Tesseract

Overview

Tesseract is an Optical Character Recognition program that Islandora uses to extract text from images to files that can then be appended to an object as datastreams. It supports HOCR standards, and when invoked, Islandora will use it to create both HOCR and raw OCR output. Tesseract supports multiple languages, the installation of which are recognized by the Islandora OCR module.

Dependencies

- Autotools (Make, etc.)
- Leptonica image processing library

Provisions

- Islandora OCR

Installation



For Linux installations: While it is likely that your distribution's package manager may contain Tesseract in one of its repositories, it is **EXTREMELY** unlikely that it will be the correct version. For the Islandora OCR module to create OCR derivatives, Tesseract 3.02.02 or higher is required. At the time of writing, this is the latest stable version. **THIS MEANS THAT IT IS LIKELY THAT YOU WILL HAVE TO COMPILE IT FROM SOURCE**.

Tesseract is managed by a team at Google; the latest stable release can be found on the downloads page of their website, https://code.google.com/p/tesseract-ocr/downloads/list. A binary installer exists for Windows, and specific instructions for installing on a Mac through homebrew can be found in the Tesseract readme here: https://code.google.com/p/tesseract-ocr/wiki/ReadMe. For Linux users, or any others compiling it from source, you will need to make sure that you also have the Leptonica library installed, and that you have appropriate source building tools.

Configuration

Additional Language Support

Tesseract requires little configuration out of the box; that being said, Islandora supports the installation of multiple languages for OCR processing, and may even require English language support.. These additional languages can be found on Tesseract's download page.

To install additional languages into Islandora, you will need to know the path to your Tesseract installation's 'tessdata' folder. On Windows, this will tend to be C:\Program Files (x86)\Tesseract OCR\tessdata, and on Mac, this will tend to be /usr/local/Cellar/tesseract /<version>/share/tessdata - in both cases, if you've used the Tesseract website's own installation case. On Linux, the path will vary from distribution to distribution, but will often be /usr/local/share/tessdata or /usr/share/tessdata. Once you have found the correct folder,

- Download one of the language tarballs from the website
- Extract it
- Copy the contents of the 'tessdata' folder inside the tarball to the 'tessdata' folder on your computer
- With the Islandora OCR module installed in your site, navigate to http://**path.to.your.site**/admin/islandora/ocr and check off the new language
- Click 'Save configuration'

Your new language should now be available to perform OCR on Paged Content.