Archivematica DIP upload to CONTENTdm for Digitized Materials

Introduction

Most of the materials created at the Digitization Centre will be added to our content management system CONTENTdm for public access. To ensure we are meeting our commitment to preservation, all items should be processed through Archivematica before uploading to CDM. If you follow the general guidelines for structuring directories and file naming conventions, this should be a fairly smooth workflow. After processing, Archivematica will provide a metadata file that will allow batch upload to the CDM Project Client. This document describes the practical steps for this workflow, including preparing the transfer packages, processing through Archivematica, and uploading to the CDM Project Client.

> For a full introduction to Archivematica, please see Archivematica_Introduction.docx, or the full documentation available at the Archivematica Wiki: https://www.archivematica.org/wiki/Documentation
>
> To explore UBC Digital Collections, visit: http://digitalcollections.library.ubc.ca
>
> For more information about CONTENTdm, try the online help site: http://www.contentdm.org/help6/projectclient/index.asp

Check CDM Collection

Before starting work in Archivematica, the CDM collection and transfer directory must be set up correctly. First, ensure that the collection has been created on CDM. Only an administrator can create a new collection (i.e. your supervisor or Leslie). Second, check the metadata fields and Dublin Core (DC) mappings on the CDM Administration site:

- Log in to the CDM admin
- Click on the Collections tab and use the drop down menu to select the collection.
- Click on the Fields tab. Refer to the most up-to-date Metadata Manual to ensure the template is set up correctly. Check the field names, order, and DC

map—each should exactly match the Metadata Manual. Occasionally a collection will have additional fields unique to the project.

CONTENTdm Administration

Current collection: Uno Langmann Family Collection of BC Photographs ▾ change

Metadata fields

View and configure collection and administrative fields.

Collection field properties

View, add, edit and delete fields. Enable full text searching and controlled vocabulary. After you have added, changed, or deleted fields, index the collection to update changes.

	Field name	DC map	Data type	Large	Search	Hide	Required	Vocab		add field
1	Title	Title	Text	No	Yes	No	Yes	No	move to ▾	edit | delete
2	Alternative Title	Title-Alternative	Text	No	Yes	No	No	No	move to ▾	edit | delete
3	Creator	Creator	Text	No	Yes	No	No	No	move to ▾	edit | delete
4	Contributors	Contributors	Text	No	Yes	No	No	Yes	move to ▾	edit | delete
5	Publisher - Original	Publisher	Text	No	Yes	No	No	No	move to ▾	edit | delete
6	Date Created	Date-Created	Text	No	No	No	No	No	move to ▾	edit | delete
7	Date Issued	Date-Issued	Text	No	No	No	No	No	move to ▾	edit | delete
8	Sort Date	Date	Text	No	Yes	No	No	No	move to ▾	edit | delete
9	Description	Description	Text	Yes	Yes	No	No	No	move to ▾	edit | delete
10	Extent	Format-Extent	Text	No	No	No	No	No	move to ▾	edit | delete
11	Subject	Subject	Text	No	Yes	No	No	No	move to ▾	edit | delete
12	Subject - Geographic	Coverage-Spatial	Text	No	Yes	No	No	Yes	move to ▾	edit | delete
13	Personal Names	Subject	Text	No	Yes	No	No	Yes	move to ▾	edit | delete
14	Genre	Type	Text	No	No	No	No	Yes	move to ▾	edit | delete
15	Type	Type	Text	No	Yes	No	No	No	move to ▾	edit | delete
16	Format	Format	Text	No	Yes	No	No	No	move to ▾	edit | delete
17	Language	Language	Text	No	Yes	No	No	No	move to ▾	edit | delete
18	Notes	Description	Text	Yes	Yes	No	No	No	move to ▾	edit | delete
19	UBC Call Number	Identifier	Text	No	Yes	No	No	No	move to ▾	edit | delete
20	Access Identifier	Identifier	Text	No	Yes	No	No	No	move to ▾	edit | delete
21	Digital Identifier	Identifier	Text	No	Yes	No	No	No	move to ▾	edit | delete
22	Is Part Of	Relation-Is Part Of	Text	No	No	No	No	No	move to ▾	edit | delete
23	Source	Source	Text	No	No	No	No	No	move to ▾	edit | delete
24	Date Available	Date-Available	Text	No	No	No	No	No	move to ▾	edit | delete
25	Publisher - Digital	Publisher	Text	No	No	No	No	No	move to ▾	edit | delete
26	Rights	Rights	Text	No	No	No	No	No	move to ▾	edit | delete
27	License	Rights-License	Text	No	Yes	No	No	No	move to ▾	edit | delete
28	Transcript	None	Text	No	Yes	No	No	No	move to ▾	edit | delete
29	Catalogue Record	Relation-Is Referenced By	Text	No	No	No	No	No	move to ▾	edit | delete
30	Project Website	Relation-Is Referenced By	Text	No	No	No	No	No	move to ▾	edit | delete
31	AIP UUID	Identifier	Text	No	No	No	No	No	move to ▾	edit | delete
32	File UUID	Identifier	Text	No	No	No	No	No	move to ▾	edit | delete
	Field name	DC map	Data type	Large	Search	Hide	Required	Vocab		add field

Prepare the Transfer Packages

Next, the transfer directory needs to be correctly structured for upload to Archivematica. This step is a significant amount of work and special attention must be paid to accuracy of the metadata and contents of the packages. The transfer packages will be set up on the network storage drives by reorganizing the working directory used for the project. If the directories of digitized files have been organized following our general guidelines, the process will be fairly simple. Some workflow automations may be available to help create the necessary structure from existing directories (using the Automator app).
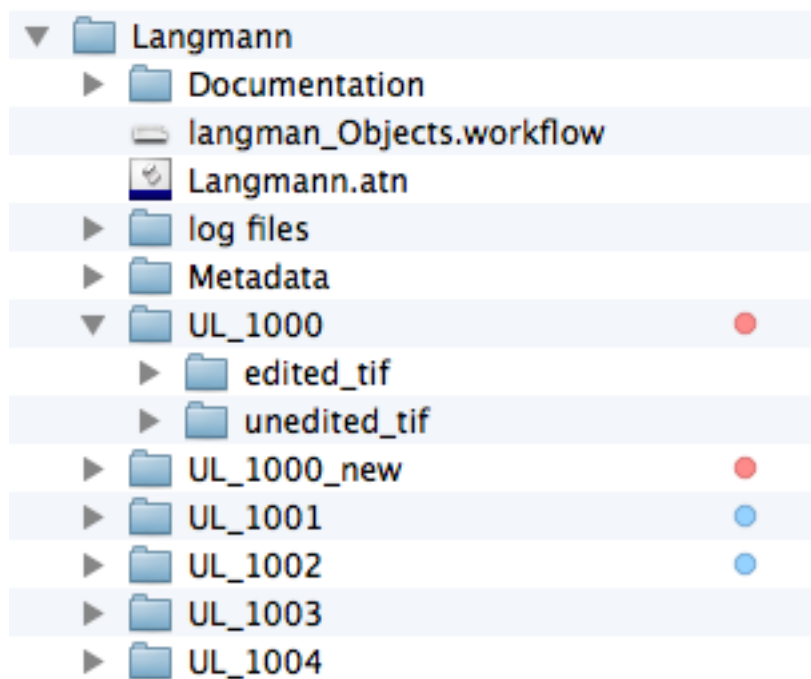
Figure 1: example working directory.

First, we must decide on a convenient package size to divide up the collection. Most often the transfer packages will simply follow the existing organization created during the scanning process.

> For example, in the Langmann working directory pictured above, individual albums of the collection are represented as a single directory named by the identifier, such as UL_1000. In this case, each album is a compound object, but every page has unique metadata. Because bulk uploading compound objects with page-level metadata is not supported by CDM, it was decided to create a separate transfer directory for each album. If there was not page-level metadata, it would be more efficient to group several albums together into a single transfer package.

Factors to consider for transfer package size:

- **Total file size:** our pipelines should comfortably process up to 100 GB packages (we have tested it to about 50 GB). It is estimated that a package requires processing space at least three times the package size. The total number of files has a bigger impact on processing speed than the file size (for example, 1000 small text files would take longer than 1 large video file at every step except the initial transfer). Contact IT to get information about the current resources allocated to each pipelines.
- **Intellectual structure:** it will often make sense to divide the files following the existing organization, for example, by the box number.
- **Simple vs. compound objects:** Individual items and compound objects will need to be contained in separate packages, because the CDM project client can only bulk import one at a time. If you would prefer to create a transfer package with a mix of object types, the resulting metadata txt file could be manually split into two before import to CDM.
- Occasionally, there will be other considerations, such as the need to add page level metadata to compound objects, that will constrain the contents of a package.

Now, we prepare the individual transfer packages. Ensure the transfer directories are named using the standard naming convention for the collection. If necessary, directories can be bulk renamed using Renamer on the Windows side (when scanning, uncheck files to rename only directories).
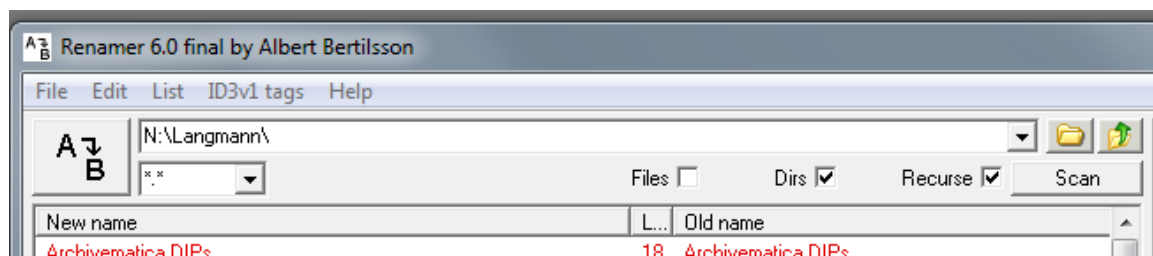

Figure 2: uncheck Files and check Dirs to rename directories.

For example, an item from RBSC is usually named following the pattern rbsc_arc_####_box#_folder#. Thus, a package containing only one folder would be named *rbsc_arc_1300_01_01*, a series of folders *rbsc_arc_1300_01-03*, or a full box *rbsc_arc_1300_01*. When transferring the entire collection in one package, it may also be desirable to have the collection's standard name included, for example, *Peter_Anderson_fonds_rbsc_arc_1007*.

Each transfer package will be a directory with two subdirectories named exactly *metadata* and *objects* (all lower case).

The *objects* directory contains the scanned materials. Usually, the objects will already be contained in unedited and edited directories. These should simply be

moved in to the *objects* directory. Subdirectories inside *objects* can be named anything, but should not contain spaces or non-standard characters. Ensure that each compound object is contained in its own subdirectory.

The *metadata* directory contains information about the package. Add a subdirectory named exactly *submissionDocumentation*. Move any relevant documentation to this directory, such as a copy of the catalogue record (saved as PDF/A), finding aids, project documentation, or notes that would be important for understanding the contents in the future. Add a copy of the XLSX spreadsheet containing the metadata for the items in the package to this directory as well. The top level of the metadata directory will contain a single file, *metadata.csv*, described in the next step.
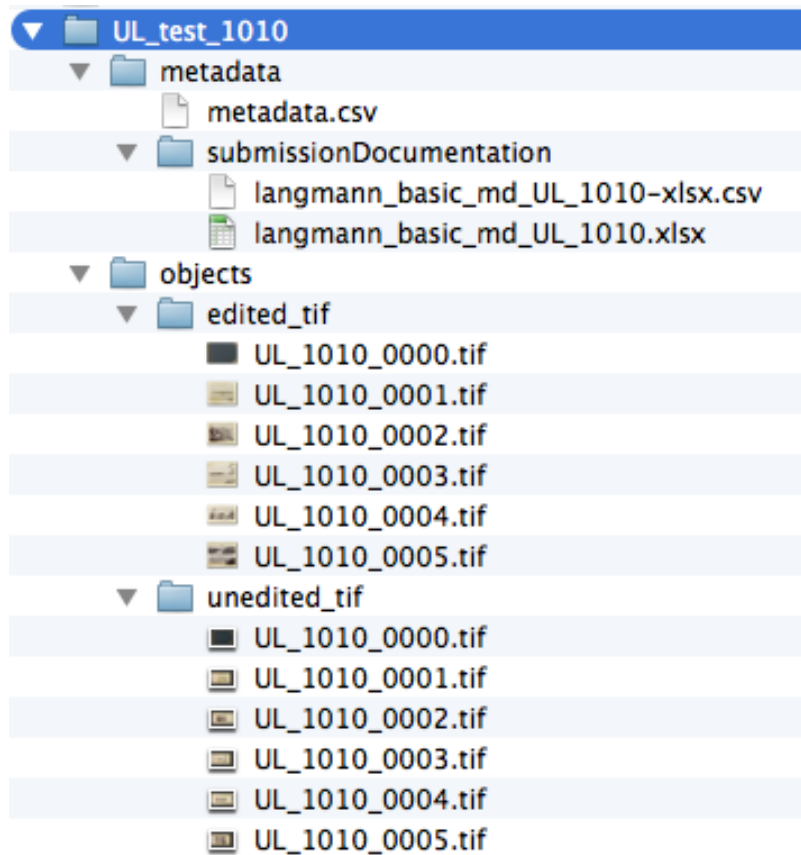
Figure 3: example transfer package of simple objects.
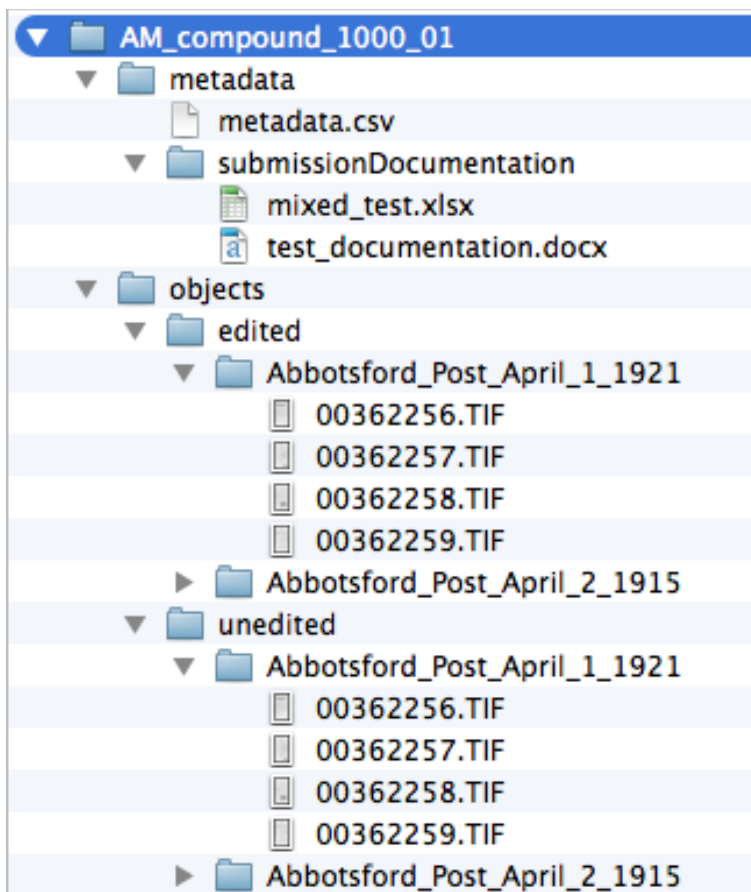
Figure 4: example transfer package of compound objects.
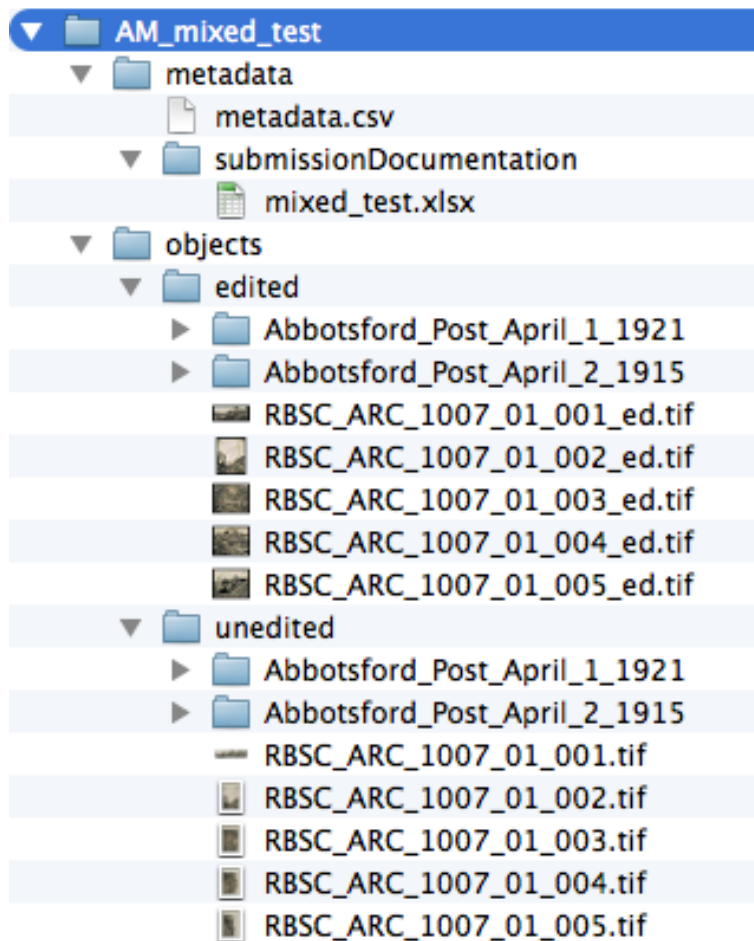
Figure 5: example transfer package with mix of simple and compound objects.

Considerations when setting up the transfer directories:

- **Proceeding Zeros:** directories will not sort correctly in Archivematica without proceeding zeros in the digital file names (example, rbsc_arc_1300_01 versus rbsc_arc_1300_1). It is best practice to always use proceeding zeros. The Renamer application (on the Windows side) can be used to add zeros or renumber both digital files and directories. Please ask a supervisor for training and documentation on using Renamer.
- **Deep Directory Structure:** The file explorer used to start transfers in Archivematica shows only 50 files for a directory at a time. To bring up more it is necessary to click "see next 50". Thus, it is very slow to navigate to the transfer package you need if it is in a directory of more than 50 items. It is best to create deeper directory structures to divide directories containing more than 50 packages.
- **Directory Name:** Each time you browse the file explorer to start a transfer, Archivematica collapses the tree—so you will have to navigate through the same directory structure over and over. To make the explorer more efficient, temporarily rename your collection's main directory "1_…" to put it at the top of the list.

- **Quality Assurance:** Setting up the transfer packages is an opportunity to quality control the materials. For example, does your metadata file have exactly the same number of items as you have edited TIFFs or compound object directories? Familiarize yourself with the collection so that you have a sense of what *SHOULD* be there during the transfer.
- **Keep Notes!** When starting the transfer process, you should create a Word or text file to keep notes. Save the file in the collection's transfer directory. Include all the information necessary for other workers to understand what you are doing and where you are in the process. Be sure to clearly record all errors you encounter so that they can be figured out and fixed! Because of the way the Archivematica Dashboard is set up, it is difficult to backtrack to figure out what was done in previous sessions—so please ensure you are documenting problems!
- **Other Options:** Archivematica will recognize a few other special types of directories. However, we do not currently use these directories in the workflow. For example, an *access* directory can contain manually created DIP access copies of the objects. A *service* directory can be created to hold edited versions of the masters that are used to create display images. Items in the *access* or *service* directory must exactly match a file name in the objects directory, although the extension can be different (for example, access rbsc_arc_1007_01_01.jpg and object rbsc_arc_1007_01_01.tif). This means if you create a single PDF as an access copy for ten TIFF images, you must place a copy of the PDF in both the access and objects directory.

Create the metadata CSV

The most important step of the DIP to CDM workflow is creating a metadata file that Archivematica can correctly parse. This ensures that the information necessary to understand the scanned images will be preserved and linked to the digital objects. Each package requires a comma-separated values file (CSV, a plain text format to store tabular information) containing the metadata for included items.

To create the *metadata.csv* from scratch, open Microsoft Excel and create a new spreadsheet. The basic structure is a first column named "Filename" for individual items or "Parts" for compound objects, followed by the columns of the standard metadata template, followed by columns for DC mapping. These elements are described in more detail below:

- **Filename/Parts:** this column will give the relative file path referring to the digital file (simple objects) or directory (compound object) included in the transfer package for each metadata record. Since the files will be put in a directory *objects* in the transfer package, the values in this field will always start with "objects/". For individual items (such as photographs), the digital file name as it appears in the objects directory will be added (e.g. "objects/rosetti_bc1903-32-39.tif"). For compound objects (such as books),

the sub-directory that contains all the files for the object will be added (e.g. "objects/020_British_Columbia_Can").

	A	B
1	Filename	Title
2	objects/edited/RBSC_ARC_1007_01_001_ed.tif	Head quarter can
3	objects/edited/RBSC_ARC_1007_01_002_ed.tif	Cater Pillor loggir
4	objects/edited/RBSC_ARC_1007_01_003_ed.tif	Load going to dur
5	objects/edited/RBSC_ARC_1007_01_004_ed.tif	Caterpillar loggin
6	objects/edited/RBSC_ARC_1007_01_005_ed.tif	Caterpillar load o
7		

Figure 6: Filename column.

- **Metadata template fields:** The next columns will match the metadata template used for the collection, as shown on CDM admin Fields tab. Fill the fields with the metadata for each simple or compound object. This metadata will be preserved with the AIP and also used to upload the objects to CDM.
- **DC mapping:** To help ensure future understandability, it is best practice to map metadata to a standard schema. Since our template is customized for use at UBC Digital Collections, the field names do not necessarily correspond directly to the Dublin Core they are based on. Thus, we copy each relevant field into a DC mapped column:
  - o Check the DC mappings for each field in the Metadata Manual and in the CDM admin Fields tab.
  - o For every field that has a DC map, create a column named based on the DC XML naming convention (dropping the < > and replacing the colon with a period). For example, CDM "Title" is mapped to DC "Title", thus add a new column named "dc.title". More than one CDM field may be mapped to the same DC field—create separate DC fields for each (i.e. DC field names can be repeated).
  - o Each DC map field will be filled with a copy of the metadata from the related field. Yes, you will repeat everything.

	Z	AA	AB	AC	AD	
Re	dc.title	dcterms.alternative	dc.creator	dcterms.created	dc.date	d
lv	Head quarte	Headquarter's camp, r	W.J. Moore F	[between 1930 and	1939	
lv	Cater Pillor l	Caterpillar logging	W.J. Moore F	[between 1930 and	1939	
lv	Load going to dump		W.J. Moore F	[between 1930 and	1939	
lv	Caterpillar logging at Harrison Lake		W.J. Moore F	[between 1930 and	1939	
lv	Caterpillar lo	Caterpillar load of full	W.J. Moore F	[between 1930 and	1939	

Figure 7: DC map columns.

- More information about the DC schema can be found at the namespace http://purl.org/dc/terms, or the DC in XML guidelines http://dublincore.org/documents/dc-xml-guidelines

Once the metadata is complete, save the file as an XLSX following the naming convention for the collection (this makes it easier to modify if corrections need to be made later on). Next, the file must saved to CSV for Archivematica to parse. Click Save As, choose "Windows Comma separated .CSV", and name the file *metadata*. Please note that Excel also offers Mac CSV and MS_DOS CSV which may NOT compatible with the Archivematica. Excel will give you a warning that features will be lost by saving in the new format—click though these warnings to save it. The file must be named exactly *metadata.csv*.
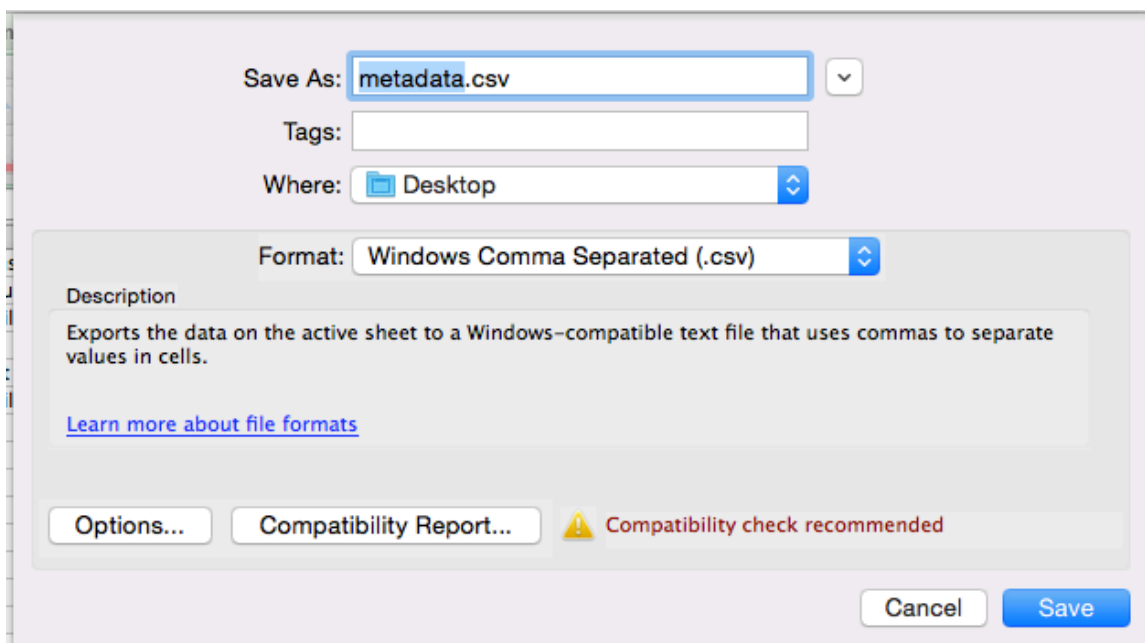
Figure 8: Save As Windows CSV.

Place *the metadata.csv* into the *metadata* subdirectory in the transfer package.
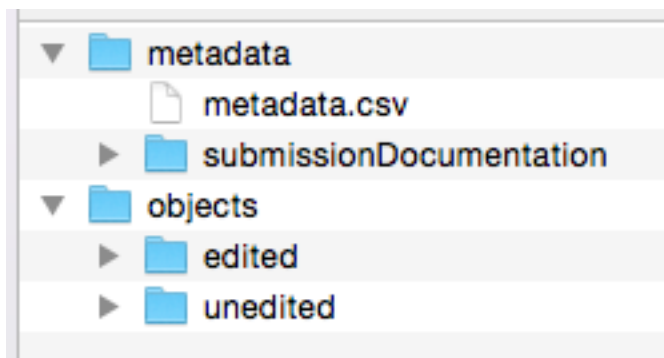
Figure 9: Transfer package structure.

It is possible to partially automate this process using a template or an OpenRefine routine. For example, if your collection uses exactly the standard DI metadata template, and you have the collection metadata as an Excel spreadsheet:

- Copy all columns and paste into the following columns.
- Replace the field names of the copied metadata with the DC map names by copying them from the file archivematica_metadata_template.xlsx.
- Delete any blank columns in the DC map and columns that are not mapped (transcript).
- Create the Filename/Parts column using a formula.

If your collection will involve many packages it is best to use OpenRefine, where a set of JSON commands can be used to automate the transformation.

Archivematica Setup: Processing Configuration

Once the transfer packages are prepared, you will start processing them using Archivematica. Log into the Archivematica Dashboard for the pipeline you are working on by navigating to the URL on Chrome or Firefox (do not use Safari or IE). The Dashboard behaves slightly different on Chrome and Firefox, so try both and decide which works best for you.

Always click on the Administration tab and check the Processing Configuration settings *before* starting work.

Processing Configuration options are used to automate the decision points during Archivematica processing when the same actions will be repeated for every package. If none of the options are checked, the Dashboard will ask for input about each step on the Transfer and Ingest tabs. Each decision point can be automated by checking the option in Processing Configuration and selecting the desired input from the drop down menu. If you make any changes to the settings be sure to click "Save".

Please note that anyone working on the same pipeline will be using the same processing resources and settings. If you change the settings it will be in effect for everyone on the pipeline.

Ensure that the automated settings match the needs of the current project, allowing some stops for QC if necessary, but also not wasting time with decision points that will always be the same. If things are running smoothly, all the steps can be automated.

For example, look at the screenshot below and the table explaining the options:

Figure 10: Processing configuration.

Processing Configuration Option	Typical setting	Notes
Send transfer to quarantine	No	Since we are creating the files in house, quarantine is unnecessary.

Approve normalization	Yes	This decision point allows you to review the results of the normalization process. If things are running smoothly this step can be automated (Yes). However, it is a good stop point to leave if you are trouble shooting (unchecked).
Store AIP	Yes	You may want to turn off this option if you are testing.
Transcribe files (OCR)	No	This option can create a text transcript of image files. We do not use it.
Generate transfer structure report	No	This option creates a record of the contents and organization of the transferred directories. It is relevant to born-digital archival collections to preserve the original order before preservation steps take place.
Remove from quarantine after	(unchecked)	Since we do not use quarantine this option is unnecessary.
Create SIP(s)	Create single SIP and continue processing	After processing completes on the Transfer tab, the package can be send directly to Ingest or to a backlog. We do not use the backlog function.
Extract packages	No	If your package contains zip files, Archivematica can extract the contents for processing. We do not have zip packages, so the decision is irrelevant.
Normalize	Normalize for Access	Normalization is the process of creating preservation and access derivatives of the original files. The workflow described in this document uses "Normalize for access" because we want Archivematica to create a DIP with metadata for CDM.

Reminder: add metadata if desired	Continue	Package metadata can be added in the Dashboard. We do not typically use this option.
Examine contents	Skip examine contents	This option relates to Extract packages, allowing you to inspect the contents. We don't use it.
Select file format identification command (Transfer)	Fido version 1	Fido is an application that examines the signatures of the digital files to determine their type. This information is essential for digital preservation.
Select file format identification command (Ingest)	Use existing data	Format identification can be run a second time during ingest, which is only necessary if zipped contents were extracted.
Select file format identification command (Submission documentation & metadata)	Fido version 1	
Delete packages after extraction	Yes	Related to the Extract packages option.
Select compression algorithm	7z using bzip2	We use the default option.
Select compression level	5 – normal compression mode	We use the default option.
Store AIP location	Production aip store	Each pipeline has one AIP storage location.
Store DIP location	(unchecked)	DIPs can be stored in a separate server for systems not integrated with Archivematica. We do not use this option.

Archivematica Processing

After checking the Processing Configuration, you are ready to start transferring packages to the system. This section describes the step-by-step process:

1. Transfer

Navigate to the Transfer tab. Transfers are initiated using the bar at the top of the page. Fill in the form using these parameters:

- **Type:** Standard
- **Transfer name:** the identifier number of the package, most likely the same as the name of the transfer directory (e.g. rbsc_arc_1300_2_34). I suggest keeping a notepad with the collection identifier to cut & paste from, to minimize the possibility of typing errors as you repeatedly enter the ID to name the transfers.
- **Accession no.:** [empty]. We do not typically use accession number in this workflow. However, accession number should be used when sending SIPs a "backlog" or creating AICs. This makes it easier to search for a collection of packages during Ingest. In general it would be the collection's number (e.g. rbsc_arc_1300).
- **Drop down (transfer source):** Select the current location of the transfer packages from the drop down menu.

After filling in the parameters, click *Browse*. This opens a file explorer to the chosen storage drive. Navigate to the directory of the transfer package and click "add". Folders may take a long time to load, particularly on drives with a lot of files, as Archivematica is counting the number of items in every folder on the drive. Click the arrow next to the folder icon for the drive so that it is pointed down. Wait for 5-10 minutes (depending on the size of the drive) and then click the folder icon. Do the same for each level in the directory if files do not appear immediately.
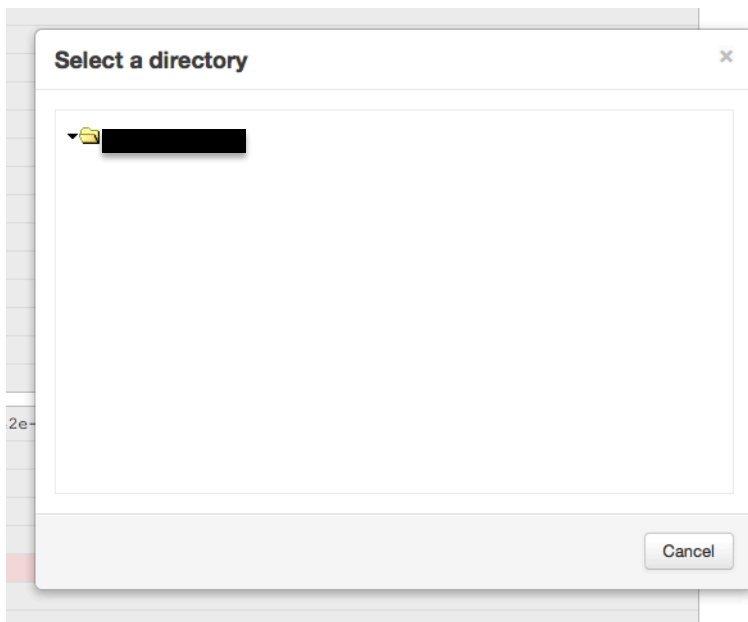
The file path and name of the directory you added will appear below the transfer name—double check to ensure you have added the correct file!

Figure 11: Transfer set up and ready to start.

If everything looks good, click Start Transfer. This begins uploading the digital files to the processing server. Depending on the size of the package, this can take a few seconds to several hours. The Dashboard will be unresponsive during the transfer. When using Chrome, you may receive a "page unresponsive" warning from the browser—click "wait" and wait…

Once the upload is complete, the package will appear on the Dashboard list. The package will be at a decision point asking for you to "approve" transfer. This step can not be automated. Ensure the package is named correctly. If all is good, choose yes from the drop down menu!

Figure 12: Approve transfer decision point.

As the packages move through processing, the micro-service steps will appear. In this basic workflow, all the Transfer steps will be automated by the settings in the Processing Configuration.

The active micro-service job is highlighted in gold. If the step is successful, the job will be highlighted in green. If there is an error, the highlight is red.

Each package on the list will have an icon next to its name on the left. A bell means it is stopped and waiting for a decision. Refresh arrows means it is currently processing. A green check means it is completed with no errors. A circle with blue square inside means the package was rejected at some point and processing was stopped permanently. Any errors or failed services should show up highlighted in red, or a red X icon.

Figure 13: Package being processed.

When the Transfer processes are done, a SIP will appear on the Ingest tab.

> Note: Artefactual suggests we work on one package at a time, waiting for the full processing to complete. More than one can be started, but care must be

taken to not exceed the limits of the processing space. Transfers take about three times as much space to process as the total package size

2. Ingest and Normalization

Ingest moves the package through the process of Normalization, creating DIPs for access and AIPs for archival storage. The package you sent through the Transfer process should now appear as SIPs on the Ingest list, and will automatically start processing. If you follow the configuration settings for this workflow given above, you will be normalizing for access and this stage will be automated except Upload DIP.

However, if it is necessary to trouble shoot issues with the collection (i.e. you are getting errors), it is best to turn off some decisions in the Processing Configuration. In particular, it may be helpful to manually complete Approve normalization. This stop is a good point to check the files in the package to ensure that there are no anomalies using the Normalization Report or Review:

Figure 14: Approve normalization stop.

- Normalization Report:
  - On the right side of the "Job: Approve normalization" entry, click the notepad icon near the action drop down menu. This will open the Normalization Report in a new window. Any problems will be highlighted in red and brought to the top of the list. Corrupted or zero byte files will probably cause an error that will be flagged. If there are more than ten digital files in the package, you will need to click "next/previous" to review them all.
  - After checking the report, close the window to go back to the Ingest page. If there is no error, approve. If there is an error reject the package—you will need to restart the transfer process (after fixing the issues in the transfer files).

File name	File format	Preservation normalization attempted	Preservation normalization failed	Already in preservation format	Access normalization attempted	Access normalization failed	Already in access format
rbsc_arc_1300_9_7_004_002.tif	TIFF	No	No	Yes	Yes	No	No
rbsc_arc_1300_9_7_002_002.tif	TIFF	No	No	Yes	Yes	No	No
rbsc_arc_1300_9_7_001_004.tif	TIFF	No	No	Yes	Yes	No	No
rbsc_arc_1300_9_7_002_001.tif	TIFF	No	No	Yes	Yes	No	No
rbsc_arc_1300_9_7_001_001.tif	TIFF	No	No	Yes	Yes	No	No
rbsc_arc_1300_9_7_003_003.tif	TIFF	No	No	Yes	Yes	No	No
rbsc_arc_1300_9_7_003_002.tif	TIFF	No	No	Yes	Yes	No	No
cts/a5fa2e0a-99fa-4dab-92ea-c056524fd66e-rbsc_arc_1300_9_7.pdf	Acrobat PDF 1.6	No	No	No	No	No	Yes
rbsc_arc_1300_9_7_001_003.tif	TIFF	No	No	Yes	Yes	No	No
rbsc_arc_1300_9_7.pdf	Acrobat PDF 1.6	No	No	No	Yes	No	Yes

Showing 1-10 of 14

Next

Figure 15: Normalization report.

- Review:
  - Click on "review" near "Job: Approve normalization" if you want to more closely inspect the contents of the package. However, Review does not highlight errors, it simply allows you to explore the approveNormalization directory of the pipeline. In most cases review is unnecessary.
- What to look for:
  - Check over the file names to ensure that everything in the package is correct. For example, check that the SIP for rbsc_arc_1300_2_36 has TIFs and PDFs with rbsc_arc_1300_2_36 file names. Occasionally incorrect files can be inserted when creating the transfer package directories.
  - Ensure that the correct number and types of files are present. If the correct number of files is not present, there is likely an error in the transfer directory.

3. Upload DIP

The Upload DIP micro-service allows you to enter parameters for sending DIPs to access systems. The decision points cannot be automated at this time. DIP upload to CDM has a series of stop points where you will have to make a selection from the drop down menus:

- Upload DIP: choose "upload DIP to CONTENTdm"

Once you have entered choices for the DIP micro-service decision points, processing in Archivematica will go on to completion. However, you will have to complete further steps outside of Archivematica to upload the objects to CDM.

NOTE: The microservice "Upload DIP" will process before the "Prepare AIP" and "Store AIP" microservices; however, it is best to wait for these microservices before proceeding to "Upload DIP" in case any errors occur in the AIP.

4. Complete Upload to CDM

To efficiently upload the batch of images to CDM, Archivematica provides a tab delimited text file named *simple.txt* or *compound.txt*. This file will contain the descriptive metadata provided in the *metadata.csv*, plus the UUIDs for the AIP and file in preservation storage. It is structured for batch uploading to the CDM Project Client.

First, you will need to download the TXT file from Archivematica which can be done by two methods:

- Option one: DIP Review on the Dashboard.

  - In the *Upload DIP* service, *Job: Move to the uploadedDIPs directory* click on "(review)". This opens a new window with a file explore for the directory "uploadedDIPs". Click on the directory to browse to uploadedDIPs and find the package you have processed.
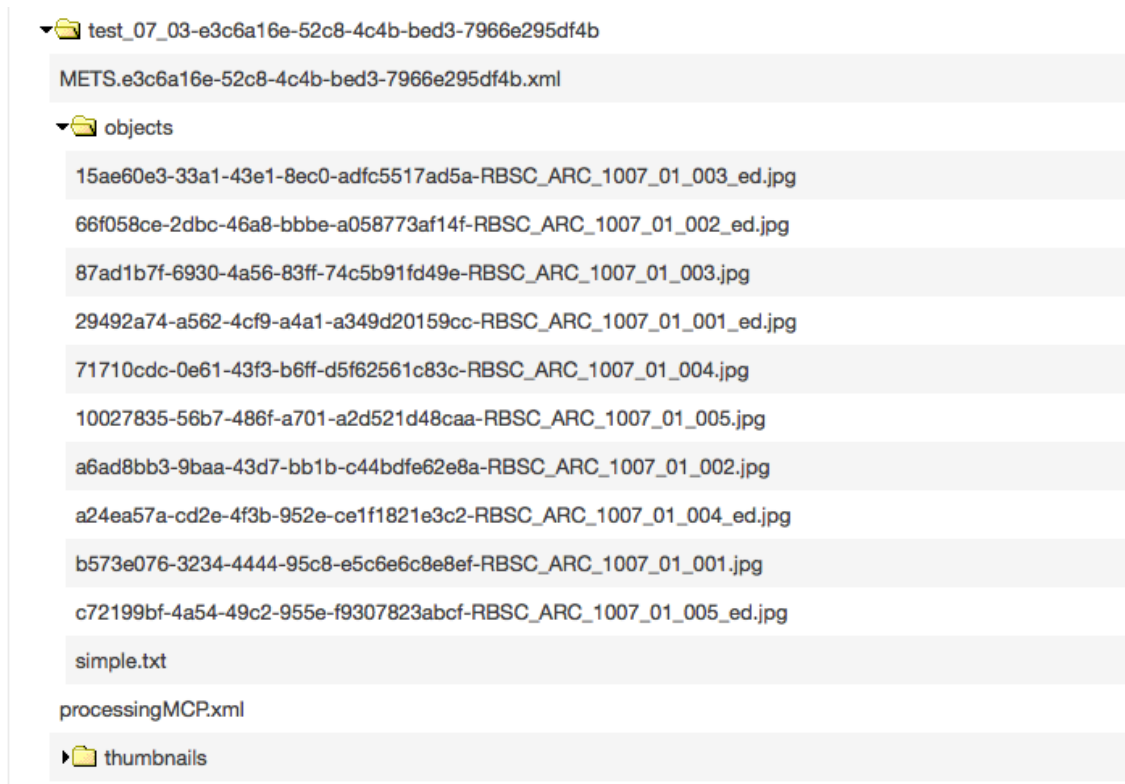
Figure 16: Navigate to the DIP using the Dashboard review.

o There will be a file "compound.txt" for compound objects or "simple.txt" for simple objects located in the "objects" folder of the DIP. Click on the name which will open the text in a new tab. In the new tab, right click and choose "save as". Change the name to something meaningful and add the extension .txt, for example "simple_anderson.txt", and click save.

Figure 17: Save the TXT from the browser.

o Finding the new DIP may be complicated if many packages have been processed as the file explorer is very rudimentary. If you have only a couple DIPs to work on, this option may be simplest. However, it is cumbersome, so when working with more than one package, it is better to use option two.

• Option two: Cyberduck SFTP.
   o Start Cyberduck and connect to DIP storage

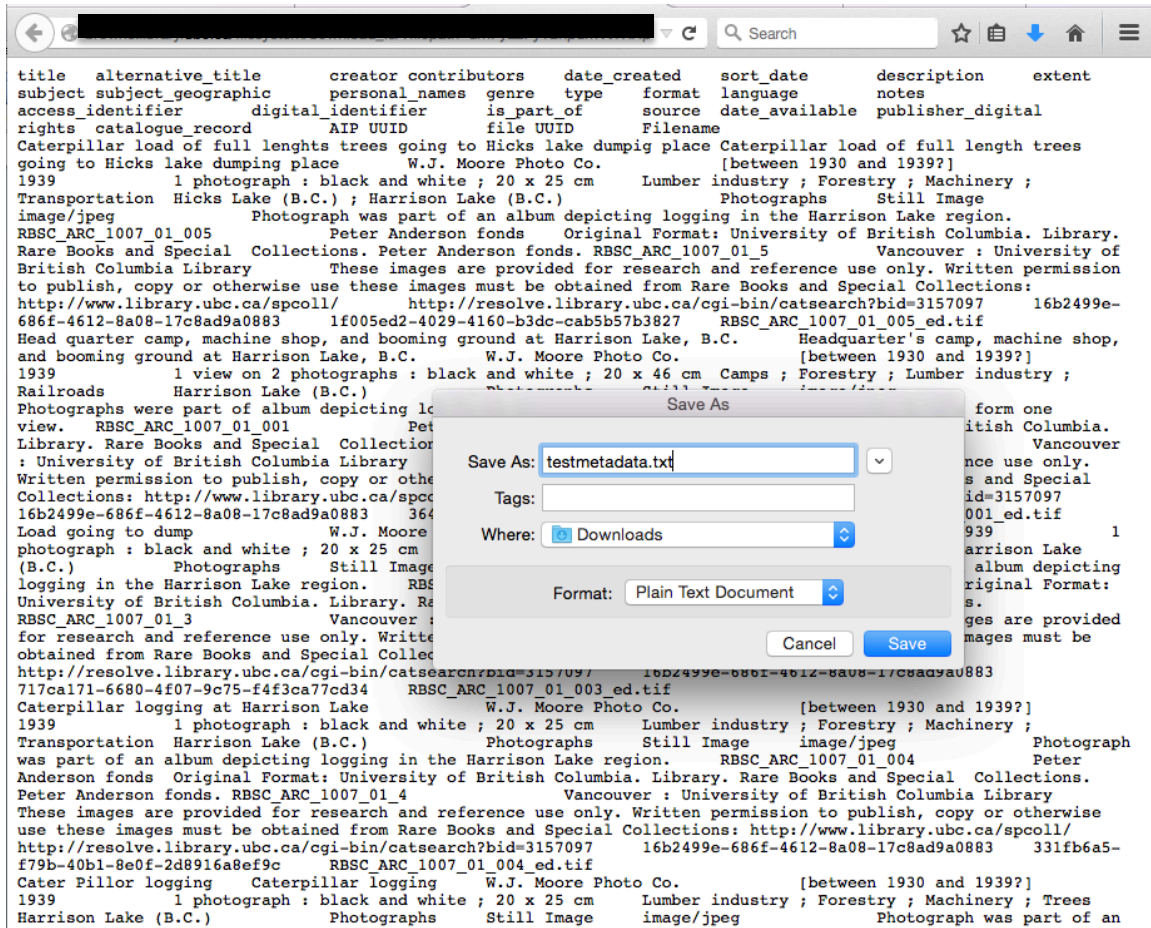o   Navigate to the DIP you have processed (the name will be the transfer package name plus a long string of characters that is the UUID). Right click on the name and download the DIP to your computer. The simple.txt/compound.txt file can be found in the "objects" folder.

o   Each time you finish processing another package, simply refresh the directory, and download the simple/compound.txt.
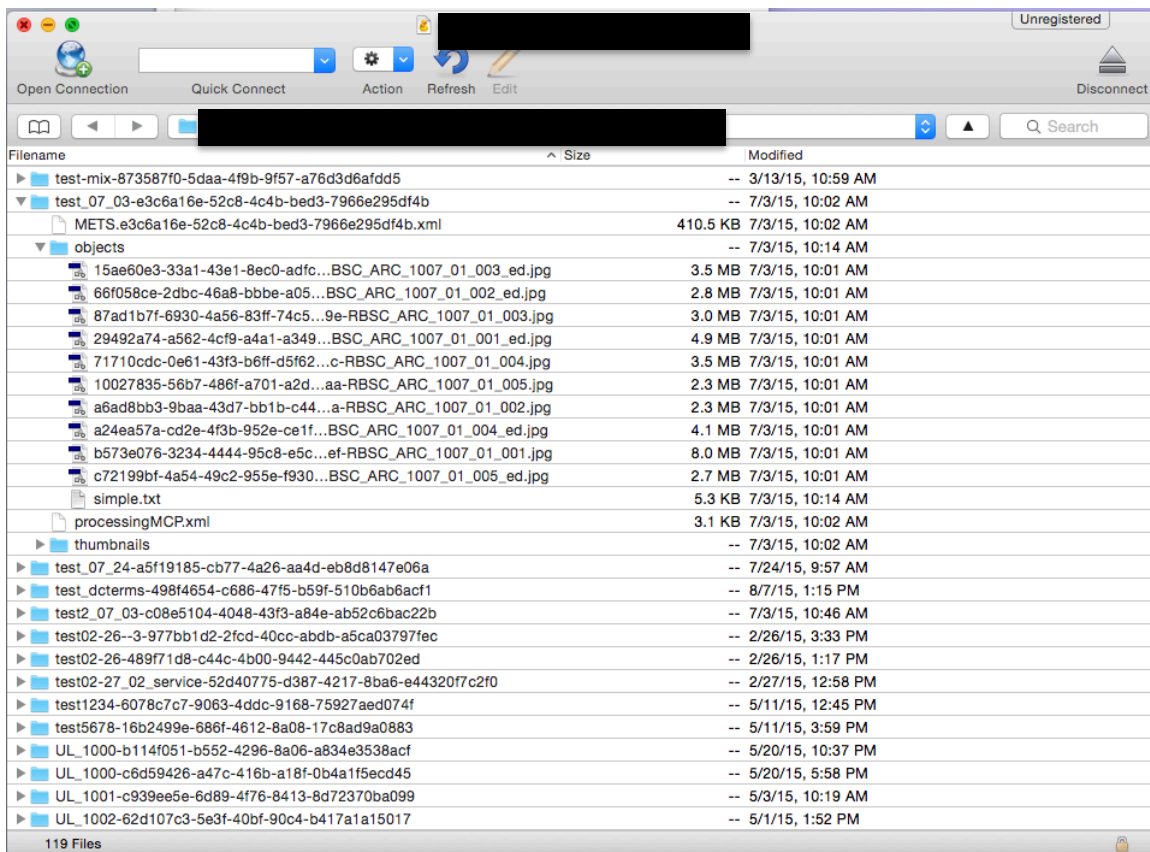

Figure 18: Cyberduck directory.

Once you have the TXT, go to the windows side, start the CDM Project Client, and open a project for the collection. You should always click "Edit Project Settings" to check the settings for images, PDF, and OCR. Ensure these options match the requirements for the collection and that there is enough pages remaining if OCR is being used (see CDM documentation for more information).

Depending on the material you are uploading, there are three possible methods: Simple objects, compound objects, or compound objects with page-level metadata.

Add the DIP items using "Add Compound Objects" (for compound objects) or "Add Multiple Items" (for simple objects) option, found on "common tasks" menu on the left side of the client.
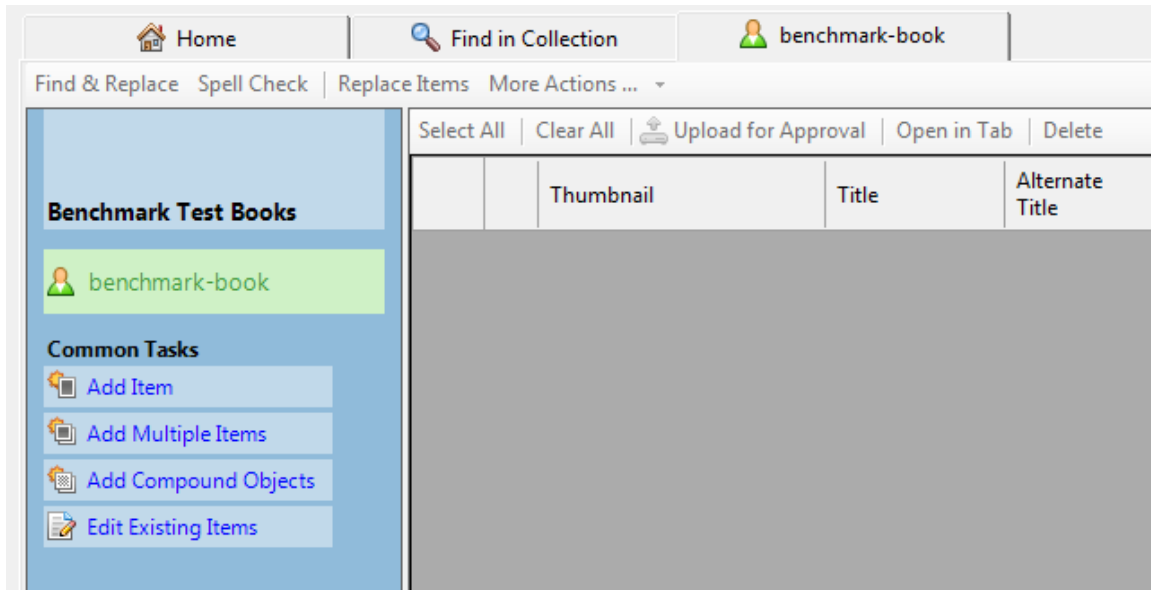
Figure 19: CDM common tasks.

- o **Simple objects: Add Multiple Items**
    - ▪ Clicking Add Multiple Items brings up the multiple items window. Choose "Import using a tab-delimited text file", and browse to the simple.txt file you downloaded from Archivematica/Cyberduck. Click next.
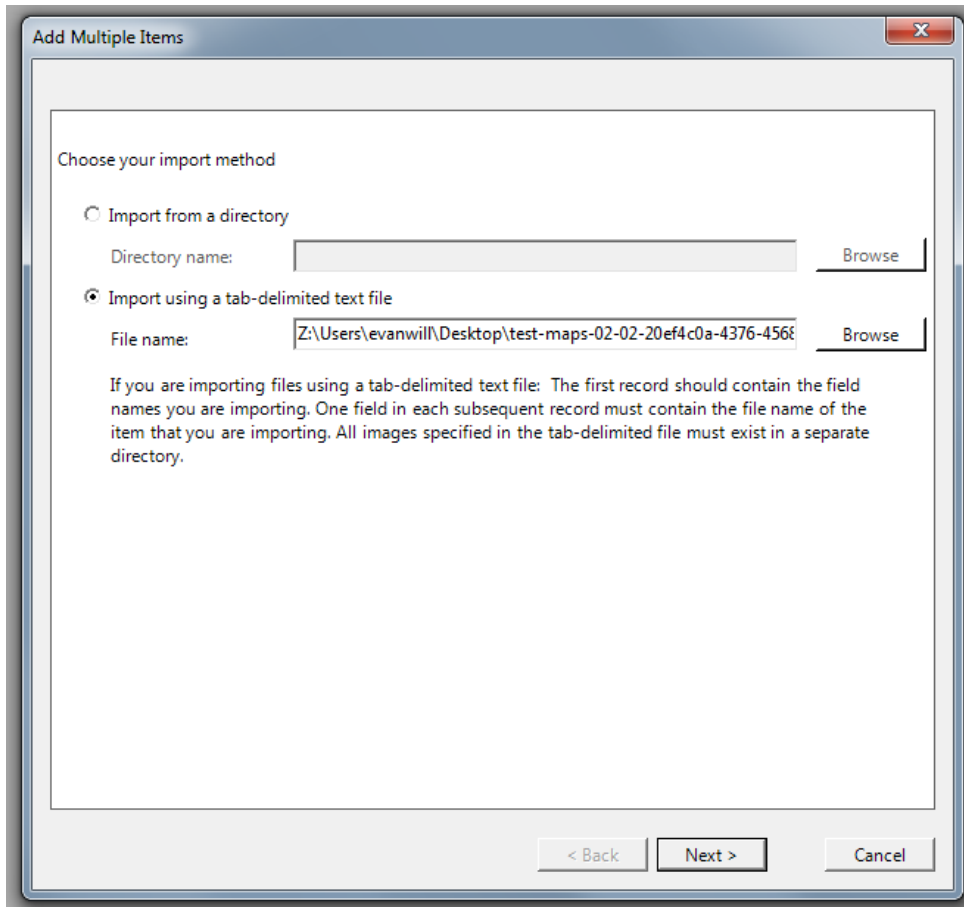
- On the next window, choose "Import files from a directory" and browse to the original transfer directory on our network storage. Click next.
- Choose Yes to generate display images, and click next.
- Check the field mappings to ensure they are correct (they should be fine). Filename should be mapped to Object File Name. Click next.
- Click Add items. CDM will process the items and generate a report. If you get any error messages, figure out what went wrong (usually an issue with the file names). Your items are now added to the client.

- **Add Compound Objects:**
  - Clicking on Add Compound Objects brings up the compound objects window. On the "Add using" drop down menu, select "Object List" and then click on "Add". [if you send only one compound object in a package Archivematica generates a different type of TXT with all the file names listed—then object

list method does not work, and you will have to use the wizard instead, pointing to the TXT for the metatdata and the scans directory for the images.]

- This opens the Object List window. For compound object type, select Document. For the Object List file name, browse to the compound.txt you downloaded. For the Scans directory, browse to the original transfer directory. Then click next.
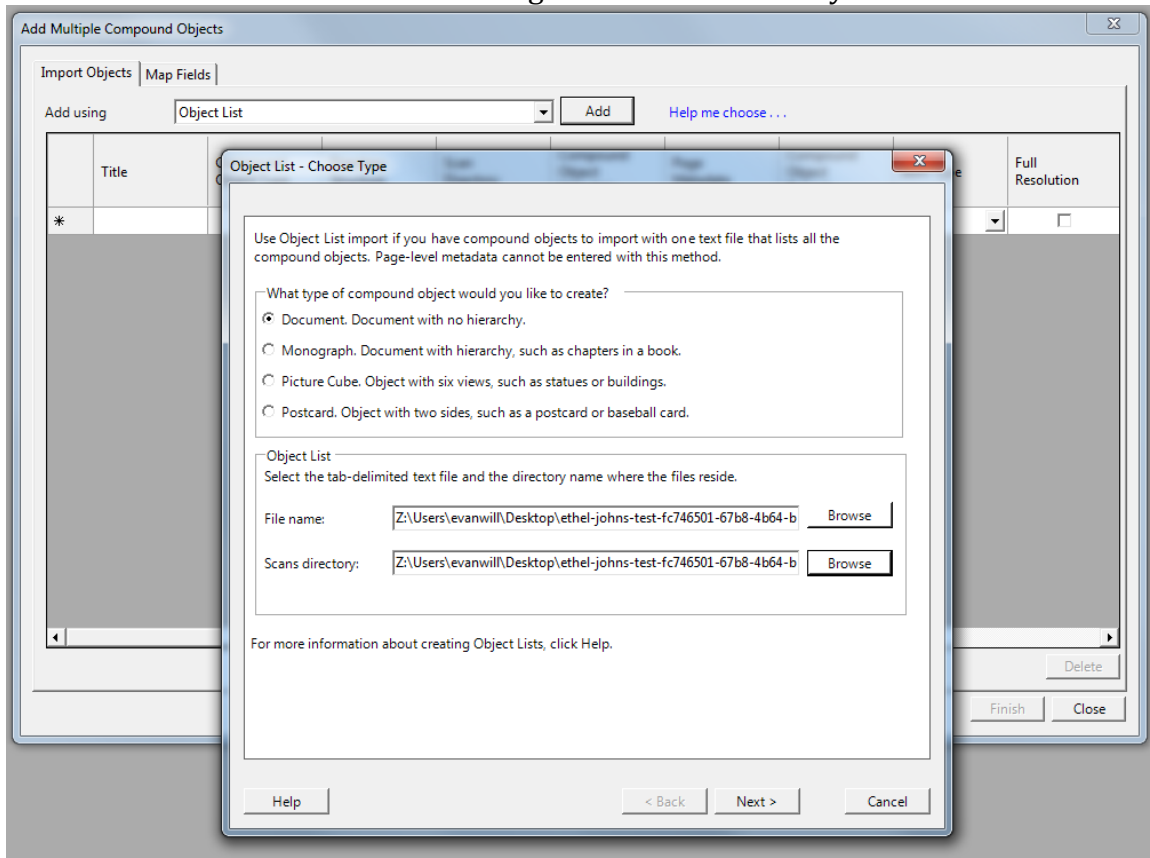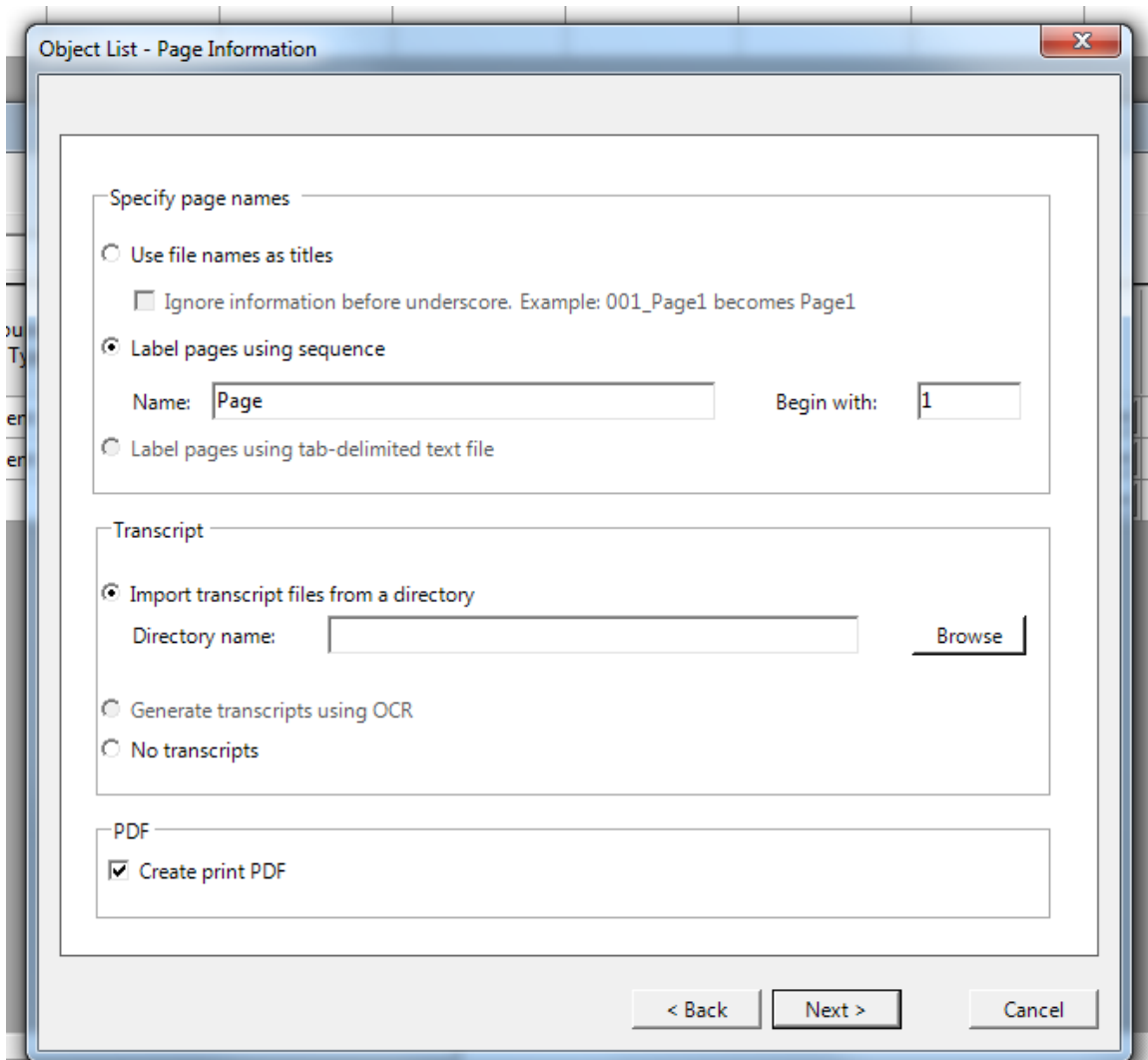
- The next window asks for the field that contains the names of the compound object directories. Use the drop down to select "directory name". Then click next.
- Choose Yes to generate display images, and click next
- On the next window, choose the options appropriate to the collection. Generally, for Specify page names, we will choose "Label pages using sequence" (name: page, begin with: 1). If the items need OCR select "Generate transcripts using OCR", otherwise select "No transcripts". If the object is a text, the option "Create print PDF" should usually be checked. Finally, click next.

Figure 22: Compound object options.

- On the next window, click finish. A summary report will appear. If there are any errors you will have to delete the object(s) and try again. After closing the report you will be returned to the Add Multiple Compound Objects window.
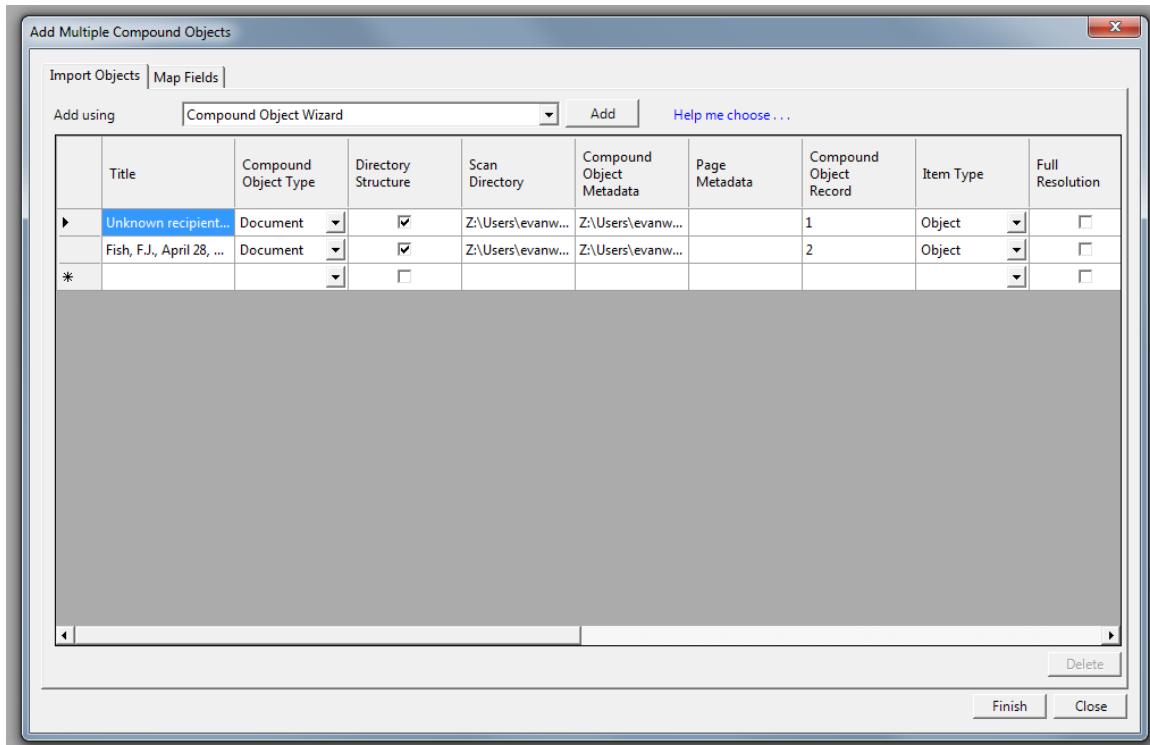
Figure 23: Compound object wizard.

- Click on the Map Fields tab and ensure that the metadata fields are correctly mapped to the CDM collection fields. Use the drop down options to correct the mapping (please note: occasionally this step does not work at first. If you cannot change the drop downs, close the Compound Objects window, then reopen it. Your objects will still be there, but you should be able to change the field mapping now.). The field "Directory name" should be set to None.

Figure 24: Mapping the fields.

- If everything is correct, click Finish. CDM will now complete the processing (creating display images, PDFs, and OCR) which may take a long time if you are adding many files. Your items are now added to the client.

- o **Add Compound Objects with item level metadata:**
    - You will need to edit the "simple.txt" file from the DIP prior to upload to CONTENTdm
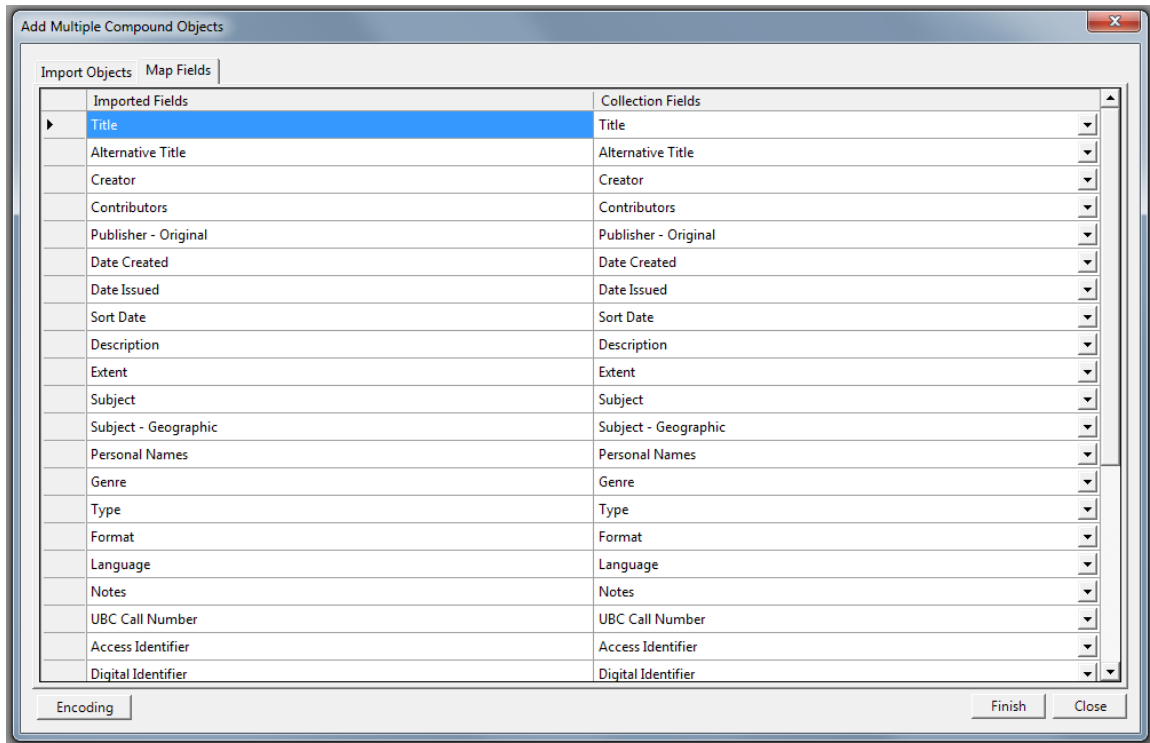   - Open the "simple.txt" file with Excel ON WINDOWS SIDE!
    - In order to gather the simple objects into a compound object you will need to insert a row before the first record and populate it with the metadata for the object.
    - The first record should contain the file name of the compound object that you are importing.
        - For example, one Langmann album (UL_1001)
    - One field in each subsequent record must contain the file name of the item you are importing.
        - For example, file names for each image in the album (UL_1001_0001, etc.)
    - The filename field for the compound object will be blank but for subsequent rows (the items' metadata) should contain the file names from the service directory.
    - You will need to manually change the files names since Archivematica populates that field with the file names from the

DIP. Currently the images in the DIP are generated from the objects directory (unedited tifs) and we want to upload images from the service directory (edited tifs)
- Once you have finished edited the "simple.txt" you are ready to upload to CONTENTdm
- In the Project client, open your project or create a new project for your upload
- Click on "Add Compound Objects" and select Add using "Compound Object Wizard" from the drop down menu and click "Add".
- Select "Document. Document with no hierarchy."
- Select "Yes" for "Are the compound object structure and metadata defined by a tab-delimited text file?"
- Browse to the "simple.txt" file
- Select "Import files from a directory and browse to the "service" directory for your upload
- Choose Yes to generate display images, and click Next
- Select Label pages using tab-delimited text file
- Select No transcripts
- Uncheck Create print PDF
- Review the Summary of settings for import and click Finish and then Click Close
- Select your object from the list and click Finish to start the Import

Once the items are added to the client, check over the metadata to ensure its accuracy. When everything is ready, click Select All and Upload for Approval.

After completing the CDM upload, you may need to log on to CDM admin to approve the added items and index the collection, although in some cases this will be done by an supervisor or other staff member.

- Before approving/indexing find the Conch picture on the magnet—only one person at DI should be approving/indexing at any time.
- Select the collection from the drop down menu and click change.
- Click on the Items tab and approve. You should find the newly added items in the approval queue.
- Click approve all, go.
- Click on Index and index now.

5. Dashboard Clean Up

The transfer and ingest tabs build up a list of everything you have processed. The Dashboard does not automatically clear itself. If the lists get too long it will

significantly slow performance and responsiveness of the page. Thus, it is necessary to clean the list manually by individually removing each completed package on the Transfer and Ingest tabs. This should be done as you work to keep the list small. Ensure that the package has completed processing successfully before removing. Keep any packages with errors so that we can refer to them when sorting out the issues.
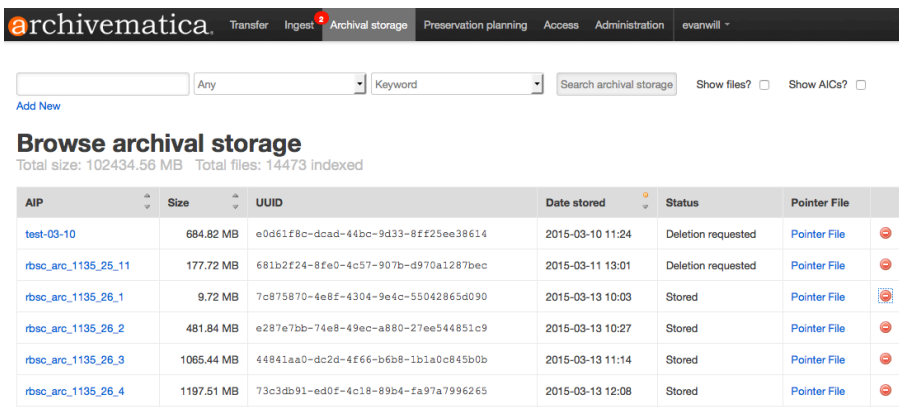
To clean the lists:

- Click on the red circle on the right side of a package listing.
- A pop up comes up, click on Confirm to remove the listing.
- Note: removing the package from the list does not delete the package from the server or stop processing. It only clears the list on the Dashboard.

Issues and Troubleshooting

In general, if errors happen at any point or you reject the package, you will need to completely redo the transfer process—there is no Archivematica back button! Be sure to document any problems so that they can be fixed.

Before re-doing the transfer, remove the AIP if it was completed.

- **Delete the AIP:** Using the Archival Storage tab, find the AIP with an error and click on the red delete icon on the right. Archivematica will ask for a reason for deletion, please enter a short description. The deletion request will be approved by a supervisor.

- **Fix DIP metadata:** If you used the DIP TXT to upload a batch of files to CDM before the AIP failed, you will need to ensure the metadata is updated when you have a successful AIP. In some cases it may be easier to delete the access images on CDM and re-upload the DIP in the Project Client. In other cases, it will be more efficient to simply update the AIP UUID and File UUID fields with information from the successful AIP.

Common issues:

- **Package Misnamed**: typing error when starting the transfer. If a misnamed package completes processing, it results in a misnamed AIP which will prevent or confuse proper retrieval. If you catch this error at any point, reject the package. If it completes processing before you catch the error, document the incorrect name and transfer time. Delete the incorrectly named AIP. Re-do the whole transfer process with the correct name.
- **Index failures**: packages are indexed by ElasticSearch during several micro-services. Occasionally ElasticSearch has issues, causing fails during indexing jobs (in Complete transfer, Prepare AIP, Store AIP). The service needs to be restarted, which requires an IT ticket. Packages with index failures should be redone.
- **Characterize and extract metadata failed**: the Job characterize and extract metadata is highlighted red, meaning one of the files in the package failed in some way. Click on the cog icon to view the error report. If the report contains only a few entries, it may be possible to identify the specific problem files by clicking on "show arguments." You may encounter files that were corrupted at some point, and appear as zero-byte files on the servers. You will need to track down a different version of the corrupted file to fix the issue. However, sometime these errors are caused by the tools used by the micro-service needing to be restarted. You will know this is the case when the same error occurs multiple times with different packages, and the error report shows all files in the package as errors. This issue requires an IT ticket.
- **Validation failed:** the Job Validate formats shows failed. Click on the cog icon to view the report. Occasionally this fail is caused by the tool JHOVE needing to be restarted (requires an IT ticket). Otherwise, the solutions are the same as Charactize and extract metadata issues.
- **Wrong file path in *metadata.csv***: If the file path given in the filename/parts column of the metadata.csv has errors, it will generate errors in several Jobs. Carefully check over the metadata.csv against the original file names to locate the issue.