - Size limit: 2MB per file or 10 pages of PDF
  - Upload Settings > Convert Text from Uploaded PDFs and Image Files
  - OCRed text will appear below the PDF or image in your document

- Tesseract
  - Command line or through third party frontend software
  - https://code.google.com/p/tesseract-ocr/
  - 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 (Windows) - http://www.paperfile.net
        - https://code.google.com/p/tesseract-ocr/wiki/3rdParty

Questions?

Mackenzie Brooks, Metadata Librarian, brooksm@wlu.edu, x8659
Alston Cobourn, Digital Scholarship Librarian, cobourna@wlu.edu, x8657
Digital Humanities Action Team, DHAT@wlu.edu