Islandora Paged Content

Overview

The Islandora Paged Content module is required by the Book Solution Pack and Newspaper Solution Pack modules, to provide numbered, individual pages as objects. This module takes files in TIFF format, and is able to create several kinds of derivatives depending on the type of collection they are being ingested into. Solution Packs that use the Paged Content module are referred to below under 'Provisions'.

Dependencies

- Islandora
- Tuque
- The Large Image Solution Pack is required to create image derivatives
- Ghostscript is used to compile PDF derivatives into a single document

Optional

- pdftotext
- pdfinfo

Install in Ubuntu/Debian with `sudo apt-get install poppler-utils`

Provisions

- The Book Solution Pack or Newspaper Solution Pack are examples of Paged Content collections. It is advisable to install one of those solution packs, and check their pages for additional dependencies.

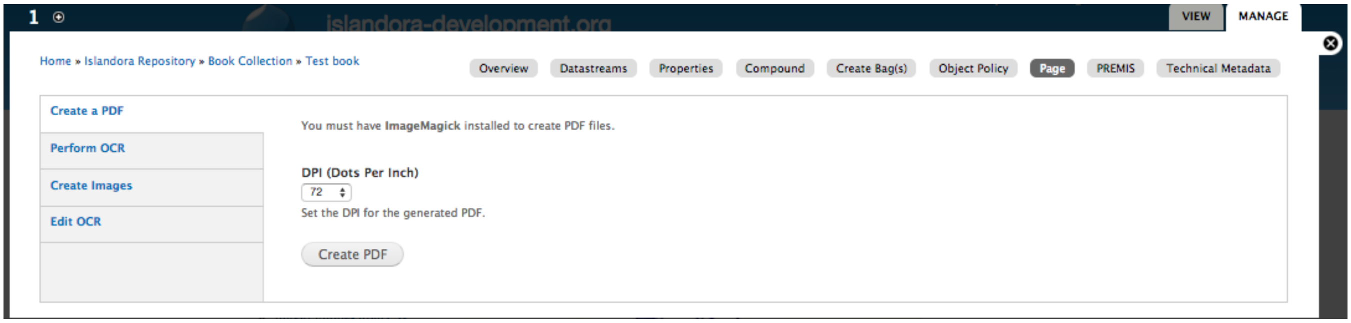
Downloads

Release Notes and Downloads

Usage

Adding a collection that uses the Paged Content module (such as Book and Newspaper Solution Pack) adds a new button to the end of the 'Manage' tab - 'Book', for the Book Solution Pack, and 'Newspaper' for the Newspaper Solution Pack, for example. It also adds a 'Page' button to the 'Manage' tab of any page objects added to a Paged Content collection.

Clicking on the "Manage" button of a page object will bring up several options, depending on what components of the module are selected and enabled:



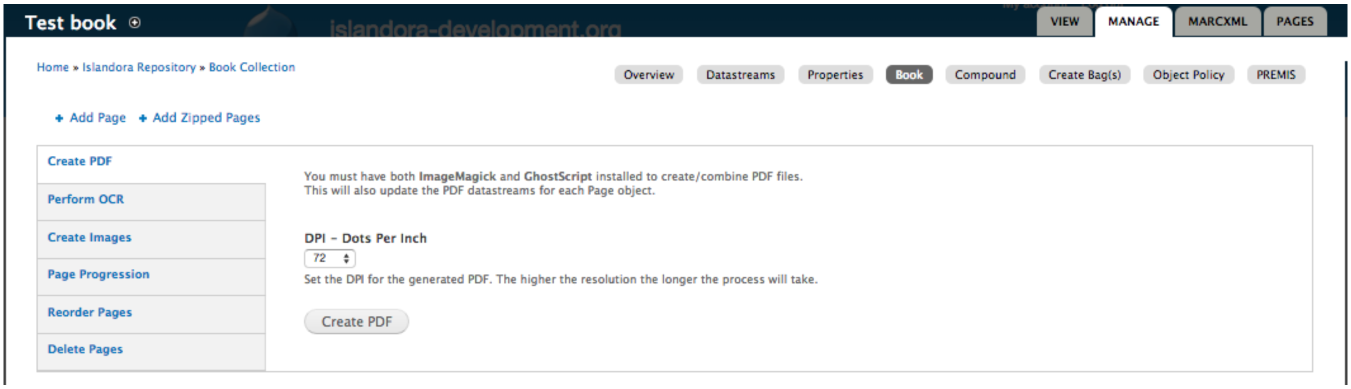
You must have **ImageMagick** installed to create PDF files.

DPI (Dots Per Inch)
72

Set the DPI for the generated PDF.

- **Create PDF** - This section includes the ability to create either a single page PDF if selected from a single page's 'Manage' tab, or a PDF of an entire Paged Content collection if selected from the collection's 'Manage' tab. The resolution of the image can also be set here. Creating a PDF will overwrite any existing PDF datastream.
- **Perform OCR** - This section includes the ability to create OCR datastreams for a single page if selected from that page's 'Manage' tab, or OCR datastreams for an entire Paged Content collection if selected if selected from the collection's 'Manage' tab. If multiple languages are installed into Tesseract, the option to switch between them will also be given here. Creating new OCR datastreams will overwrite any existing ones.
- **Create Images** - This section adds the option to create image derivatives if the Large Image Solution Pack is installed. Any existing image derivatives will be overwritten if this is used. If this option is selected from a Paged Content collection, the option will be given to create a thumbnail image for the collection from the first ordered page, updating and overwriting any existing thumbnail.
- **Edit OCR** - This allows you to manually edit the OCR datastream.

You can also perform these actions against a batch of Page objects, and also **Reorder Pages** and **Delete Pages**, from the "Manage" > "Book" tab of a Book object:

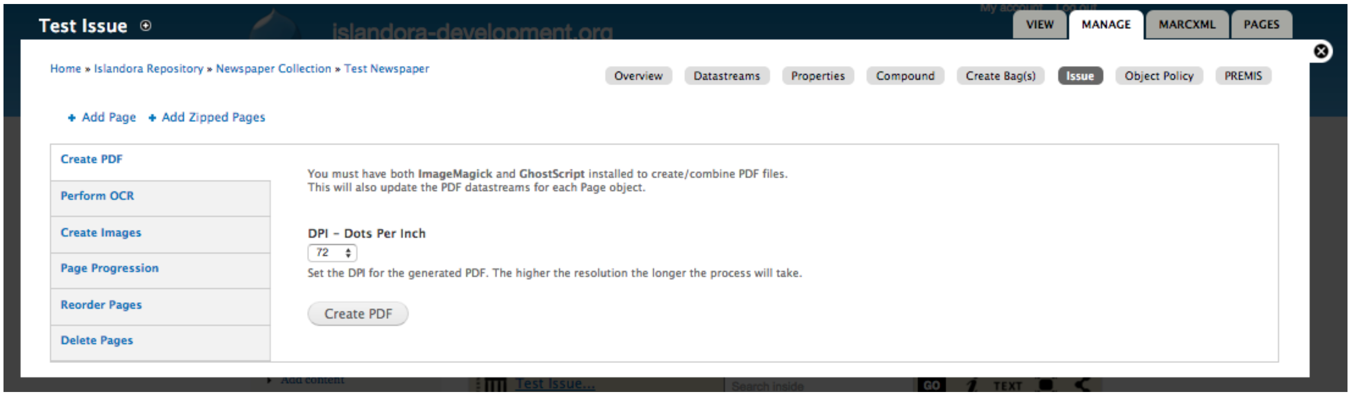


You must have both **ImageMagick** and **GhostScript** installed to create/combine PDF files.
This will also update the PDF datastreams for each Page object.

DPI - Dots Per Inch: 72

Set the DPI for the generated PDF. The higher the resolution the longer the process will take.

You can also perform these actions against a batch of Page objects, and also **Reorder Pages**, **Delete Pages**, and alter the **Page Progress** from the "Manage" > "Issue" tab of a Newspaper object:



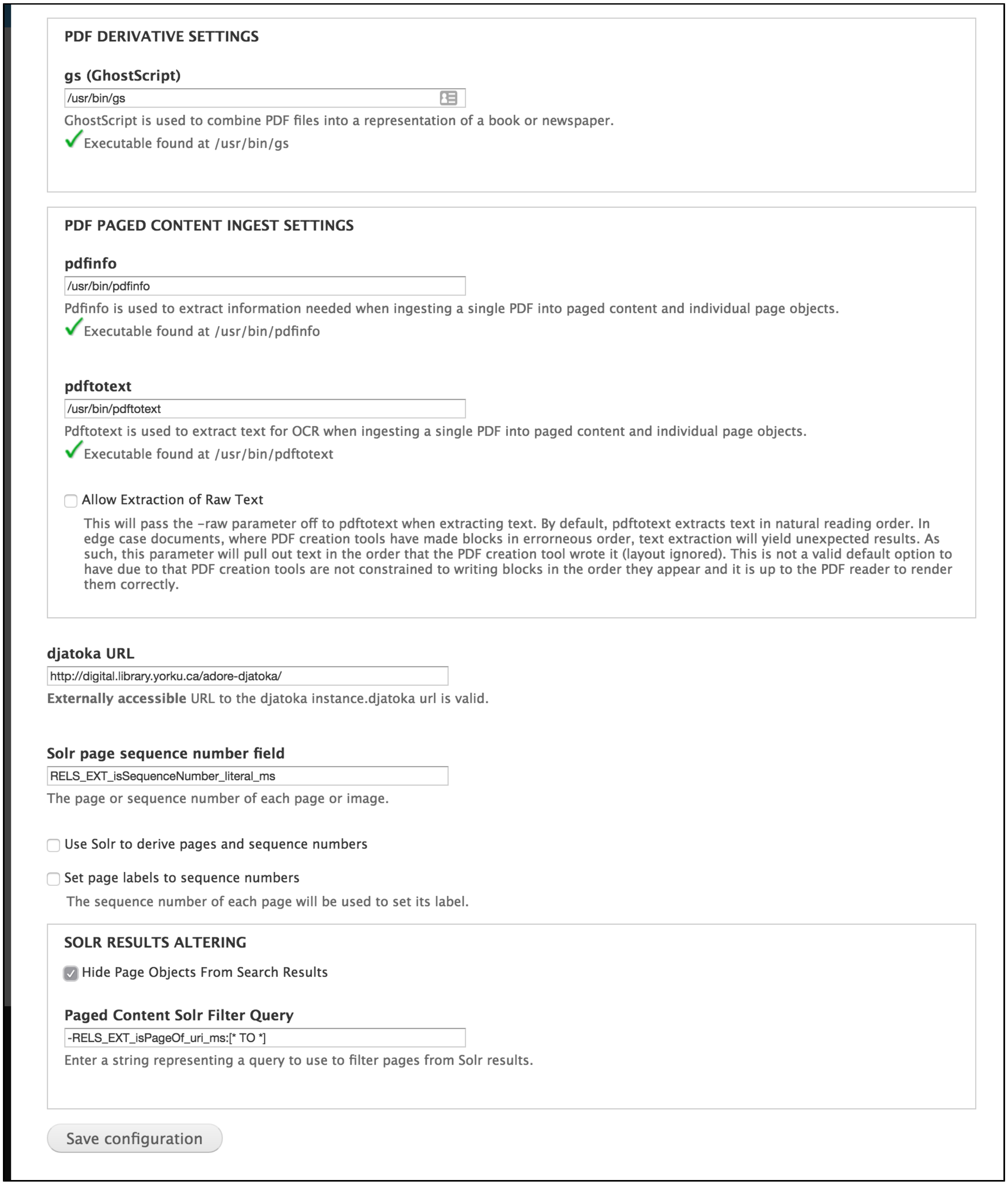
You must have both **ImageMagick** and **GhostScript** installed to create/combine PDF files.
This will also update the PDF datastreams for each Page object.

DPI - Dots Per Inch: 72

Set the DPI for the generated PDF. The higher the resolution the longer the process will take.

Configuration

Few configuration options exist for the Paged content module out-of-the-box. Most of the configuration is associated with the relevant, dependent solution pack (Book or Newspaper). The configuration page at Administration > Islandora > Solution pack configuration > Paged Content Module (admin/islandora/solution_pack_config/paged_content) has the following options:



PDF DERIVATIVE SETTINGS

gs (GhostScript)

/usr/bin/gs

GhostScript is used to combine PDF files into a representation of a book or newspaper.

✓ Executable found at /usr/bin/gs

PDF PAGED CONTENT INGEST SETTINGS

pdfinfo

/usr/bin/pdfinfo

Pdfinfo is used to extract information needed when ingesting a single PDF into paged content and individual page objects.

✓ Executable found at /usr/bin/pdfinfo

pdftotext

/usr/bin/pdftotext

Pdftotext is used to extract text for OCR when ingesting a single PDF into paged content and individual page objects.

✓ Executable found at /usr/bin/pdftotext

□ Allow Extraction of Raw Text

This will pass the –raw parameter off to pdftotext when extracting text. By default, pdftotext extracts text in natural reading order. In edge case documents, where PDF creation tools have made blocks in errorneous order, text extraction will yield unexpected results. As such, this parameter will pull out text in the order that the PDF creation tool wrote it (layout ignored). This is not a valid default option to have due to that PDF creation tools are not constrained to writing blocks in the order they appear and it is up to the PDF reader to render them correctly.

djatoka URL

http://digital.library.yorku.ca/adore-djatoka/

Externally accessible URL to the djatoka instance.djatoka url is valid.

Solr page sequence number field

RELS_EXT_isSequenceNumber_literal_ms

The page or sequence number of each page or image.

□ Use Solr to derive pages and sequence numbers

□ Set page labels to sequence numbers

The sequence number of each page will be used to set its label.

SOLR RESULTS ALTERING

☑ Hide Page Objects From Search Results

Paged Content Solr Filter Query

-RELS_EXT_isPageOf_uri_ms:[* TO *]

Enter a string representing a query to use to filter pages from Solr results.

PDF Derivative Settings

Enter the path to the Ghostscript executable here. This will allow multi-page PDFs to be compiled using each page in the book or newspaper. More information about installing Ghostscript on your server can be found at the official project website, http://www.ghostscript.com/.

There is also an option to set the page label to the page's sequence number. On ingest, each page's label will be set to its sequence number. When reordering pages, all of the page labels will be updated with the new sequence numbers.

SOLR Results

The Paged Content Module also allows you to filter individual page objects from your search results. If you use this option, you will want to make sure you enable the "Aggregate OCR?" option when ingesting your book. That appends a consolidated OCR datastream to the book object, which allows it to be searched and returned.

Content Models, Prescribed Datastreams and Forms

The Paged Content Solution Pack comes with the following objects in http://**path.to.your.site**/admin/islandora/solution_pack_config/solution_packs:

- Islandora Page Content Model (islandora:pageCModel)

A page image ingested into a Paged Content collection using ImageMagick, the Large Image Solution Pack and the Islandora OCR modules, will have the following datastreams:

OBJ	Original TIFF or JP2 file uploaded
DC	Dublin Core record
PDF	PDF derivative created by Ghostscript
JP2	JPEG 2000 derivative created by ImageMagick or Kakadu
JPG	Smaller JPEG derivative created by ImageMagick
TN	Thumbnail icon created from the image during the ingest process
RELS-INT	Internal Fedora relationship metadata defining the dimensions of the JP2 datastream
OCR	The raw output from Tesseract
HOCR	A converted version of the OCR datastream, intended to be more human-readable
RELS-EXT	Default Fedora relationship metadata

The Paged Content Solution Pack does not come with any forms.