Islandora Newspaper Batch

Introduction

The Islandora Newspaper Batch module uses the Islandora Batch framework to provide a **command-line** (drush) and a **GUI** (Drupal interface) option for adding a batch file of newspaper issues and pages to an existing Islandora Newspaper object.

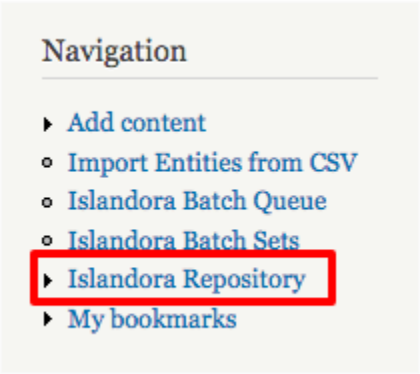
Batch-loading newspapers is a two-step process.

1. Preprocessing: Drupal creates entries in the database for each object (issue and page) that will be added.
2. Ingest: The data is ingested and derivatives are generated as part of the Islandora batch functions.

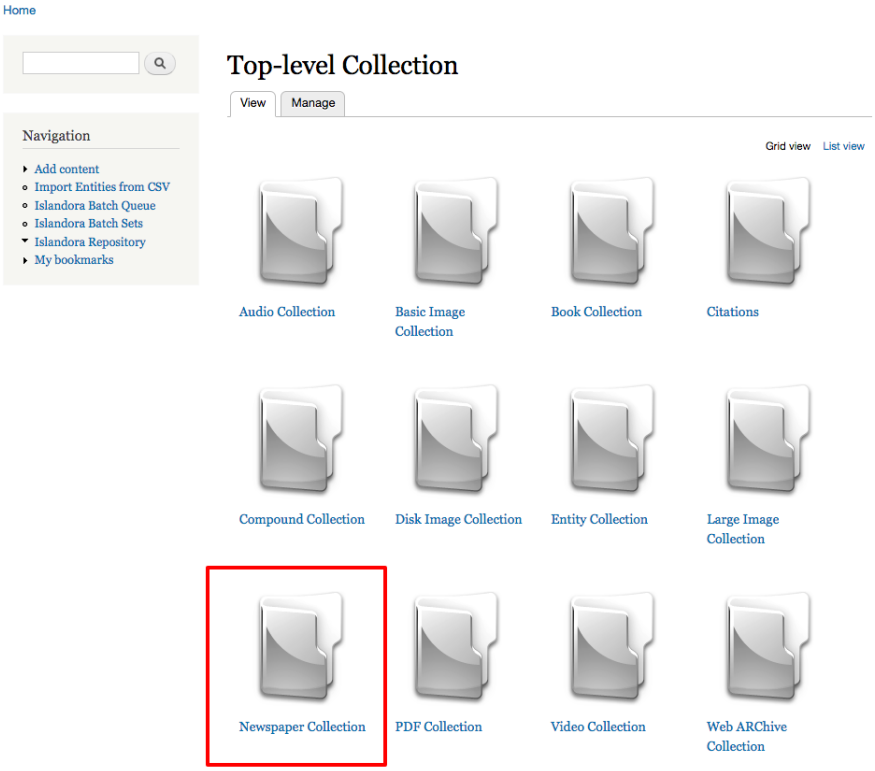
Newspaper Batch uses the value in the MODS dateIssued field on each issue to populate the issue browsing display for newspaper. The data in this field must be formatted as YYYY-MM-DD. If only YYYY is entered, the interface will use the current month and day for the issue.

Creating a new Newspaper object

- Newspaper Batch can only be used with an existing Newspaper object (islandora:newspaperCModel). To create a new Newspaper object:
  - Go to http://localhost:8000 and log in
  - Navigation > Islandora Repository

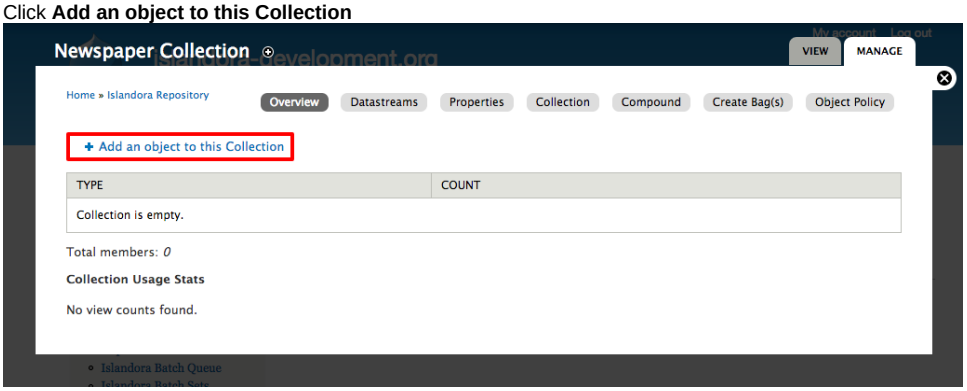


  - Click on the **Newspaper Collection**

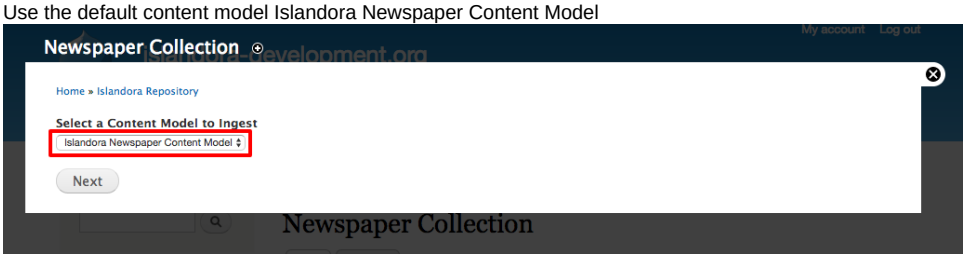


- Click **Manage** tab

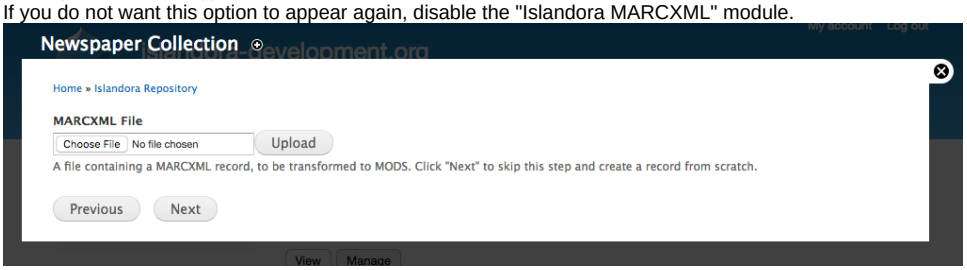
- Click **Add an object to this Collection**



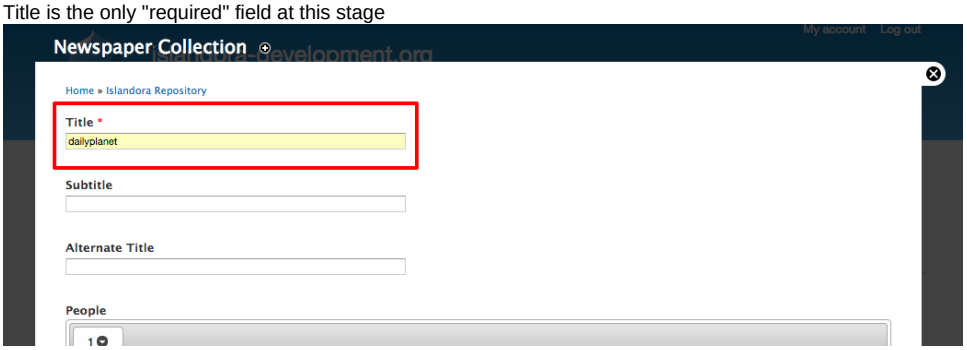
- Use the default content model Islandora Newspaper Content Model



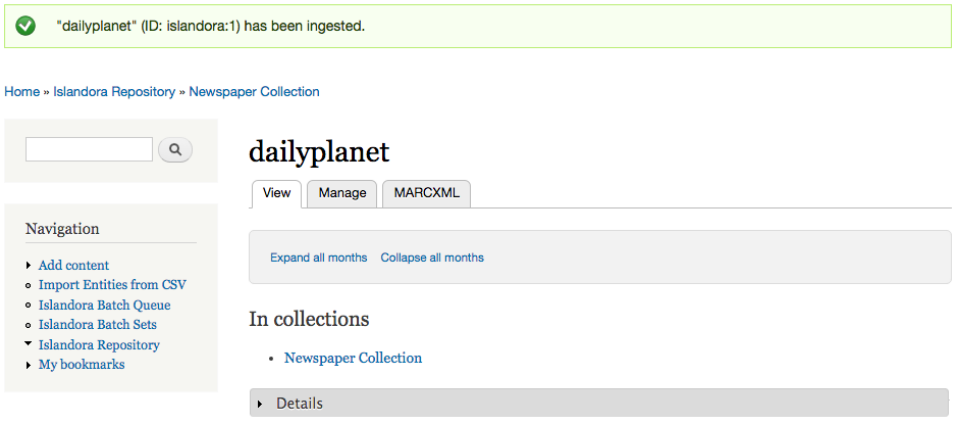
- If you have MARCXML to submit, select the file, upload it, and click Next. If you do not have MARCXML, just click **Next** (MARCXML is not required at this step).
- If you do not want this option to appear again, disable the "Islandora MARCXML" module.



- Title is the only "required" field at this stage



- Click ingest and it should confirm your ingest



Preparing files for batch ingest

The Newspaper Batch module is designed for digitized newspapers where **each page is represented by an individual TIFF image file**. These TIFF files, along with derivatives, full text, and metadata, must be arranged in directories according to a very specific structure.

Tips for preparing batch ingest files

- Generally, Islandora performs best with ZIP files smaller than 2 GB. If your files are larger, consider using drush with the --type=directory option.
- Within the zip file or target directory (if using drush and the --type-directory option), each top-level directory represents a newspaper issue.
- Files within the issue directory will become datastreams on the issue object (e.g. this is where you put issue-level metadata including a MODS file with the date of the issue).
- Directories within the issue directory contain files that will become newspaper page objects. These are usually named numerically, as they are processed in numerical order.
- File names must match their respective Islandora datastream IDs. This means that every page image (tiff) needs to be renamed "OBJ.tif". If you have created derivatives for the page objects, these can be named respectively (e.g. TN.jpg, OCR.txt, ...)

Sample single-issue batch folder hierarchy

```
batch.zip

 issue1
     001
        OBJ.tif
     002
        OBJ.tif
     MODS.xml - this becomes the MODS record for the issue-level object
```

Sample batch folder hierarchy with derivatives

```
batch02.zip

 issue1
     001
        JP2.jp2
        JPG.jpg
        OBJ.tif
        OCR.txt
        TN.jpg
     2
        JP2.jp2
        JPG.jpg
        OBJ.tif
        OCR.txt
        TN.jpg
     MODS.xml
```

Descriptive Metadata

If MODS metadata is not available for issue or page objects, the following formats can be supplied and will be automatically transformed to general MODS and DC.

- DC.xml
- MARCXML.xml
- MARC.mrc

Other things to note about metadata:

- If no MODS is provided or available from transformations for the issue-level object, the directory name (issue01 in the example above) will be used as the issue title.
- The issue browser function on each newspaper object page is populated by the "dateIssued" field in MODS from each issue record. This date must be present and formatted as YYYY-MM-DD.
- Known issue: if only YYYY is included in the dateIssued field, the batch ingest will supply a month and day matching the current date, which may result in incorrect descriptive metadata.

Sample Issue-level MODS.xml file

Here is a sample MODS file describing a newspaper issue.

```
<?xml version="1.0" encoding="UTF-8"?>
<mods xmlns="http://www.loc.gov/mods/v3" xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xsi="http://www.w3.org/2001
/XMLSchema-instance" xmlns:xlink="http://www.w3.org/1999/xlink">
    <titleInfo>
      <title>Canadian Jewish Review, June 1, 1928</title>
    </titleInfo>
    <originInfo>
      <place>
        <placeTerm>Toronto, Ontario</placeTerm>
      </place>
      <publisher>Canadian Jewish Review </publisher>
      <dateIssued encoding="iso8601">1928-06-01</dateIssued>
    </originInfo>
    <language>
      <languageTerm>eng</languageTerm>
    </language>
    <subject>
      <topic>Jews, Canadian -- Ontario -- Toronto -- History -- Newspapers</topic>
      <topic>Jews, Canadian -- Quebec -- Montreal -- History -- Newspapers</topic>
      <topic>Jews -- History -- 20th century -- Newspapers</topic>
      <topic>Jews -- Canada -- Periodicals</topic>
      <topic>Canada -- History -- 20th century -- Newspapers</topic>
      <topic>Ontario -- History -- 20th century -- Newspapers</topic>
      <topic>Quebec -- History -- 20th century -- Newspapers</topic>
      <topic>Toronto (Ont.) -- History -- 20th century -- Newspapers</topic>
      <topic>Montreal (Que.) -- History -- 20th century -- Newspapers</topic>
    </subject>
    <identifier>Cjewish-1928-06-01</identifier>
</mods>
```

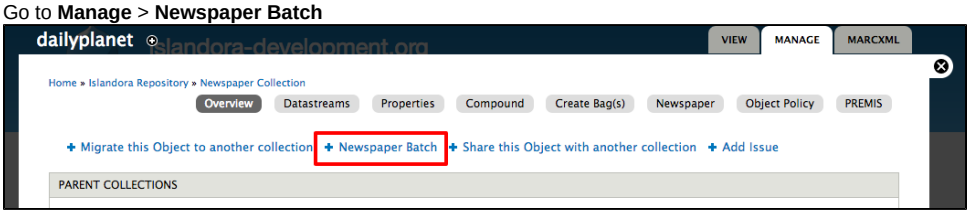
Using Newspaper Batch from the Drupal interface

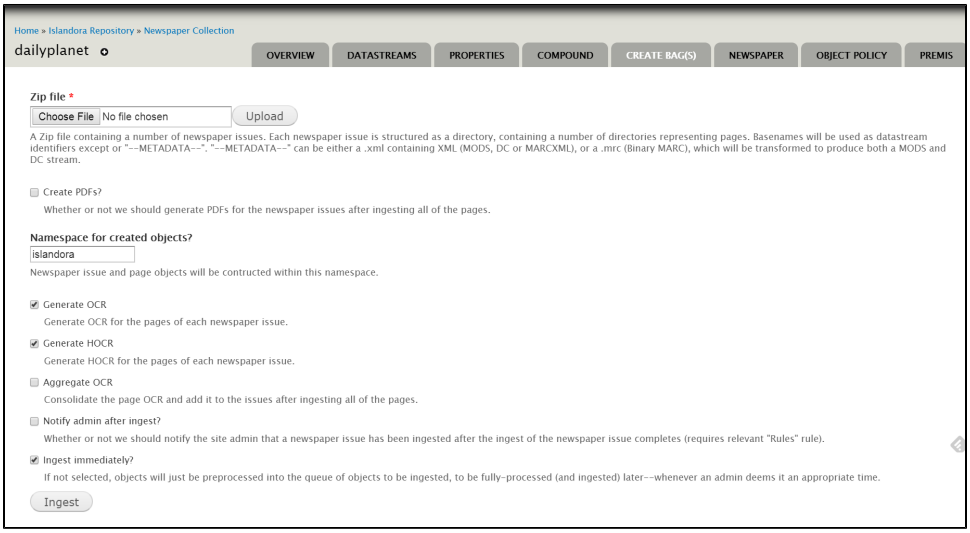
To use Newspaper Batch in Islandora:

1. Log in as a user with batch ingest permissions.
2. Navigate to a Newspaper Content Model object and click **Manage**.
3. In the **Overview** tab, click **Newspaper Batch**.
4. Upload the ZIP file and set the appropriate options for this batch, then click **Ingest**.

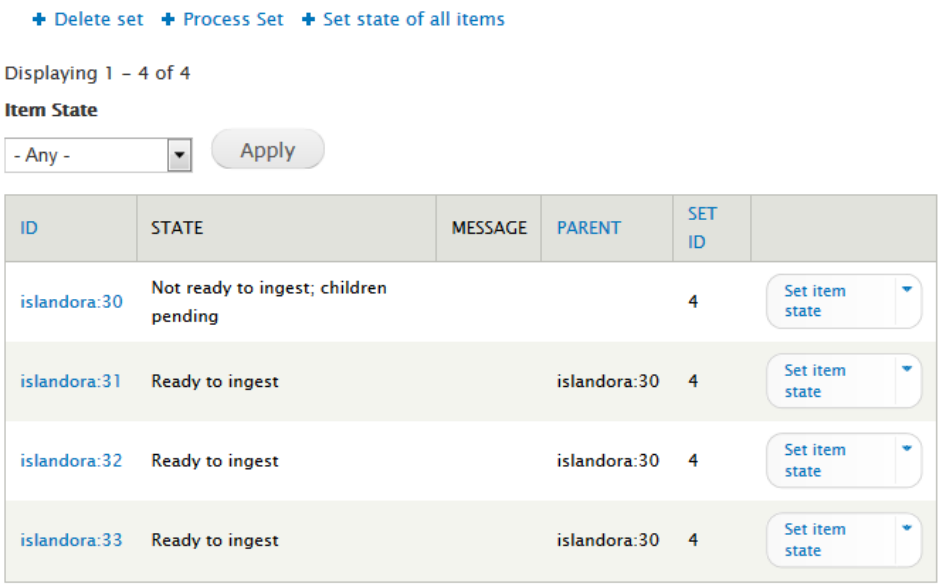
Newspaper Batch Ingest options

- Go to **Manage** > **Newspaper Batch**





- **Zip file** - Upload the ZIP file for batch ingest.
- **Create PDFs?** - Checking this box creates a PDF derivative that contains all the pages associated with a newspaper issue.
- **Namespace for created objects** - Set the namespace for the issue and page objects created for this batch ingest.
- **Generate OCR?** - Checking this box causes OCR to be generated for each Page object. OCR will be attached as a datastream to each page. If checked, another option appears below it, "Aggregate OCR?".
- **Generate HOCR?** - Checking this box causes HOCR to be generated for each Page object (text highlighting after full text search). HOCR will be attached as a datastream to each page.
- **Aggregate OCR?** Check this box to create an OCR datastream in the issue object that aggregates the OCR datastreams from all of the page-level objects in that issue.
- **Notify admin after ingest?** - Check this box to send an email to the site admin (user 1) that a newspaper batch ingest has completed. This requires the Drupal Rules module and a rule for newspaper batch notifications.
- **Ingest immediately?** - Checking this box will cause the batch to go through both steps of the ingest (pre-processing and actual ingest) immediately.
    - If you do not check "Ingest Immediately", the files will be pre-processed only and added to the Islandora batch queue for an administrator to approve.
    - 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.



Using Newspaper Batch from the command line (Drush)

If you have many ZIP files to ingest, or if the ZIP files are too large to ingest through the interface, you can also batch ingest newspapers from the Drupal command line with Drush.

First, your file(s) need to be accessible by the Drush instance. That usually means that they need to be uploaded to the Islandora server (scp, ftp, using a mounted storage drive, etc). The --scan_target option (--target option in Drush 6 and above) is either a directory of issue directories, or a zip file of

issue directories. That is, the directories representing the issues to ingest (or issue, if only one) must be one level below the directory or zip file used as the --scan_target.

Second, preprocess the file(s). For a full list of the command-line parameters, see "drush help islandora_newspaper_batch_preprocess". The batch options are also described in the Islandora Batch module.

```
drush -v -u 1 --uri=http://localhost islandora_newspaper_batch_preprocess --type=directory --scan_target=/path/to
/issues --namespace=dailyplanet --parent=islandora:dailyplanet
```

This will populate the queue (stored in the Drupal database) with PID entries. Note that the --parent parameter must be a newspaper title object, not an issue object or a collection object.

Here are the options in the drush command:

```
drush help islandora_newspaper_batch_preprocess

Preprocessed newspaper issues into database entries.


Options:

 --aggregate_ocr                          A flag to cause OCR to be aggregated to issues, 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 "newspaper issue"

                                          level
object.

 --create_pdfs                            A flag to cause PDFs to be created in newspaper issues. Page PDF
creation is dependant on the configuration

                                          within Drupal
proper.

 --directory_dedup                        A flag to indicate that we should avoid repreprocessing newspaper
issues which are located in directories.

 --do_not_generate_ocr                    A flag to allow for conditional OCR
generation.

 --email_admin                            A flag to notify the site admin when the newspaper issue is fully
ingested (depends on Rules being enabled).

 --namespace                              The namespace for objects created by this command.  Defaults to
namespace set in Fedora config.

 --parent                                 The collection to which the generated items should be added.  Only
applies to the "newspaper issue" level

                                          object. If "directory" and the directory containing the newspaper
issue description is a valid PID, it will

                                          be set as the parent. If this is specified and itself is a PID, all
newspapers issue will be related to the

                                          given PID.
Required.

 --parent_relationship_pred               The predicate of the relationship to the parent. Defaults to
"isMemberOf".

 --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
newspaper issues until we have metadata

                                          available for them at the newspaper issue level.


Aliases: inbp
```

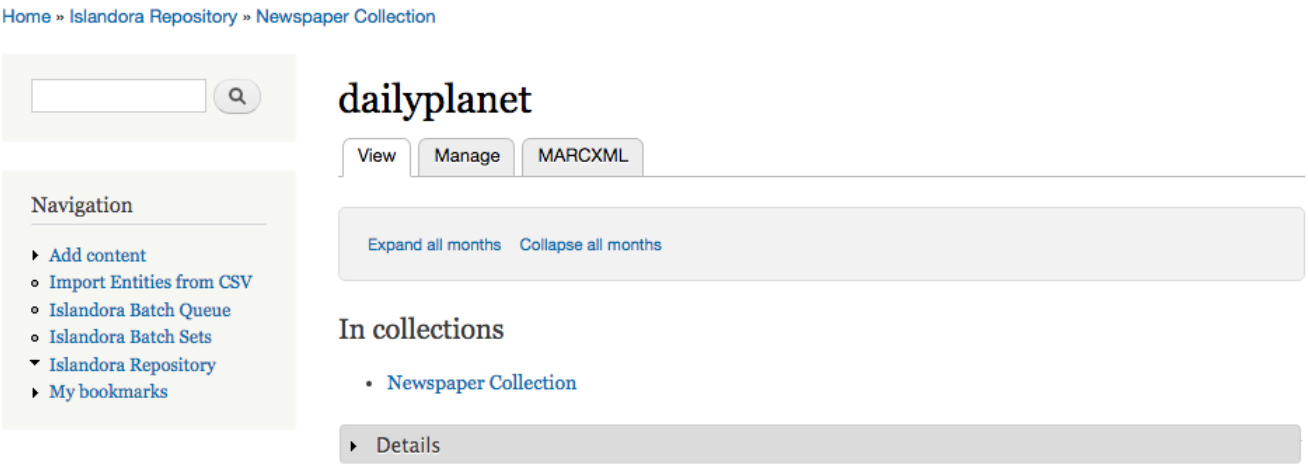
Third, process all items in the batch queue:

```
drush -v -u 1 --uri=http://localhost islandora_batch_ingest
```

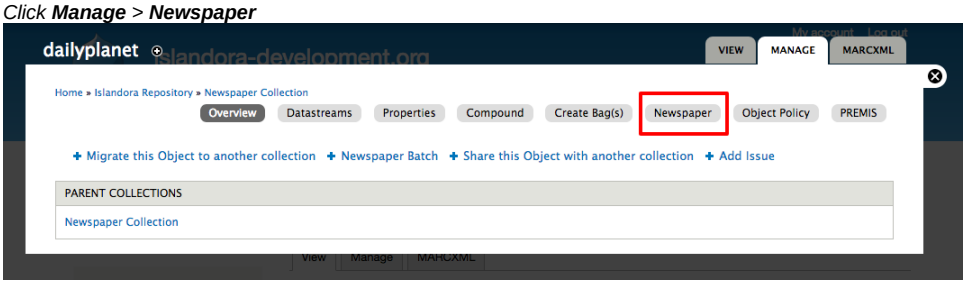
Troubleshooting

*You may get a warning. "Failed to get issued date from MODS for dailyplanet:1"
*

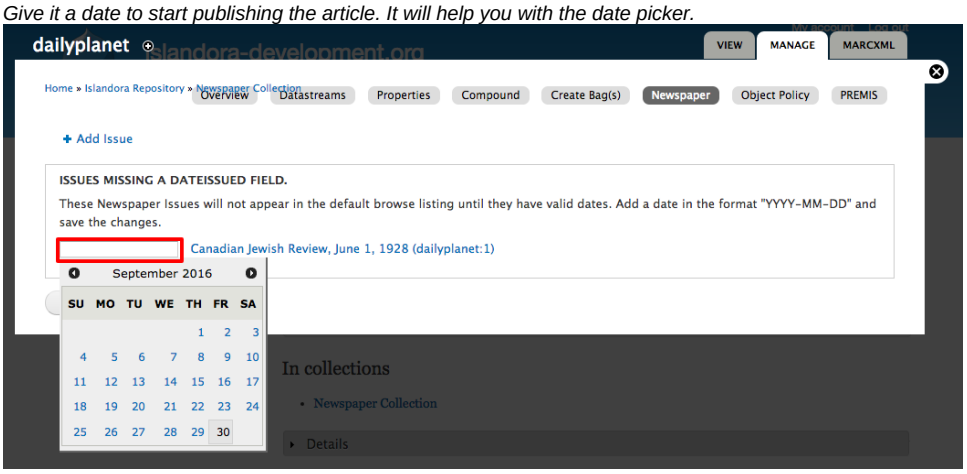
After ingesting everything looks normal but the "issue" you ingested is missing.



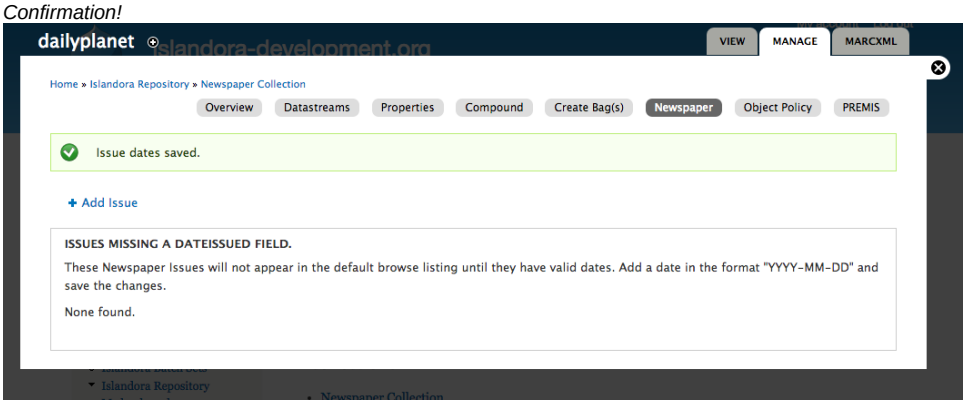
- *Click **Manage** > **Newspaper***



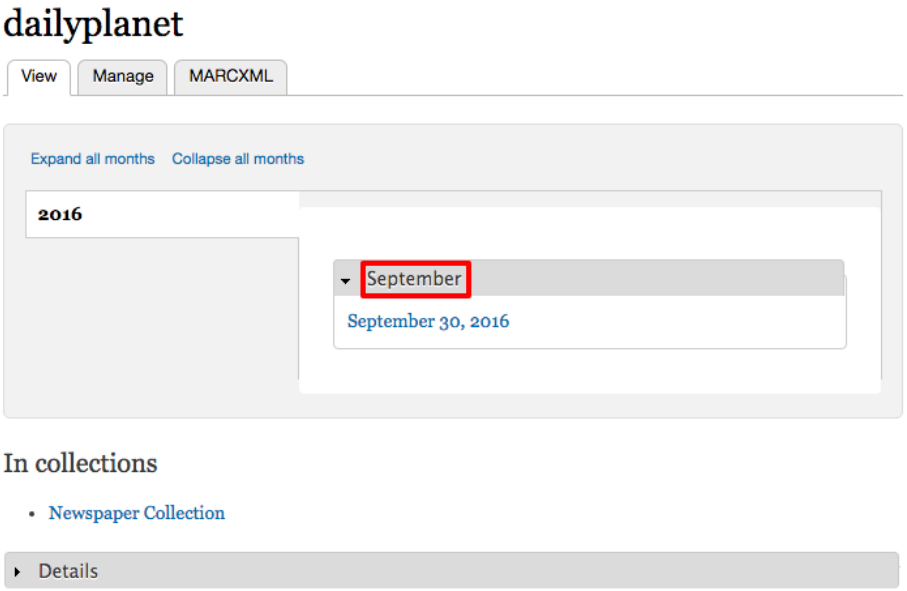
- *Give it a date to start publishing the article. It will help you with the date picker.*



- *Confirmation!*



- *Now you'll see a date on the Newspaper page*



Additional Documentation

Further documentation for this module is available at the Islandora Newspaper Batch github repository.