Islandora Book Batch

Introduction

This module implements a batch framework for importing books into Islandora.

The ingest is a two-step process:

- Preprocessing: The data is scanned and a number of entries are created in the Drupal database. There is minimal processing done at this point, so preprocessing can be completed outside of a batch process.
- Ingest: The data is actually processed and ingested. This happens inside of a Drupal batch.

Requirements

This module requires the following modules/libraries:

- Islandora
- Tuque
- Islandora Batch
- Book Solution Pack

Installation

Install module as usual, see this for further information.

Configuration

N/A

Usage

The base ZIP/directory preprocessor can be called as a drush script (see `drush help islandora_book_batch_preprocess` for additional parameters):

Books must be broken up into separate directories, such that each directory at the "top" level (in the target directory or Zip file) represents a book. Book pages are their own directories inside of each book directory containing an OBJ and additional datastreams can be added here manually.

Files are assigned to object datastreams based on their basename, so a folder structure like:

Note: A Metadata file (*.xml OR *.mrc) and a Object file is required, all other datastreams are optional.

Option 1

```
/tmp/batch_ingest/
  book/
       1/
          OBJ.tif
       2/
          OBJ.tif
          OCR.asc
          HOCR.shtml
       DC.xml
```

Option 2

```
/tmp/batch_ingest/
 book/
     1/
        DC.xml
        OBJ.tif
     2/
         DC.xml
         OBJ.tif
```

Either would result in a two-page book.

Each page directory name will be used as the sequence number of the page created.

A file named --METADATA--.xml can contain either MODS, DC or MARCXML which is used to fill in the MODS or DC streams (if not provided explicitly). Similarly, --METADATA--.mrc (containing binary MARC) will be transformed to MODS and then possibly to DC, if neither are provided explicitly.

If no MODS is provided at the book level - either directly as MODS.xml, or transformed from either a DC.xml or the "--METADATA--" file discussed above - the directory name will be used as the title.

Drush made the `target` parameter reserved as of Drush 7. To allow for backwards compatibility this will be preserved.

Drush 7 and above: (Examples of Zip and Directory batch preprocessing)

```
drush -v -u 1 --uri=http://localhost islandora_book_batch_preprocess --type=zip --scan_target=/path/to
/archive.zip

drush -v -u 1 --uri=http://localhost islandora_book_batch_preprocess --namespace=book --type=directory --
scan_target=/tmp/batch_ingest/
```

Drush 6 and below: (Examples of Zip and Directory batch preprocessing)

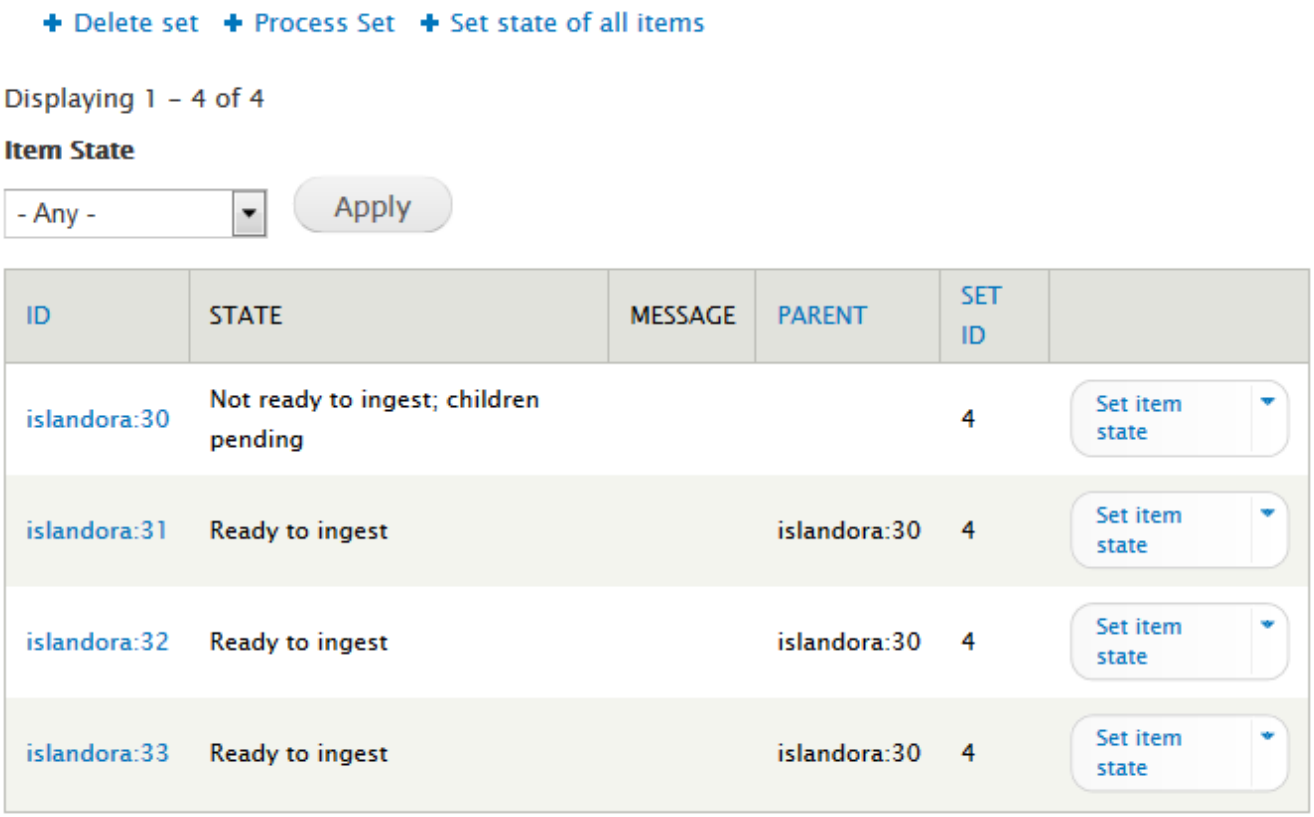
```
drush -v -u 1 --uri=http://localhost islandora_book_batch_preprocess --type=zip --target=/path/to/archive.zip

drush -v -u 1 --uri=http://localhost islandora_book_batch_preprocess --namespace=book --type=directory --
target=/tmp/batch_ingest/
```

This will populate the queue (stored in the Drupal database) with base entries for an administrator to approve and start the processing. The queue of preprocessed items can then be processed either through a drush command or the admin console.

Drush(6 and 7):

```
drush -v --user=admin --uri=http://localhost islandora_batch_ingest
```

To approve the batch, go to Administration > Reports > Islandora Batch Sets and select "View Items in Set" next to an unprocessed set. To process the set, click "Process Set" and process all items.



Customization

Custom ingests can be written by extending any of the existing preprocessors and batch object implementations.

Troubleshooting/Issues

Having problems or solved a problem? Check out the Islandora google groups for a solution.

- Islandora Group
- Islandora Dev Group

Maintainers/Sponsors

Current maintainers:

- Rosie Le Faive

Development

If you would like to contribute to this module, please check out CONTRIBUTING.md. In addition, we have helpful Documentation for Developers info, as well as our Developers section on the Islandora.ca site.

License

GPLv3

Additional Usage

Drush 7

Drush made the `target` parameter reserved as of Drush 7. To allow for backwards compatibility, this will be preserved. The parameter has been renamed `scan_target`.

Web options

Options are available for processing include:

- Page progressions (left-to-right or right-to-left)
- Create PDF
- Namespace (default is repository's default namesapce)
- Generate OCR (run Tesseract on individual pages)
- Aggregate OCR (place individually generated OCR pages into one text file, whose datastream will be listed with the **book object** itself . This is helpful for searching and allows repository managers to disable individual page listings in Solr output.)
- Notify admin after ingest (if Rules have been set)
- Ingest immediately (ingest or delay queue for processing)

Command-line Book Batch options interaction with the Book Solution Pack ingest settings:

```
Options:
 --aggregate_ocr                          A flag to cause OCR to be aggregated to books, if OCR is also being
generated per-page.
 --content_models                         A comma-separated list of content models to assign to the objects.
Only applies to the "book" level object.
 --create_pdfs                            A flag to cause PDFs to be created in books. Page PDF creation is
dependant on the configuration within Drupal proper.
 --directory_dedup                        A flag to indicate that we should avoid repreprocessing books which
are located in directories.
 --do_not_generate_hocr                   A flag to allow for conditional HOCR generation.
 --do_not_generate_ocr                    A flag to allow for conditional OCR generation.
 --email_admin                            A flag to notify the site admin when the book is fully ingested
(depends on Rules being enabled).
 --namespace                              The namespace for objects created by this command.  Defaults to
namespce set in fedora config.
 --output_set_id                          A flag to indicate whether to print the set ID of the preprocessed
book.
 --page_progression                       A flag to indicate the page progression for the book. If not
specified will default to LR.
 --parent                                 The collection to which the generated items should be added.  Only
applies to the "book" level object. If "directory" and the directory containing the book description is a valid
PID, it will be set as the parent. If this is specified and itself is a PID, all books will be related to the
given PID.
 --parent_relationship_pred               The predicate of the relationship to the parent. Defaults to
"isMemberOfCollection".
 --parent_relationship_uri                The namespace URI of the relationship to the parent. Defaults to
"info:fedora/fedora-system:def/relations-external#".
 --target                                 The target to directory or zip file to scan. Required.
 --type                                   Either "directory" or "zip". Required.
 --wait_for_metadata                      A flag to indicate that we should hold off on trying to ingest books
until we have metadata available for themat the book level.

Aliases: ibbp


Example that will work inside of islandora_vagrant:
/var/www/drupal$ drush -v -u 1 --uri=http://localhost islandora_book_batch_preprocess --
content_models=islandora:bookCModel --namespace=islandora --parent=islandora:1 --type=directory --target=
/vagrant/dir_of_books --create_pdfs=TRUE
```

- The options --create_pdfs and --aggregate_ocr will have no effect if the box for the corresponding option is not checked on the Book Solution Pack configuration page (admin/islandora/solution_pack_config/book).
- So, if "PDF datasteam" is checked on the SP configuration page, then the option --create_pdfs will create a book-level (aggregated) PDF datastream.
- Likewise, if "OCR datastreams" is checked, then the option --aggregate_ocr will create a book-level (aggregated) OCR datastream.