Islandora OCR

Overview

This module acts a Toolkit for generating OCR and word coordinate information (stored in an HOCR datastream) that is required for on-page highlighting. The module relies on Tesseract to generate this information.

Tesseract

Tesseract is an OCR engine that was developed at HP Labs from 1985 and by Google from 1995. Recognized as one of the most accurate open source OCR engines available, Tesseract will read binary, grey, or colour images and output text.

A TIFF reader that will read uncompressed TIFF images is also included. Islandora Book Solution Pack requires at least Tesseract version 3.02.02, which can be obtained from the project home page.

Dependencies

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

Islandora recommends this module for text search of OCR-ed material:

- Islandora Solr Search

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

Downloads

Release Notes and Downloads

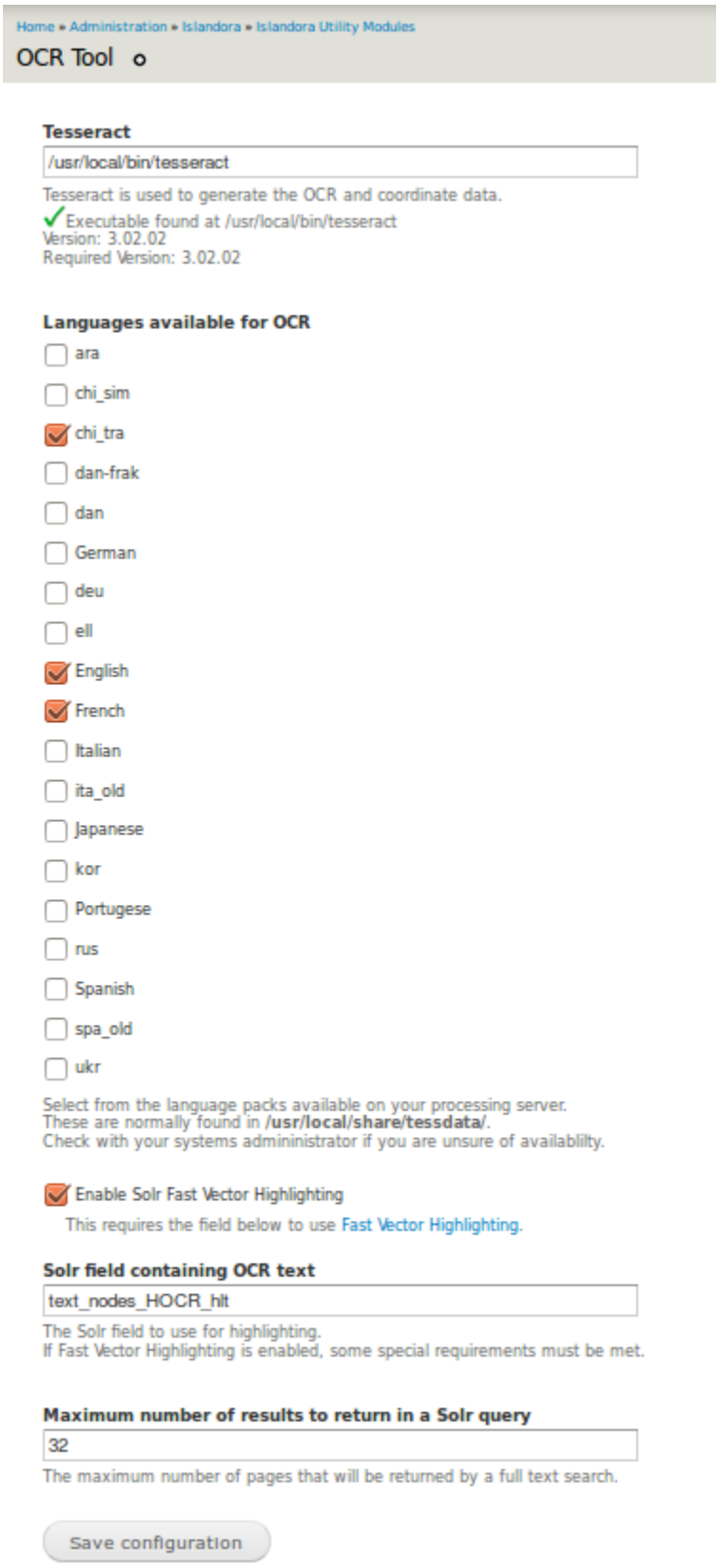
Installation

Install as usual, see this for further information.

Configuration

In Administration » Islandora » Islandora Utility Modules » OCR Tool (admin/islandora/tools/ocr), you can do the following:

- Set the path for Tesseract
- Select languages available for OCR from your Tesseract installation
- Enable/disable Solr Fast Vector Highlighting
- Set Solr field containing OCR text and the maximum number of results to return in a Solr query



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.

XSLT Reference Implementation for GSearch

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.

Note: If you running Linux distribution Tesseract software and language might be available in your repo, you can check it (e.g. on Ubuntu **apt-cache search tesseract** or on CentOS **yum search tesseract** for Tesseract and available languages,

for specific language: Ubuntu **apt-cache search tesseract | grep Greek** , CentOS **yum search tesseract | grep Greek**) when checked and found just install it on your server.