PDF Solution Pack

Overview

The PDF Solution Pack module adds functionality to Islandora for ingesting and viewing PDF files. It uses the ImageMagick library and module to create derivative thumbnail and preview images. Because of the text-based nature of PDF files, it can also be used to create or append easily searchable text datastreams to the object, which can later be configured through Solr to appear in searches.

Dependencies

- Islandora
- Tuque
- ImageMagick is required to create derivatives. (Debian/Ubuntu `sudo apt-get install imagemagick`)
- pdftotext is required to automatically create a FULL_TEXT data stream. (Debian/Ubuntu `sudo apt-get install poppler-utils`)
- ghostscript (Debian/Ubuntu `sudo apt-get install ghostscript`)
- ImageMagick Drupal module
    - ensure that the full path to Imagemagick's convert is specified in the Image Toolkit (admin/config/media/image-toolkit)

Downloads

Release Notes and Downloads

Configuration

The configuration options for the PDF Solution Pack module can be found at http://**path.to.your.site**/admin/islandora/solution_pack_config/pdf, and include the following:

Text

Users can either upload a text file of their own, or allow Islandora to extract one from the PDF. Text accompanying the PDF is stored as the FULL_TEXT datastream. If both options are checked under the **Text** configuration section, and a valid path to pdftotext is entered, preference will be given to a supplied text file on ingest.

- **Allow users to upload text file with PDFs**: This file must be plain text stored in .txt format.
- **Extract text streams from PDFs using pdftotext**: Checking this box, will display an option to enter a path to the pdftotext executable. This package is not native to most server setups and will need to be installed manually for this option to be functional. Check the pdftotext dependency page for more information.
- **Create PDF/A archival derivative from PDF:** Create a PDF/A version of any uploaded PDF. PDF/A is a restrictive standard that prohibits more easily broken components of the PDF spec, such as fillable forms and DRM. The PDF/A derivative will not be used for display. Requires ghostscript to be installed on the server.
- **Use dUseCIEColor when generating PDFA datastream:** whether the dUseCIEColor switch should be used for GhostScript when creating a PDF/A version. Not recommended for GhostScript versions 9.11 or higher.

Thumbnail and Preview

These options set the width, height, and colorspace parameters that will be sent to ImageMagick when generating Thumbnail and Preview derivatives for the PDF. ImageMagick will attempt to create these using the first page of the document. Changing the width or height will change the image size, but not the aspect ratio, of the derivatives being created.

PDF Solution Pack

TEXT

- [ ] Allow users to upload .txt files with PDFs

  Uploaded text files are appended to PDFs as FULL_TEXT datastreams and are indexed into Solr.

- [x] Extract text streams from PDFs using pdftotext

  Extracted text streams are appended to PDFs as FULL_TEXT datastreams and are indexed into Solr. Uploading a text file takes priority over text stream extraction.
  Note: PDFs that contain visible text do not necessarily contain text streams (e.g. images scanned and saved as PDFs). Consider converting text-filled images with no text streams to TIFFs and using the Book Solution Pack with OCR enabled.

- [x] Create PDF/A archival derivative from PDF

  Create a PDF/A version of any uploaded PDF. PDF/A is a restrictive standard that prohibits more easily broken components of the PDF spec, such as fillable forms and DRM. The PDF/A derivative will not be used for display. Requires ghostscript to be installed on the server.

- [ ] Use dUseCIEColor when generating PDFA datastream.

  As of GhostScript 9.11, the use of the dUseCIEColor switch is not recommended. See https://ghostscript.com/pipermail/gs-devel/2014-July/009693.html.
  Version installed: 9.26.

Path to pdftotext executable

/usr/bin/pdftotext

✓ pdftotext executable found at **/usr/bin/pdftotext**

Path to ghostscript executable

/usr/bin/gs

✓ ghostscript executable found at **/usr/bin/gs**

THUMBNAIL

Settings for creating PDF thumbnail derivatives

Width

200

The width of the thumbnail in pixels.

Height

200

The height of the thumbnail in pixels.

Colorspace

sRGB

The colorspace for the derived thumbnail.

PREVIEW IMAGE

Settings for creating PDF preview image derivatives

Max width

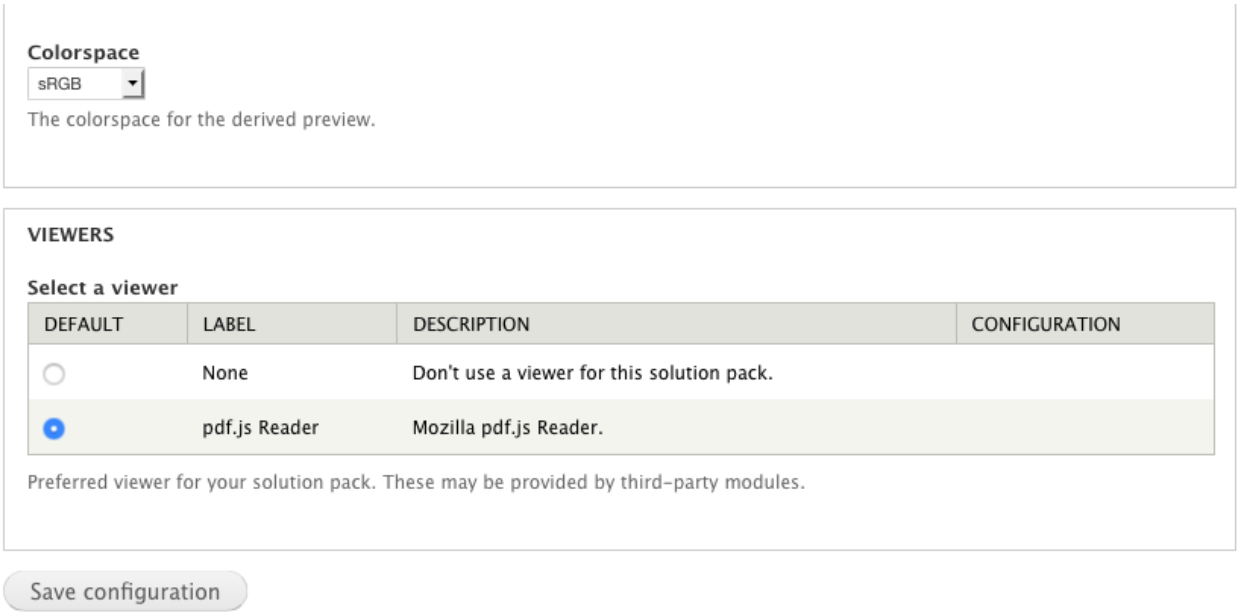
500

The maximum width of the preview in pixels.

Max height

700

The maximum height of the preview in pixels.



Viewers

The PDF Solution Pack can utilize the PDF.js viewer to display PDF documents inline. To enable, navigate to the PDF Solution Pack's configuration page (admin/islandora/solution_pack_config/pdf) and select the PDF.js as the viewer.

Content Models, Prescribed Datastreams and Forms

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

- Islandora PDF Content Model (islandora:sp_pdf)
- PDF Collection (islandora:sp_pdf_collection)

An object created using the PDF Solution Pack's content model will have the following datastreams:

RELS-EXT	Default Fedora relationship metadata
MODS	MODS metadata record created during ingest
DC	Dublin Core record
OBJ	Original PDF file uploaded
TN	Thumbnail image created by ImageMagick during ingest
PREVIEW	Preview image created by ImageMagick during ingest
FULL_TEXT	Optional datastream either uploaded during ingest, or created by the pdftotext executable
PDFA	Optional archival datastream created by the ghostscript executable

The PDF Solution Pack comes with the PDF MODS Form.