Islandora OCR

Overview

The Islandora OCR module integrates Tesseract into the Islandora Paged Content module. It allows for creation of OCR and HOCR derivatives that can be appended to a page as a datastream. Check the instructions for the OCR-compatible module you wish to use for specifics on how to create OCR derivatives.

Dependencies

- Islandora
- Tuque
- Tesseract (3.02.02 or later)
- ImageMagic (Optional, Required for OCR preprocessing)
- Islandora Paged Content (Optional)

Tesseract installation will differ depending on your operating system; please see the Tesseract README Wiki for detailed instructions.

Downloads

Release Notes and Downloads

Configuration

Configuration options for the Islandora OCR module can be found at http://**path.to.your.site** /admin/islandora/ocr, and include the following options:

- **Tesseract**: Islandora OCR requires the path to your Tesseract binary to function correctly. It also requires Tesseract to be version 3.02.02 or higher to function correctly.
- **Languages available for OCR**: Islandora can look for any additional OCR languages you have installed; these are chosen from a drop-down menu at time of ingest or derivative creation.

blocked URL

Solr result highlighting

To have Islandora viewers recognize Solr search results and highlight them one will need to configure Solr to index the HOCR in a particular fashion.

The field that the HOCR is stored in must have the following attributes: `indexed="true" stored="true" termVectors="true" termPositions="true" termOffsets="true"`

Each text node of each element in the HOCR datastream must be placed in order in a single value for the Solr field with all whitespace sub strings normalized to a single space.

Any objects that were previously ingested but require this functionality will need to be re-indexed.

Reference Implementation

Tesseract

Tesseract provides many languages which can be downloaded from here.

To install just unzip them in your tessdata directory, typically located at `/usr/local/share/tessdata`

If you want to add your own languages or train your Tesseract for your specific needs please review the documentation here

It is recommended to check the Tesseract page for more information on these options.