

How to Batch Ingest Files

Before you can batch ingest objects, you will need to have downloaded and installed the [Islandora Importer](#) module (web interface) or the [Islandora Batch](#) module (command line). If you want to batch ingest books, you will need to have downloaded and installed the [Islandora Book Batch](#) module; if you want to batch ingest newspaper issues, you will need to have downloaded and installed the [Islandora Newspaper Batch](#) module. It is also strongly encouraged that you review the `mods_to_dc.xsl` within the Islandora Book Batch module if you plan to ingest MODS metadata. Reviewing the `mods_to_dc.xsl` will help you to understand what type of Dublin Core will be produced by the `mods_to_dc.xsl`. For example, you may notice that the `mods_to_dc.xsl` will not produce clean Dublin Core subject tags - all individual MODS subject tags will be expressed as one Dublin Core subject tab. The `mods_to_dc.xsl` will also not map names tags if no `roleTerm` with a `type` attribute has been specified. The `mods_to_dc.xsl` is a Library of Congress XSLT and the Islandora community does not make modifications to this file. You are encouraged to make your own edits to the `mods_to_dc.xsl` if you need to modify the XSLT.

For larger collections, Islandora is able to pull multiple files out of a zipped archive and ingest them into Fedora as a batch. There are a few ways that this can be done. You can upload .zip archives full of:

- [files \(content\) such as images or PDFs](#)
- [XML metadata that can later have content files added to it](#)
- [both content files to be ingested and XML metadata to be appended to files](#)
- [books, formatted with a specific directory structure](#)
- [newspaper issues and pages, formatted with a specific directory structure](#)
- [batch ingest cleanup](#)

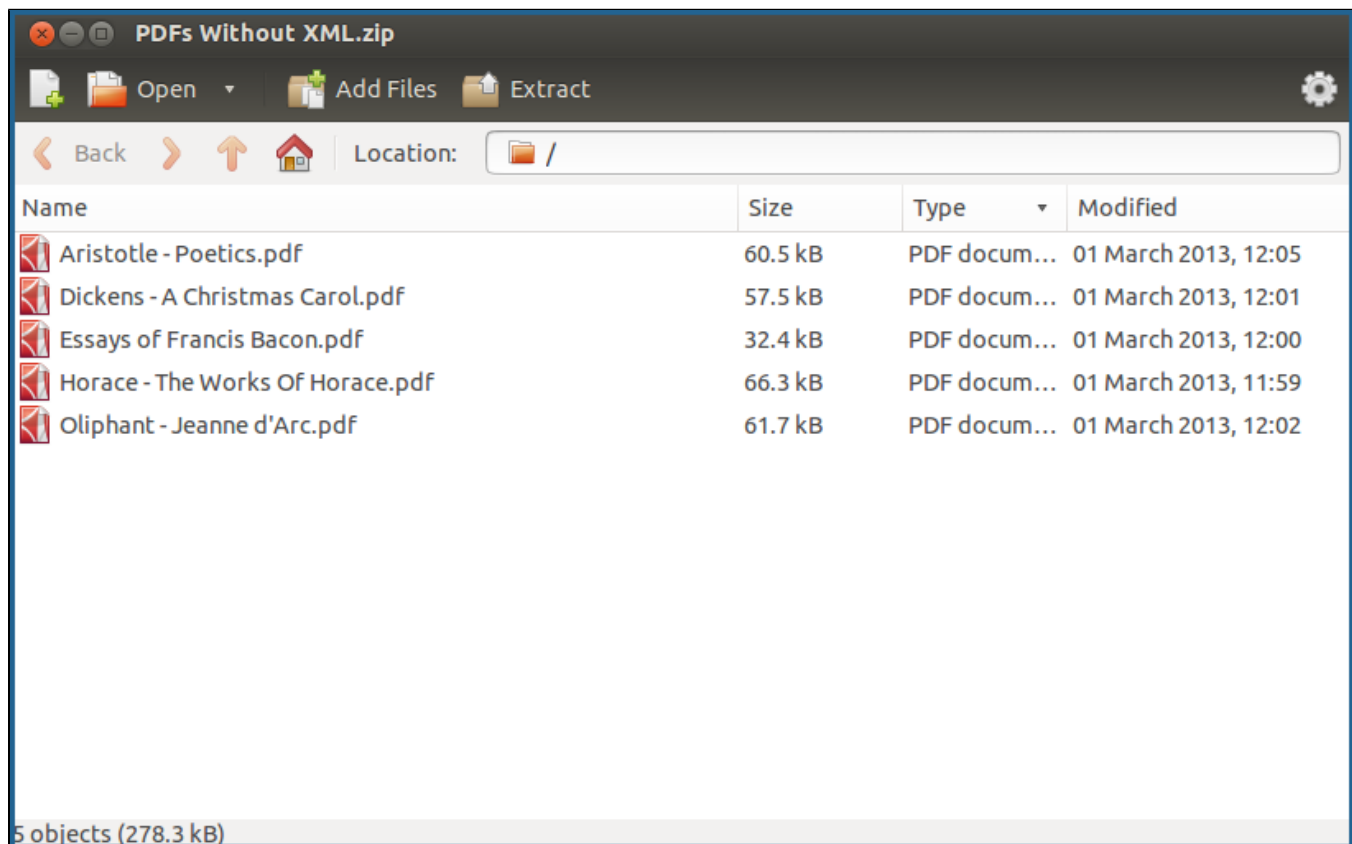
This page will run through the specifics of each one. In these examples, we will be batch-ingesting PDF files into a collection with the 'PDF Solution Pack' content model applied to its collection policy.

Ingest zipped content

1. Create a .zip archive with your files in it

The process for doing this will vary from operating system to operating system, but on PC, Mac and Linux at least, a .zip archive can be made in your file browser by highlighting the file or files you would like to zip, right-clicking, and finding an option similar to 'compress', 'create archive', 'create zipped folder', and so on.

In our example, opening the zipped archive shows our PDFs grouped together:

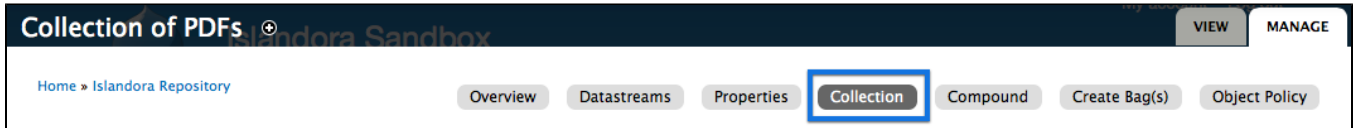


2. Navigate to the destination collection and click 'Manage'



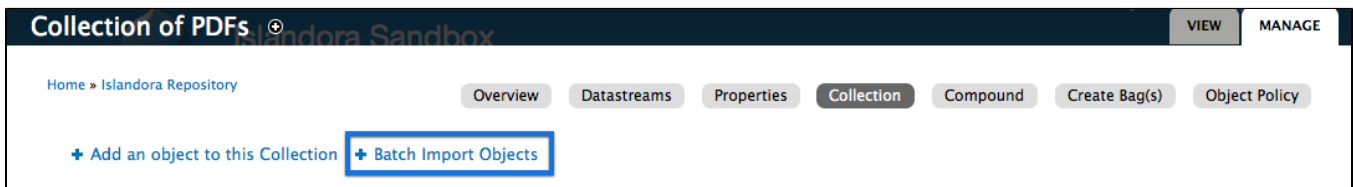
This will take you to the collection's management page.

3. Click on the 'Collection' button



This will take you to that collection's options page.

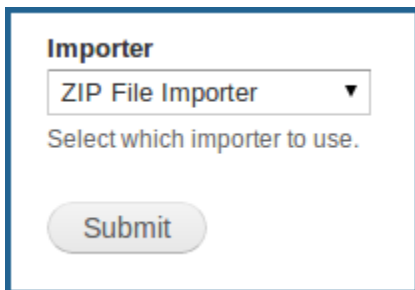
4. Click on 'Batch Import Objects'



This will start the batch import process.

5. If other batch import modules are active, choose 'ZIP File Importer' from the drop-down menu

The Islandora Batch Importer module comes with several different modules for handling different types of content. If more than one is enabled, you will need to select the correct one.



In this case, we will be importing objects from a .zip file, so we are going to select that option.

6. Choose the correct options for your batch import

There are a few options on this screen that will need to be set up:

ZIP BATCH IMPORTER

Select the file containing the assets and metadata to import. Assets and metadata will be matched together based on the portion of the filename without the extension, so my_file.xml and my_file.pdf will be combined into a single object.

Zip file of files to import

1

CONTENT MODEL

The content model(s) to assign to the imported objects.

<input type="checkbox"/>	NAME
<input type="checkbox"/>	Islandora Collection Model ~ islandora:collectionCModel
<input checked="" type="checkbox"/> 2	Islandora PDF Content Model

Object Namespace

3

The namespace in which the imported objects will be created.

4

1. Browse to the .zip file you would like to upload, and then click the 'Upload' button. It may take a while to move the file to the server.
2. Choose the content models you would like to apply to the objects.

All checked content models are applied to **each** object. So you can't mix different types of objects in one .zip file.

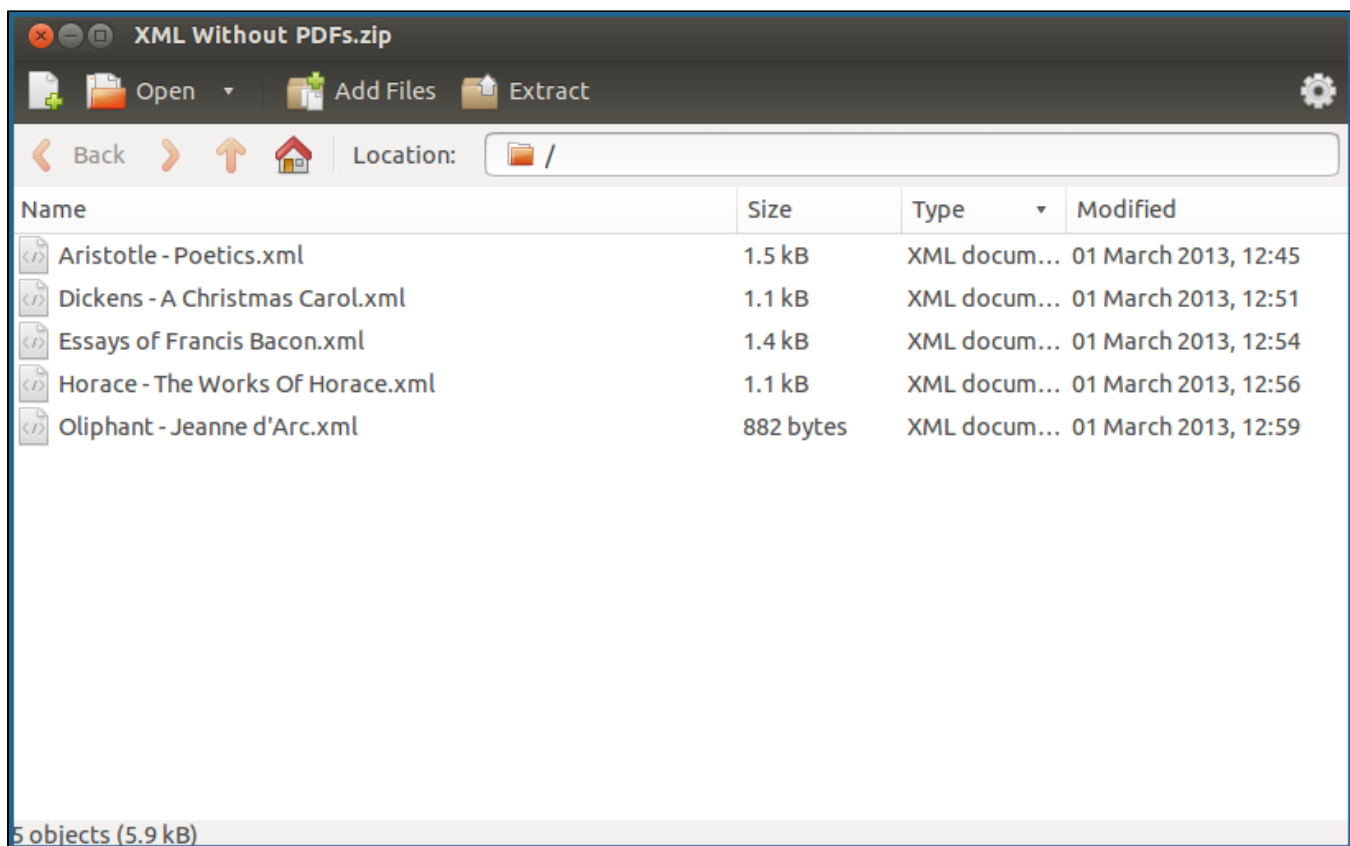
3. Choose the namespace to be applied to the objects. ("Islandora" is given only as an example.)
4. Click the 'Import' button to begin the batch import process.

This will import all the files from your zipped archive during which new Fedora objects are created and associated with the specified collection.

Ingest XML files with metadata

If you wish to ingest objects as simply metadata without a file datastream attached, you may do so by using the batch importer to import a .zip file full of XML forms.

In our example, a .zip file full of XML files has been created. Once the metadata files are fully ingested, PDF files can be added to the objects as datastreams.



In this case, we can simply follow the same steps as in the previous example to perform the batch ingest.

Most Solution Packs look for a file with a MIME type associated with the corresponding content model. In these cases, when uploading XML-only .zip files, the Book Batch Importer will show an error saying that derivatives could not be created. This can be ignored in these cases.

To create the XML files, you can either design them manually in a text editor or XML editor or use the Form Builder built into Islandora.

To use the Form Builder to create XML records, navigate to <http://path.to.your.site/admin/islandora/xmlform>, find the type of form you would like to fill out, and click the 'view' link beside it:

+ Create Form + Import Form					
TITLE	TYPE	OPERATIONS			
Audio MODS form	Built-in	Copy	View	Export	Associate
Basic image MODS form	Built-in	Copy	View	Export	Associate
Citation MODS form	Built-in	Copy	View	Export	Associate
Compound Object MODS form	Built-in	Copy	View	Export	Associate
Islandora Book MODS Form	Built-in	Copy	View	Export	Associate
Large image MODS form	Built-in	Copy	View	Export	Associate
Newspaper	Built-in	Copy	View	Export	Associate
Newspaper Issue	Built-in	Copy	View	Export	Associate
PDF MODS form	Built-in	Copy	View	Export	Associate
Thesis MODS form	Built-in	Copy	View	Export	Associate
Video MODS form	Built-in	Copy	View	Export	Associate
Web ARChive MODS form	Built-in	Copy	View	Export	Associate

In this case, we will be creating an XML file for a PDF.

Fill out the form with the metadata values you would like, and click the 'Submit' button at the bottom of the form. This will output raw XML to your browser that you can then paste into a text editor, similar to the following:

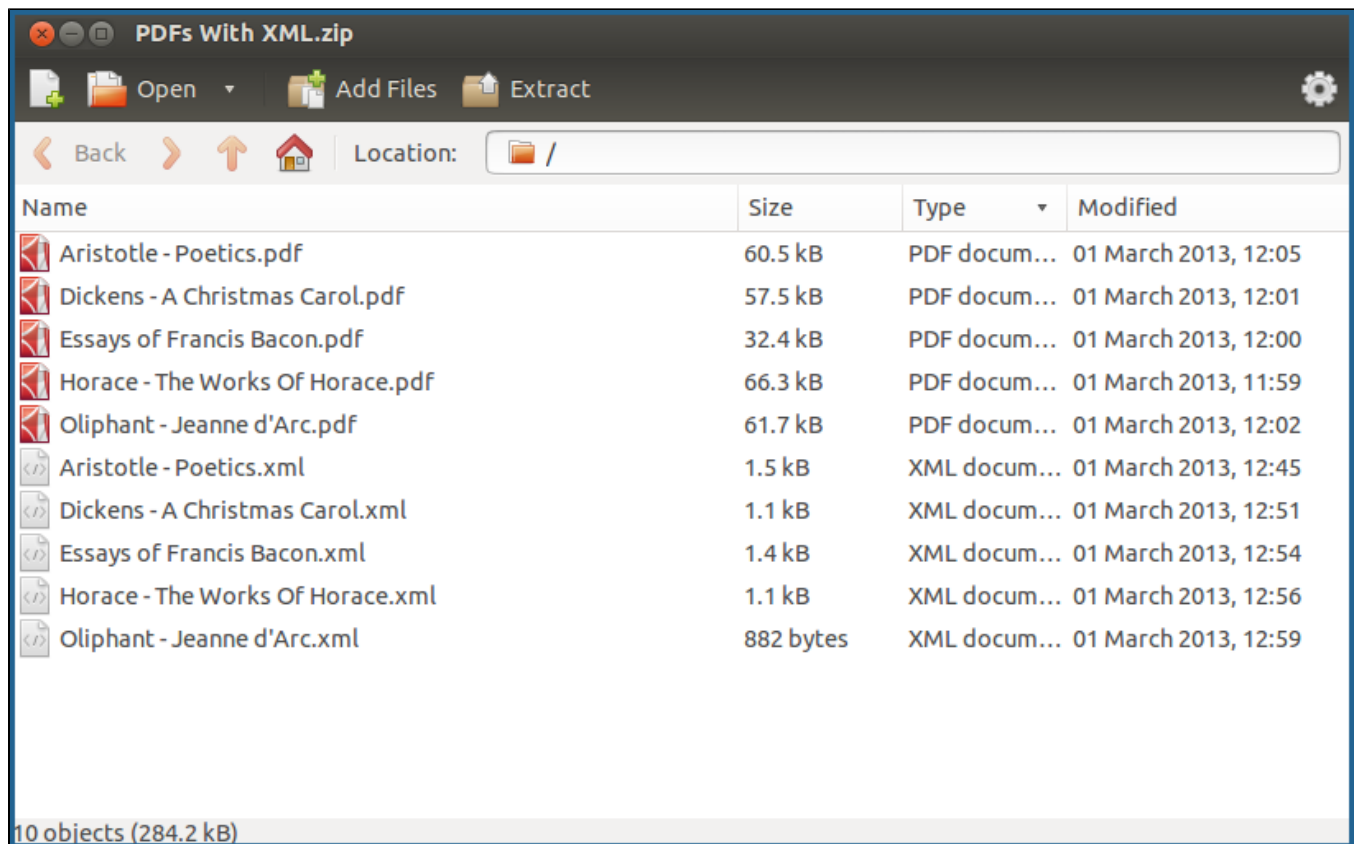
```
<?xml version="1.0"?>
<mods xmlns="http://www.loc.gov/mods/v3" xmlns:mods="http://www.loc.gov/mods/v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xlink="http://www.w3.org/1999/xlink">
  <titleInfo>
    <title>Poetics</title>
    <subTitle/>
  </titleInfo>
  <name type="personal">
    <namePart>Aristotle</namePart>
    <role>
      <roleTerm authority="marcrelator" type="text">Author</roleTerm>
    </role>
  </name>
  <typeOfResource>text</typeOfResource>
  <abstract>Aristotle's philosophical treatise on the subjects and methods of poetry.</abstract>
  <genre>article</genre>
  <note/>
  <subject>
    <topic/>
    <geographic/>
    <temporal/>
  </subject>
  <location>
    <url/>
  </location>
</mods>
```

Save this output as an XML file, and add it to your .zip archive.

Make sure you choose the correct content model that you would like to be associated with the files that will eventually be ingested

Ingest metadata and files at the same time

Zip archives can contain both the content files to be ingested and the metadata files at the same time. This can be accomplished in the same way as the first example, with a few changes to the .zip archive itself:





Name	Size	Type	Modified
Aristotle - Poetics.pdf	60.5 kB	PDF docum...	01 March 2013, 12:05
Dickens - A Christmas Carol.pdf	57.5 kB	PDF docum...	01 March 2013, 12:01
Essays of Francis Bacon.pdf	32.4 kB	PDF docum...	01 March 2013, 12:00
Horace - The Works Of Horace.pdf	66.3 kB	PDF docum...	01 March 2013, 11:59
Oliphant - Jeanne d'Arc.pdf	61.7 kB	PDF docum...	01 March 2013, 12:02
Aristotle - Poetics.xml	1.5 kB	XML docum...	01 March 2013, 12:45
Dickens - A Christmas Carol.xml	1.1 kB	XML docum...	01 March 2013, 12:51
Essays of Francis Bacon.xml	1.4 kB	XML docum...	01 March 2013, 12:54
Horace - The Works Of Horace.xml	1.1 kB	XML docum...	01 March 2013, 12:56
Oliphant - Jeanne d'Arc.xml	882 bytes	XML docum...	01 March 2013, 12:59

10 objects (284.2 kB)

In the above example, you will notice that each PDF file has a corresponding XML file, and that the filenames of the PDFs and the XML files are identical in every way – including capitalization – except for the extension. Files with matching filenames will be ingested together into the same object.

After creating a .zip archive like above, you can simply follow the steps from the first example to ingest the batch into the repository.

Uploading multiple datastreams in the PDF content model

 The PDF solution pack has an option to allow/disallow users to upload text files with PDFs for index into Solr. Note that this option applies to loading individual objects, not for batch processing: batch uploading PDF content model objects with multiple datastreams (such as uploading an object's PDF, XML, and full-text files through the zip importer) will effectively ignore this checkbox.  Not every file in the archive needs to have corresponding datastreams. You could potentially upload an archive that contains some objects without metadata, some objects with only metadata, and some objects with both.

Batch Ingest Books

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.

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

- my_cool_book/
 - MODS.xml
 - 001/
 - OBJ.tiff
 - 002/
 - OBJ.tiff

The above 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.

Text files for individual pages can also be supplied to provide a plain-text representation of the materials. For example, handwritten items can have a transcribed text file uploaded in the batch process as --TEXTFILE--.txt.

Batch Ingest Newspapers

When batch ingesting newspapers, you must already have an existing newspaper-level object. Each ingest folder contains folders that represent issues of the newspaper, and each issue directory contains folders that represent separate page images.

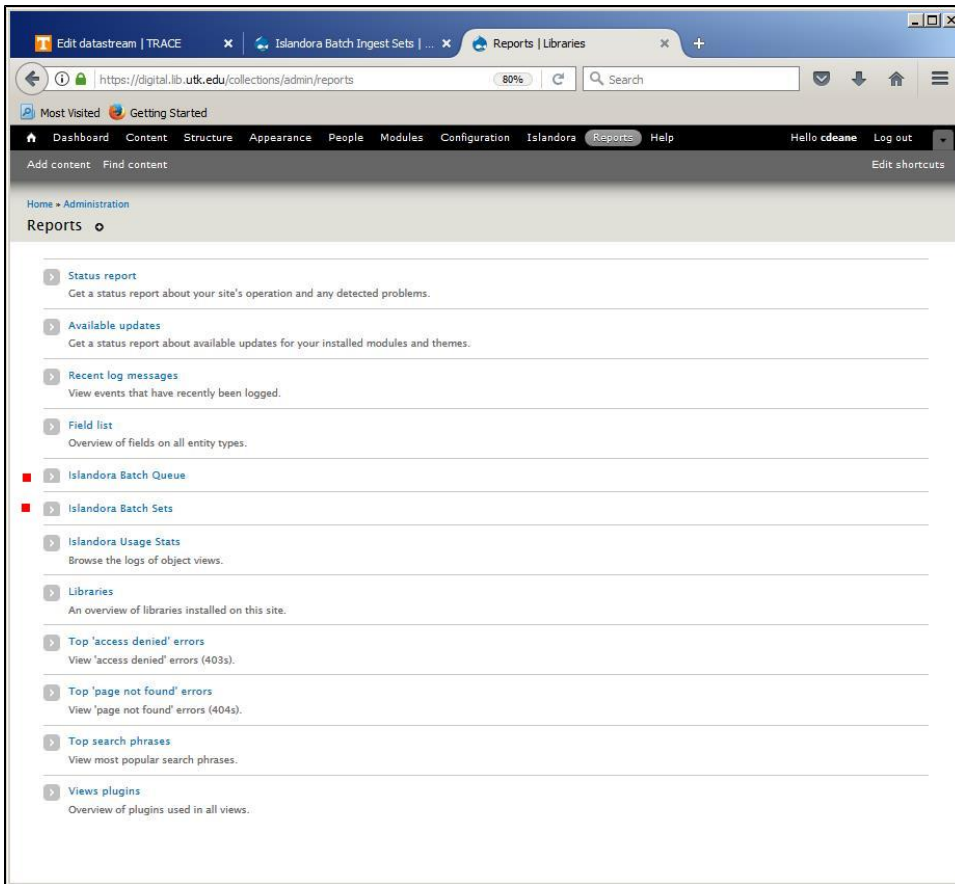
For sample directory structures and configuration options, see the [Newspaper Batch Ingest instructions](#).

Batch Ingest Cleanup

Islandora creates detailed reports for each Batch Ingest.

These reports can be very helpful for debugging and tracking, but they also take up hard drive space on your Islandora server.

- The easiest way to find these reports is to click on **Reports**.



- Click on the link for the report you want.

- **Islandora Batch Ingest Queue**
- Note that the first row has **SET ID 4**.

Displaying 1 - 18 of 18

Item State:

ID	STATE	MESSAGE	SET ID
cdf:10401	Ingested	Ingested cdf:10401.	4
cdf:10402	Ingested	Ingested cdf:10402.	5
cdf:10403	Ingested	Ingested cdf:10403.	6
cdf:10404	Ingested	Ingested cdf:10404.	7
cdf:10405	Ingested	Ingested cdf:10405.	8
cdf:10406	Ingested	Ingested cdf:10406.	9
cdf:10407	Ingested	Ingested cdf:10407.	10
cdf:10408	Ingested	Ingested cdf:10408.	11
cdf:10409	Ingested	Ingested cdf:10409.	12
cdf:10410	Ingested	Ingested cdf:10410.	13
scopes:2731	Ingested	Ingested scopes:2731.	36
scopes:2732	Ingested	Ingested scopes:2732.	36
scopes:2733	Ingested	Ingested scopes:2733.	36
scopes:2734	Ingested	Ingested scopes:2734.	36
scopes:2735	Ingested	Ingested scopes:2735.	37
scopes:2736	Ingested	Ingested scopes:2736.	37
scopes:2737	Ingested	Ingested scopes:2737.	37

- **Islandora Batch Ingest Sets**
- Note that the creator of each Batch Set is identified so your site admin can prod you to clean up your stuff!
- In this report, **SET ID 4** is at the bottom.

Displaying 1 - 13 of 13

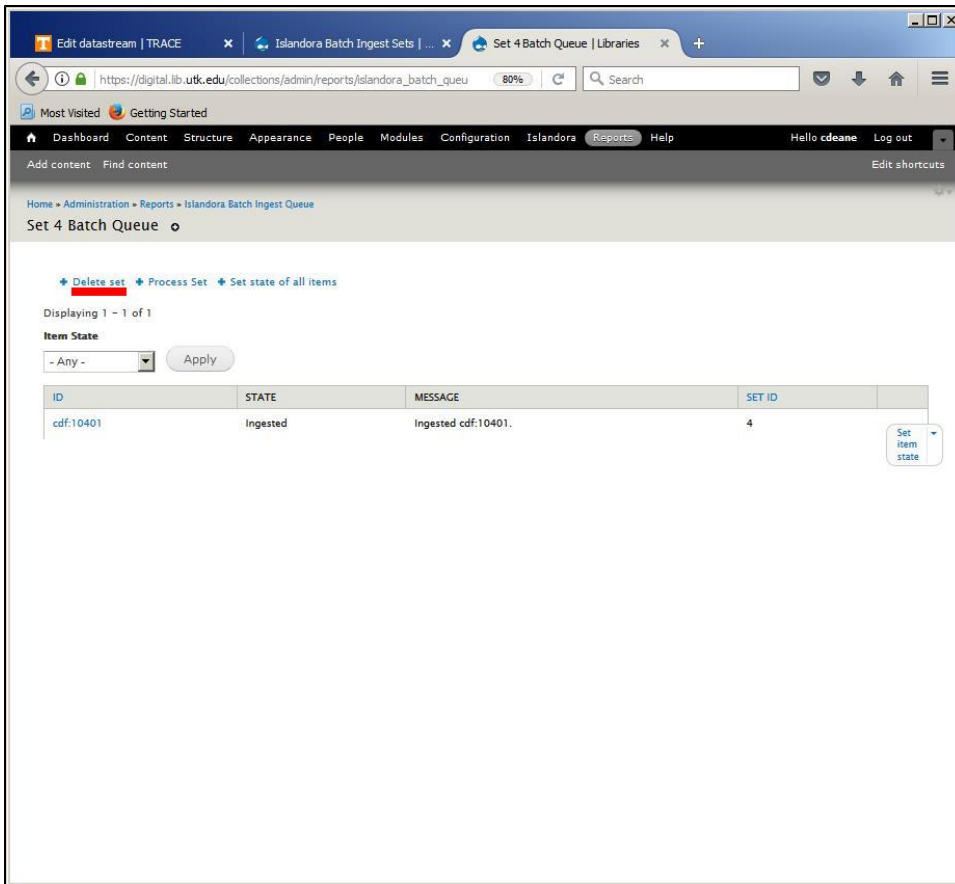
SET ID	SET CREATOR	CREATED DATE	SET SIZE
37	mbagget1	Wednesday, March 1, 2017 - 14:45	4
36	mbagget1	Wednesday, March 1, 2017 - 14:36	4
35	mbagget1	Wednesday, March 1, 2017 - 14:35	0
13	cdeane	Tuesday, February 21, 2017 - 14:55	1
12	cdeane	Tuesday, February 21, 2017 - 14:53	1
11	cdeane	Tuesday, February 21, 2017 - 14:51	1
10	cdeane	Tuesday, February 21, 2017 - 14:49	1
9	islandora	Tuesday, February 21, 2017 - 13:26	1
8	islandora	Tuesday, February 21, 2017 - 13:25	1
7	islandora	Tuesday, February 21, 2017 - 13:23	1
6	islandora	Tuesday, February 21, 2017 - 13:21	1
5	islandora	Tuesday, February 21, 2017 - 13:19	1
4	islandora	Tuesday, February 21, 2017 - 13:17	1

- The easiest way to clean out these files is to use the gui provided on the **Islandora Batch Ingest Sets** report.
- First click on the **dropdown menu** for the row you wish to delete. (Here, the row containing **SET ID 4**.)

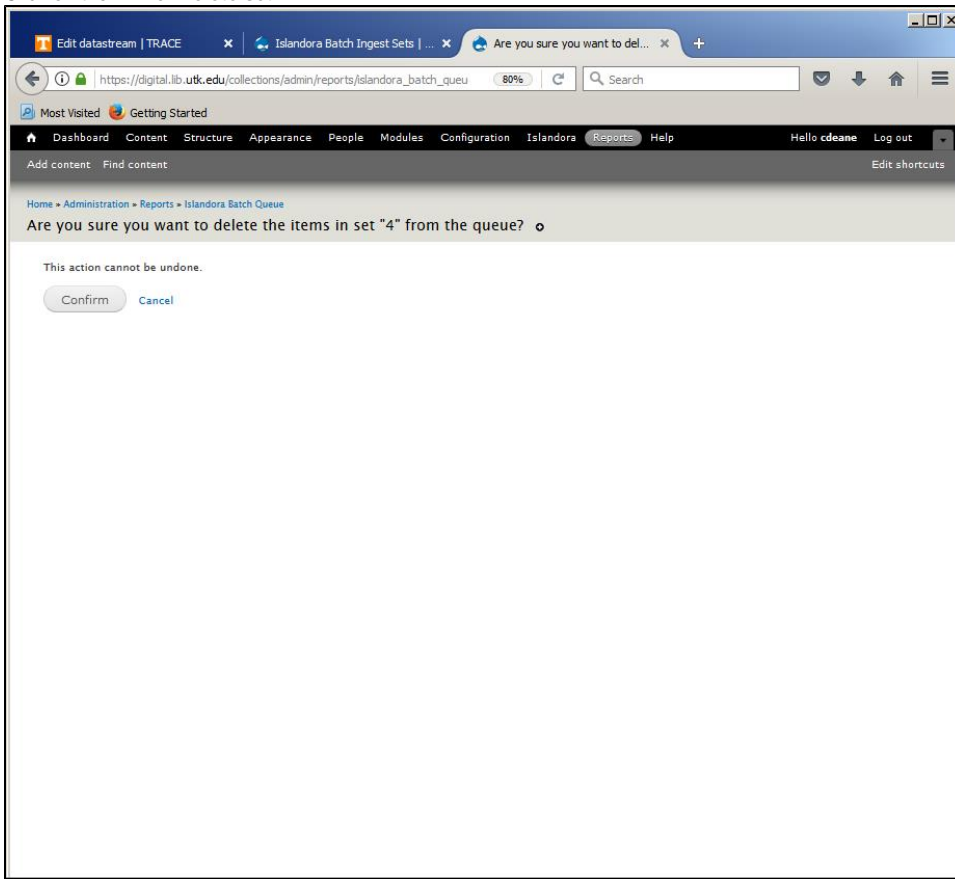
Displaying 1 - 13 of 13

SET ID	SET CREATOR	CREATED DATE	SET SIZE
37	mbagget1	Wednesday, March 1, 2017 - 14:45	4
36	mbagget1	Wednesday, March 1, 2017 - 14:36	4
35	mbagget1	Wednesday, March 1, 2017 - 14:35	0
13	cdeane	Tuesday, February 21, 2017 - 14:55	1
12	cdeane	Tuesday, February 21, 2017 - 14:53	1
11	cdeane	Tuesday, February 21, 2017 - 14:51	1
10	cdeane	Tuesday, February 21, 2017 - 14:49	1
9	islandora	Tuesday, February 21, 2017 - 13:26	1
8	islandora	Tuesday, February 21, 2017 - 13:25	1
7	islandora	Tuesday, February 21, 2017 - 13:23	1
6	islandora	Tuesday, February 21, 2017 - 13:21	1
5	islandora	Tuesday, February 21, 2017 - 13:19	1
4	islandora	Tuesday, February 21, 2017 - 13:17	1

- Although there is an option for **Delete set**, the most prudent action is to click on **View items in set** to verify which **Batch Set** you have chosen.
- Here I have chosen to **View items for Set 4** which is the last Set in the **Islandora Batch Set** report shown above.



- This brings up the **Set 4 Batch Queue** with a link for **Delete set**.
- Note that the **SET ID 4** is in the only row above.
- Click on the link for **Delete set**.



- Islandora gives you one more chance to change your mind.
- Upon clicking the **Confirm** Button, you return to the **Islandora Batch Ingest Sets** page. Note the message about the **Deleted item**.
- **Set ID 4** is no longer on the report.
- Three of these sets belong to someone else, so I only have 9 to go.

Edit datastream | TRACE

Islandora Batch Ingest Sets | ...

Islandora Batch Ingest Sets | ...

https://digital.lib.utk.edu/collections/admin/reports/islandora_batch_sets

80%

Search

Most Visited

Getting Started

Dashboard

Content

Structure

Appearance

People

Modules

Configuration

Islandora

Reports

Help

Hello cdeane

Log out

Add content

Find content

Edit shortcuts

Home » Administration » Reports

Islandora Batch Ingest Sets

Deleted 1 item from queue.

Displaying 1 - 12 of 12

SET ID	SET CREATOR	CREATED DATE	SET SIZE
37	mbagget1	Wednesday, March 1, 2017 - 14:45	4
36	mbagget1	Wednesday, March 1, 2017 - 14:36	4
35	mbagget1	Wednesday, March 1, 2017 - 14:35	0
13	cdeane	Tuesday, February 21, 2017 - 14:55	1
12	cdeane	Tuesday, February 21, 2017 - 14:53	1
11	cdeane	Tuesday, February 21, 2017 - 14:51	1
10	cdeane	Tuesday, February 21, 2017 - 14:49	1
9	islandora	Tuesday, February 21, 2017 - 13:26	1
8	islandora	Tuesday, February 21, 2017 - 13:25	1
7	islandora	Tuesday, February 21, 2017 - 13:23	1
6	islandora	Tuesday, February 21, 2017 - 13:21	1
5	islandora	Tuesday, February 21, 2017 - 13:19	1

View items

View items

View items

View items

View items

View items

View items

View items

View items

View items

View items

View items in set

- Returning to the **Islandora Batch Ingest Queue** from **Reports**, the report shows the **SET ID** numbers in ascending order.
- **SET ID 4** no longer appears in the **Islandora Batch Ingest Queue**.

Islandora Batch Ingest Queue

Displaying 1 - 17 of 17

Item State:

ID	STATE	MESSAGE	SET ID
cdf:10402	Ingested	Ingested cdf:10402.	5
cdf:10403	Ingested	Ingested cdf:10403.	6
cdf:10404	Ingested	Ingested cdf:10404.	7
cdf:10405	Ingested	Ingested cdf:10405.	8
cdf:10406	Ingested	Ingested cdf:10406.	9
cdf:10407	Ingested	Ingested cdf:10407.	10
cdf:10408	Ingested	Ingested cdf:10408.	11
cdf:10409	Ingested	Ingested cdf:10409.	12
cdf:10410	Ingested	Ingested cdf:10410.	13
scopes:2731	Ingested	Ingested scopes:2731.	36
scopes:2732	Ingested	Ingested scopes:2732.	36
scopes:2733	Ingested	Ingested scopes:2733.	36
scopes:2734	Ingested	Ingested scopes:2734.	36
scopes:2735	Ingested	Ingested scopes:2735.	37
scopes:2736	Ingested	Ingested scopes:2736.	37
scopes:2737	Ingested	Ingested scopes:2737.	37
scopes:2738	Ingested	Ingested scopes:2738.	37

Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item
Set Item

- This is the entire process for deleting a single Batch Set.