Archivematica DIP upload to AtoM for Digitized Materials

Introduction

RBSC Archives utilizes AtoM to provide public access to archival descriptions and digital objects. UBC Archivematica pipelines are integrated with AtoM to allow direct upload of DIPs. This ensures that a link is made between the files in archival storage and the description hosted on AtoM. This document describes the practical steps for processing digitized materials created at the Digitization Centre through Archivematica with DIP upload to AtoM.

> 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 RBSC Archives instance of AtoM, visit: http://rbscarchives.library.ubc.ca
>
> For more information about AtoM, visit the project site: https://www.accesstomemory.org

Archival Description on AtoM

To better understand the organization of content on AtoM, it is necessary to have a basic familiarity with the concepts of archival description. Description can be divided into a series of hierarchical levels proceeding from general to specific. Fonds is the highest unit of organization in the archives, referring to the entire body of records from a single source.[1] The description of a fonds in AtoM is subdivided into levels following archival terminology, with each element having its own page:

Fonds > Series > Subseries > File > Item.

These levels of description are represented in the "Treeview" navigation pane on the left side of each page. Clicking the arrow next to an item in the treeview reveals levels of description nested below the element.

[1] See: SAA Glossary, http://www2.archivists.org/glossary/terms/f/fonds.

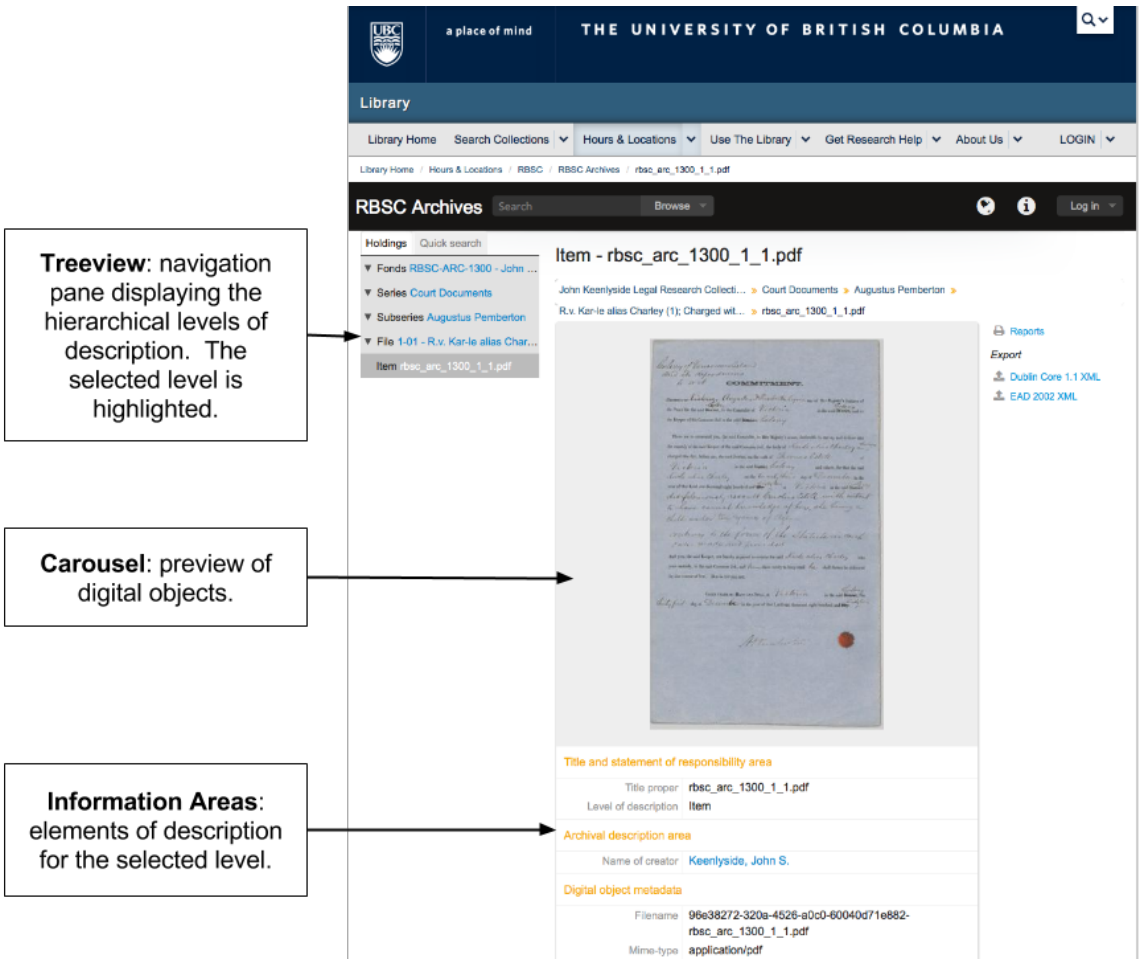


Figure 1: an item displayed on RBSC AtoM.

In general, for textual materials the Digitization Centre will scan the documents creating master TIF images for each page. To create a more usable format for access, the page images for a single document are combined into a PDF. These access derivatives are uploaded to AtoM and will generally appear at the Item level of description. AtoM calls them digital objects.

> As an example, you can explore the John Keenlyside Legal Research Collection (RBSC-ARC-1300) on AtoM: http://rbscarchives.library.ubc.ca/index.php/john-keenlyside-legal-research-collection
>
> The fonds is divided into seven Series. If you click on Series: Court Documents, you will see that it is divided into four Subseries. If you click on Subseries: Augustus Pemberton, you will find a list of around 80 Files. Click on one of the Files and you will see the Item (a PDF) that we digitized.

Note About Terminology

One complication of working with Archivematica and AtoM is the overlap of archival and computer terminology.

To avoid confusion, in this document:

- "digital file" refers to computer files such as PDF or TIF.
- "directory" refers to what is often loosely called a folder on computers.
- "file" refers to the level of archival description.
- "package" refers to the group of digital files sent to Archivematica. This will usually represent a single archival file.

Preliminary Steps

Before starting work in Archivematica, a skeleton description must be complete in AtoM *and* the transfer packages must be set up.

RBSC Archives staff will create the AtoM description. Our digital objects must be attached to an existing "parent" level of description. In general, this means we will send our digital objects to a File level description. The objects will appear as Items nested below that File. Thus, the transfer packages we create will correspond to the parent level of description.

> For example, one archival file in the Keenlyside collection is titled "Regarding the Queen Charlotte Mining Company" with the identifier "RBSC-ARC-1300-8-42". The transfer package would be a directory named *rbsc_arc_1300_08_42* that contains all the digitized documents in the archival file. The package would be sent to the description of "Regarding the Queen Charlotte Mining Company" on AtoM.[2]

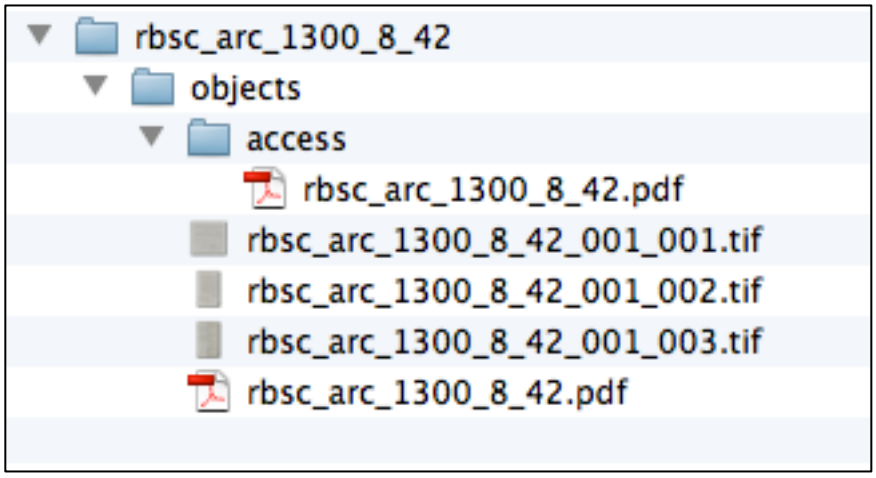
The transfer packages need to be set up on the network storage drives. If the directories of digitized files have been organized following our general guidelines, the process of creating the packages will be fairly simple. Some workflow automations may be available to help create the necessary structure from existing directories (using the Automator app).

In general, we will create a top level directory for the collection/fonds with subdirectories containing all the packages to be transferred.

---

[2] http://rbscarchives.library.ubc.ca/index.php/regarding-queen-charlotte-mining-company

Each transfer package will be a directory named with the indentifier and will contain a subdirectory named *objects*.[3] The *objects* folder should contain all the digital objects for the archival file, including the master images (TIF) and access derivatives (PDF). It will also contain a subdirectory named *access*. The *access* folder should contain a copy of the access derivatives that we want to display on AtoM. This means there is two copies of the access PDF, one in *objects* and one in *access*, with file names that exactly match. For example, look at the screenshot below:



In the example, the transfer package is named *rbsc_arc_1300_8_42*. The objects directory contains three TIF master images representing the three pages of the original print document. Those TIF images have been combined into a single PDF for access (the access derivative). Both the *objects* and *access* directory contain copies of the PDF.

Keep in mind that more complex files may require more than one access PDF. Thus, your access directory may have multiple PDFs. They will all appear under the parent level of description.

Since we manually create the access derivatives, rather than have Archivematica create derivatives during processing, this set up is called "manual normalization." The key points: the access directory contains the digital files we want to display on AtoM; and there are copies of the access derivatives in both the *access* and *objects* directory with exactly the same file names.

Considerations when setting up the transfer directories:

- **Proceeding Zeros:** directories will not sort correctly in Archivematica without proceeding zeros in the digital file names. It is best practice to always use proceeding zeros. The Renamer application (on the Windows

[3] Archivematica will automatically create additional subdirectories such as "logs" and "metadata" so it is unnecessary for us to create them.

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 directory 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 collection. For example, you may check to ensure that the PDFs have the same number of pages as you have TIF images. You may check the digital files against the finding aid to ensure everything is organized correctly. 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!

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 appears 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.

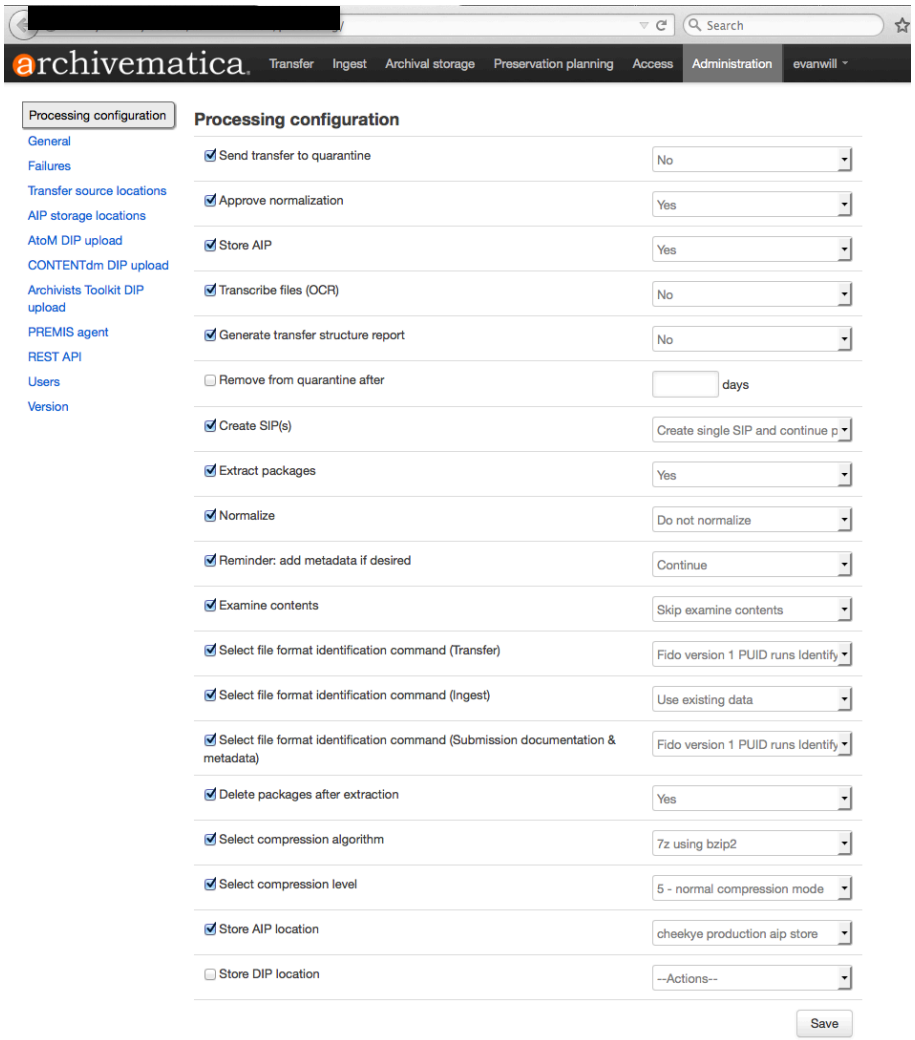
Processing Configuration options are used to automate the decision points during Archivematica processing when the same actions will be repeated for every package. If none of the options are checked, the Dashboard will ask for input about each step on the Transfer and Ingest tabs. Each decision point 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:



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	Do not normalize	Normalization is the process of creating preservation and access derivatives of the original files. The workflow described in this document uses “Do not normalize”

		because we have manually normalized the files, i.e. created PDF access copies and put them in the *access* directory.
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 since it will be sent to

		AtoM directly.

Archivematica Processing

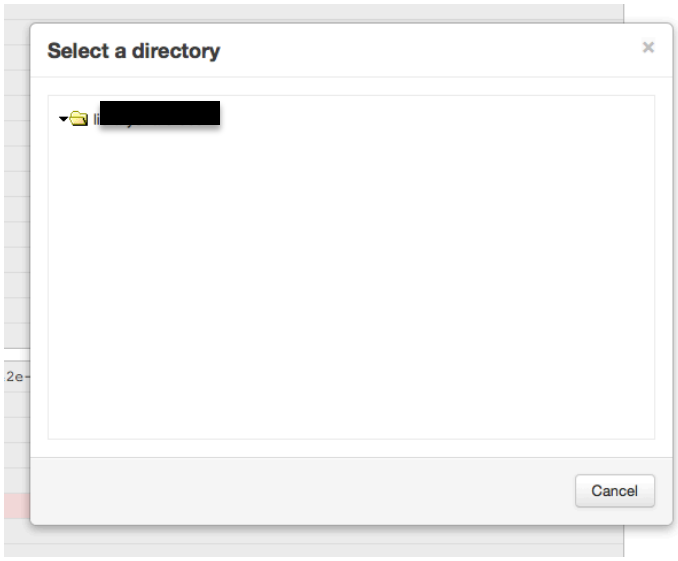
After checking the Processing Configuration, you are ready to start transferring packages to the system. This section describes the step-by-step process:

1. Transfer

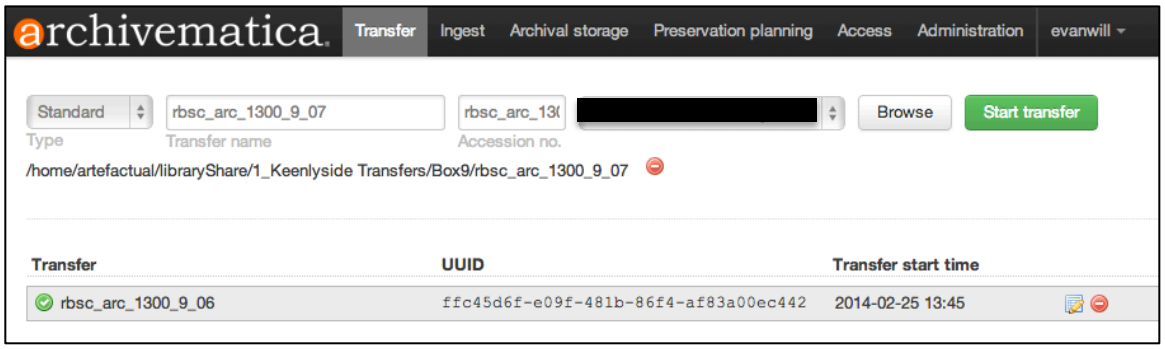
Navigate to the Transfer tab. Transfers are initiated using the bar at the top of the page. Fill in the form using these parameters:

- **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.

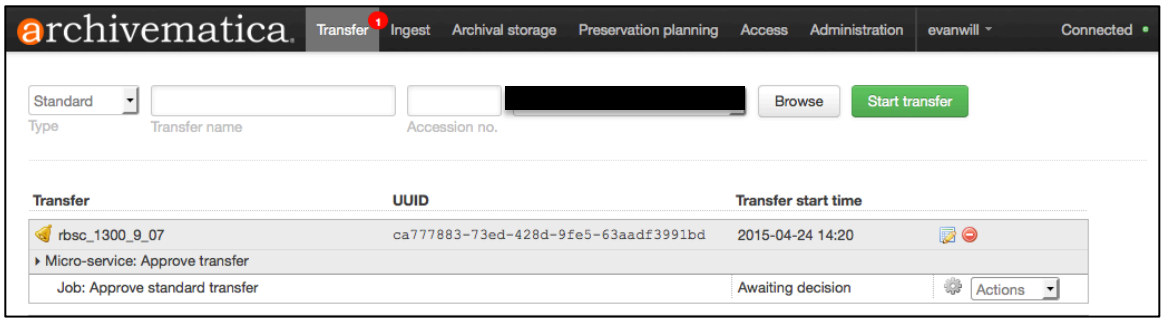


The file path and name of the directory you added will appear below the transfer name—double check to ensure you have added the correct file!



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…

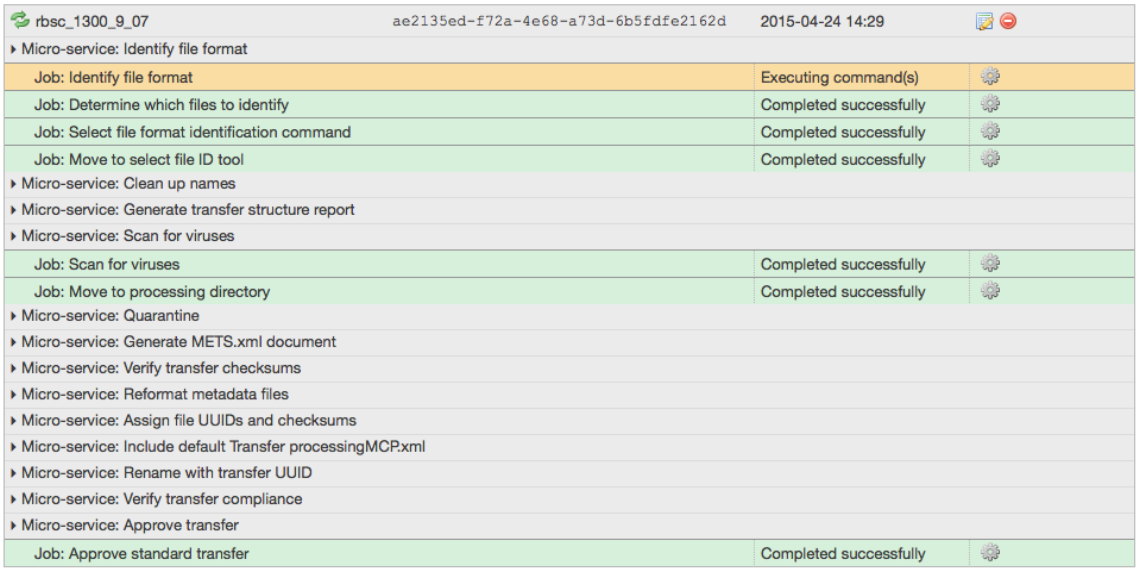
Once the upload is complete, the package will appear on the Dashboard list. The package will be at a decision point asking for you to "approve" transfer. This step can not be automated. Ensure the package is named correctly. If all is good, choose yes from the drop down menu!



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.



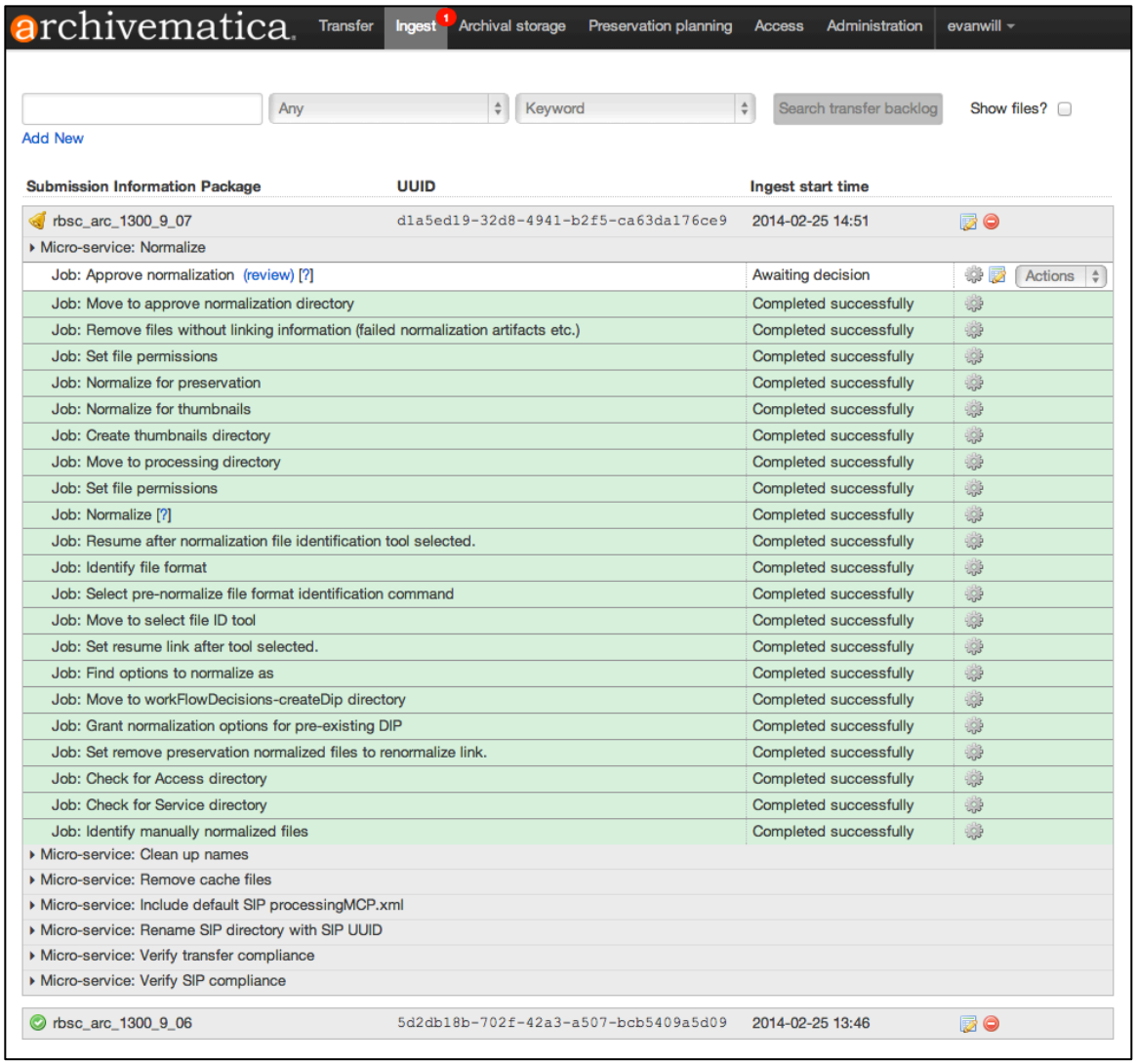
When the Transfer processes are done, a SIP will appear on the Ingest tab.

> 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.

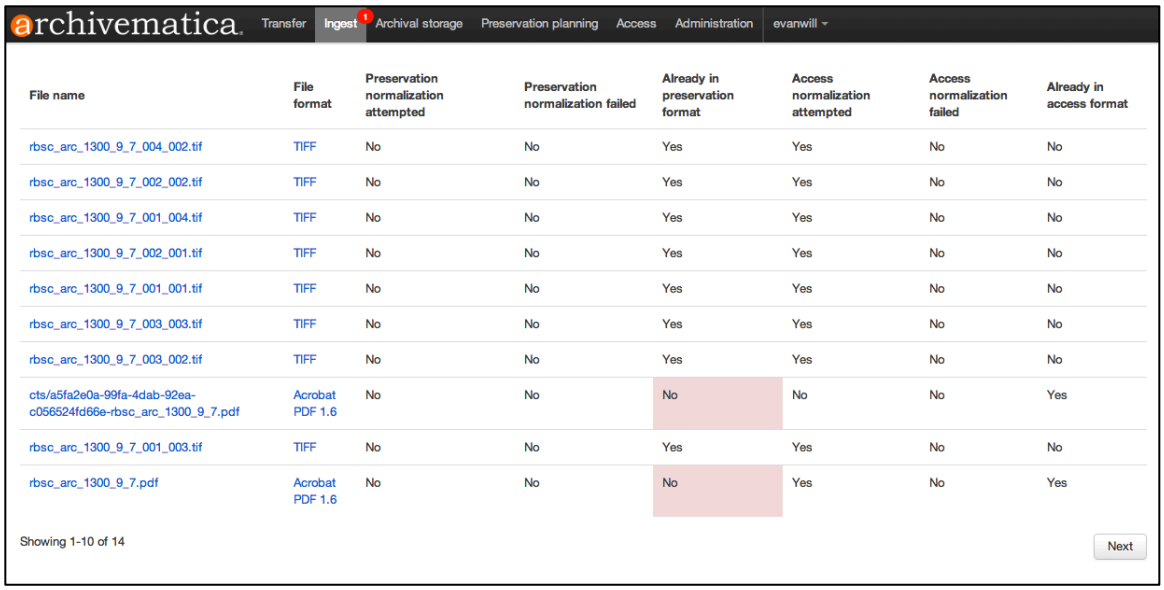
2. Ingest and Normalization

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, this stage will be automated except Upload DIP.

If everything is running smoothly, you will skip to the next step. 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:



- 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).



- 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 file types are present. For example, in this workflow we should see TIF masters for each page plus a PDF for storage and a PDF for access. If the correct number of files is not present, there is likely an error in the transfer directory. The most common error is an empty access directory, resulting in less PDFs than expected.

3. Upload DIP

Since we provided Archivematica with manually normalized access derivatives, the next step is to send those access copies to AtoM. This links the archival description on AtoM to the Archival storage with UUIDs. It also uploads the access derivatives to the AtoM server for public access. Since Archivematica and AtoM are both developed by Artefactual, this linkage is fairly straightforward and robust. Please

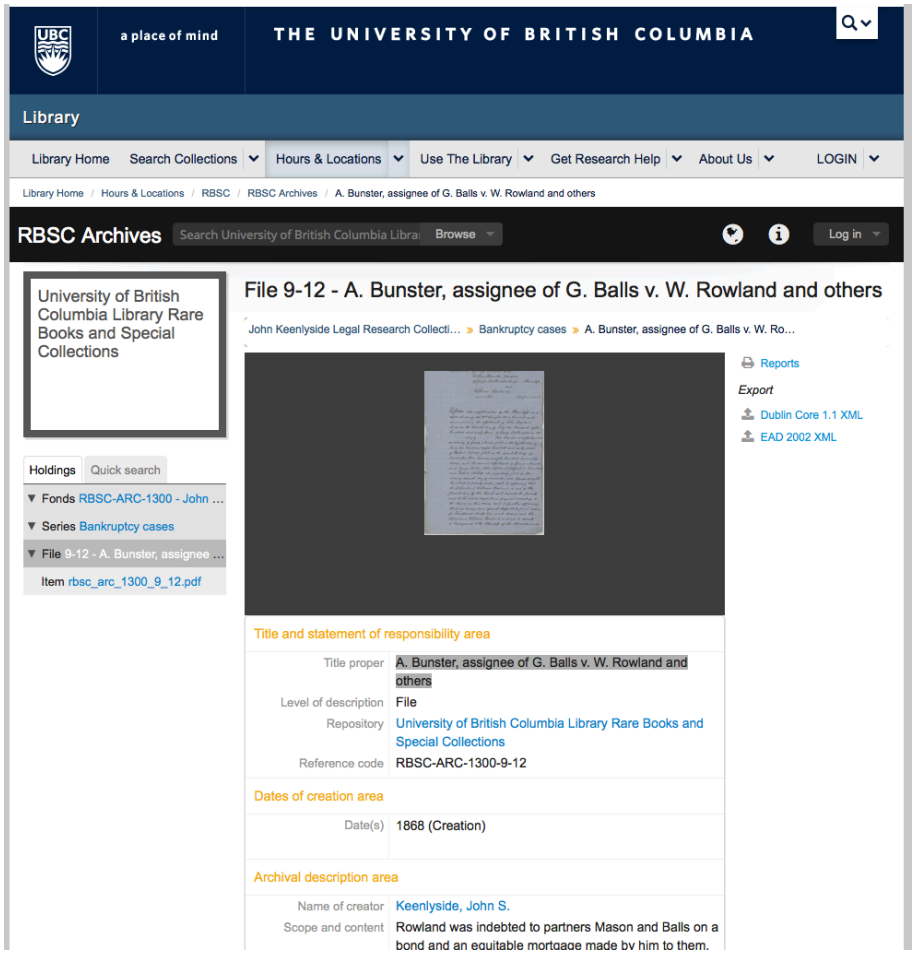
note that you will need an account to log into AtoM. This will be set up by the RBSC archivist, ask your supervisor to request one for you.

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.

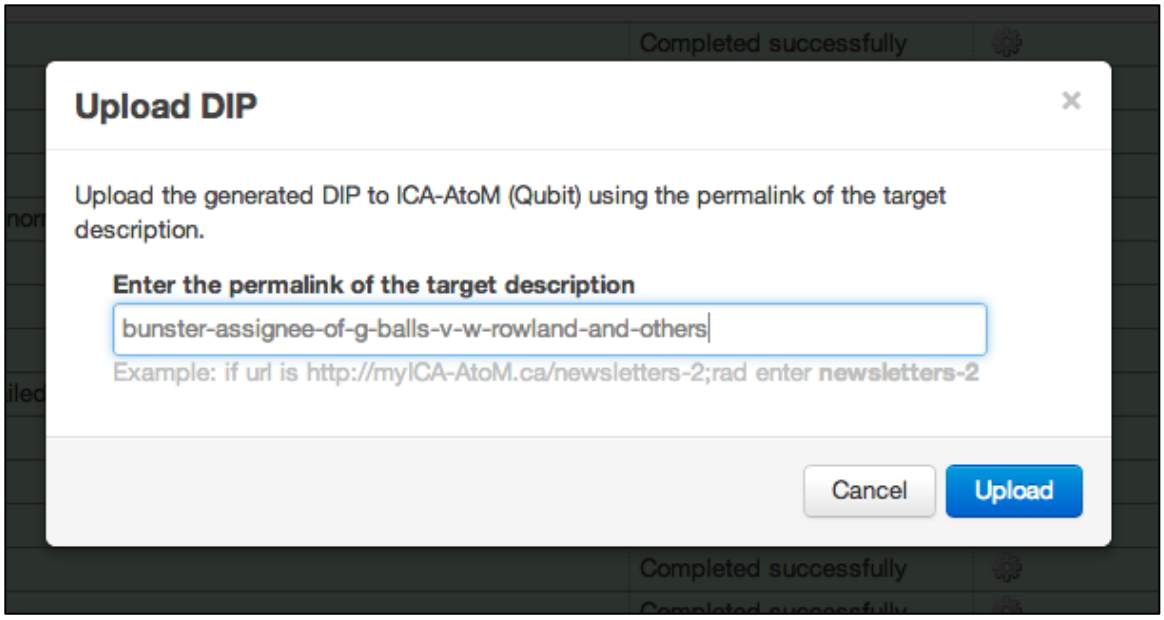
In the Micro-service: Upload DIP entry:

- Choose the option "upload DIP to AtoM" from the drop down menu.
- A pop up window will appear, add the permalink for the parent level of description on AtoM. The permalink is the unique section of the URL identifying the page. It is easiest to have the AtoM page for the parent open in the browser so that the permalink can be cut & pasted.
- Click upload.
- When the upload completes on Archivematica, refresh AtoM to ensure the digital files appear in the correct location.

For example, if we are uploading the DIP for the package rbsc_arc_1300_9_12, visit the page for the File level description, "A. Bunster, assignee of G. Balls v. W. Rowland and others", http://rbscarchives.library.ubc.ca/index.php/bunster-assignee-of-g-balls-v-w-rowland-and-others



In this example, the permalink is “bunster-assignee-of-g-balls-v-w-rowland-and-others.” Copy the permalink, paste it into the Upload DIP popup window, and click upload.

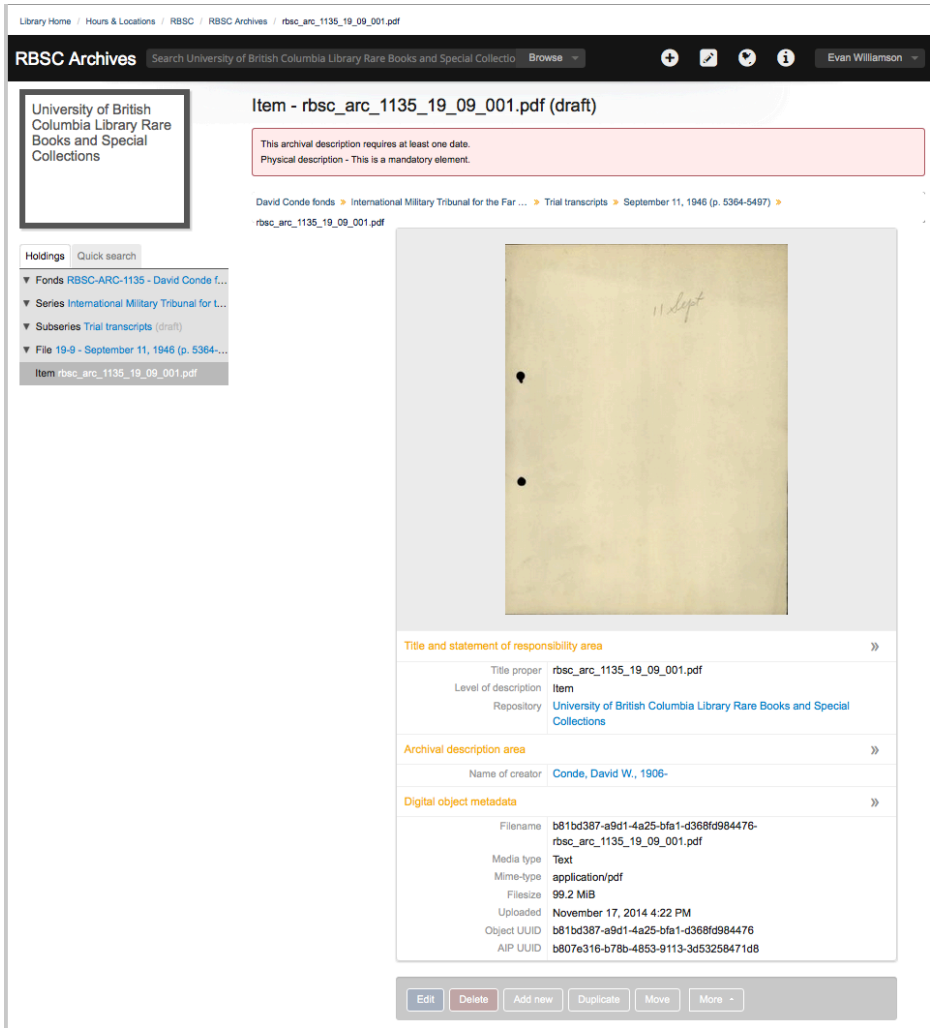


Since you will repeatedly return to AtoM to find slugs, it is easiest to keep open the collection level page with the tree navigation on the left expanded. Open new tabs for each file from the tree as you work. Copy the slug, and keep the tab open (since you will need it for the next step).

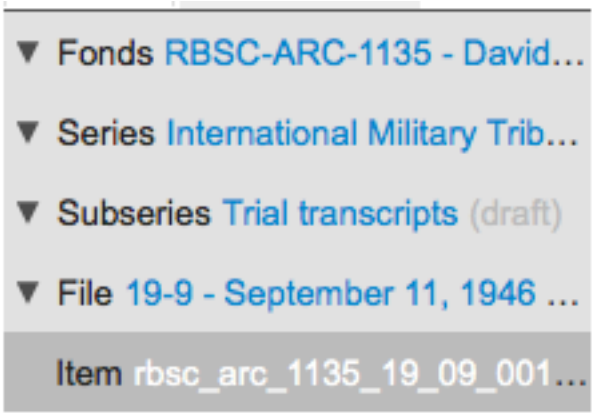
Occasionally the Upload to DIP step gets stuck after you enter the slug. You will notice that nothing happens after ten seconds and that the "Job: upload DIP" does not change color. Simply go back and choose "upload DIP to AtoM" again, the pop up window will come up with the slug you entered earlier, and click Upload again. It will usually start working.

4. Check DIP on AtoM

Once the Upload DIP microservice completes, it is best practice to immediately QC the upload on AtoM. If you left the page open, simply refresh the AtoM description you sent the DIP to (it may take a few minutes for the server to process the files you sent), otherwise navigate back to the correct parent description. The access file(s) you sent should now appear nested below that element of description— ensure the correct digital object shows up in the correct location (i.e the PDF has the correct file name and appears at the item level below the correct file level description). It will be marked "draft" (if you are not logged into AtoM, you will not see it).



For example, with David Conde fonds we send a PDF to the file level description—the digital object will appear as the item level nested below this element. The PDF file name includes the full access identifier, which should match the file level description.



Please note, that we want a link between the AtoM description and the archival storage—therefore, if something goes wrong with the DIP upload (you accidentally reject it, send it to the wrong place, etc.), you will need to redo the entire transfer process. See Issues and Troubleshooting below for more information.

At this point, the digital objects on AtoM are drafts. They will need to be published before they are accessible to the public. In most cases we will leave this step to employees of RBSC. For reference, the digital objects on AtoM can be published individually by clicking the Edit button at the bottom of the description. Next click "administrative options", change the status from "draft" to "published", then click "save". The item will now be publicly available. Furthermore, if you change the published/draft status of any element of description, the status of all children of that element will also change. Therefore, it is possible to publish a batch of draft digital objects by changing the status of a higher level of description to draft, then back to published.

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.

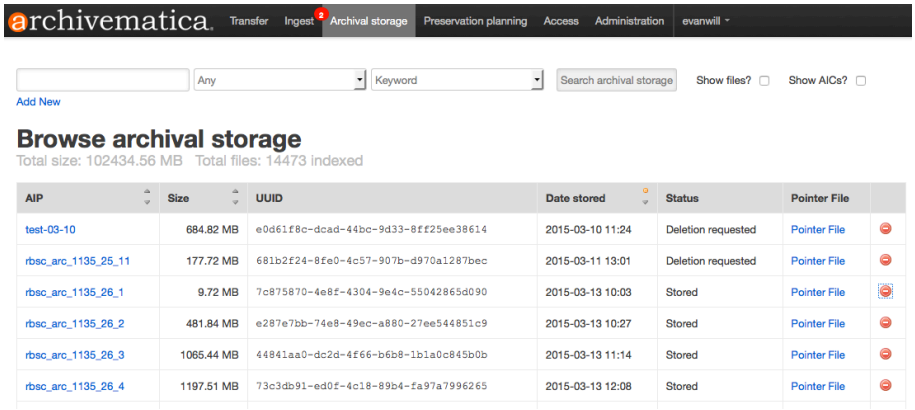
Issues and Troubleshooting

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.

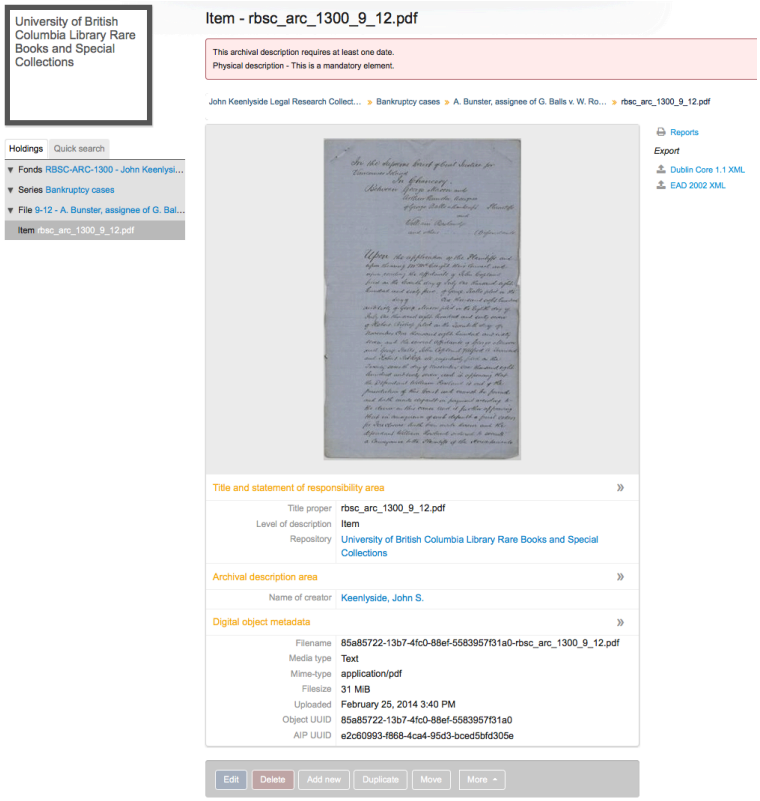
Before re-doing the transfer, remove the AIP and DIP (digital object on AtoM) if either was completed.

- **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.



- **Delete the DIP:** Please delete any access files on AtoM that are NOT linked to a successful AIP package. To delete a digital object from AtoM, navigate to the object. This will usually be an Item level description with a title equal to the digital file name including .pdf. Ensure that you are not on the parent level description! At the bottom of the description, click on the Delete button and follow the dialog boxes. Redo the complete transfer process.



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 and the digital object on AtoM. Re-do the whole transfer process with the correct name.
- **DIP upload failure+AIP successful *or* DIP upload successful+AIP failure**: it is important for the description on AtoM to be correctly linked to the archival storage. Since there is no "back" button, an error in either DIP or AIP requires us to re-do the whole package. Delete the AIP and the digital object. Re-do the transfer process.
- **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.
- **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." For example, the argument may look like:

  ```
  characterizeFile_v0.0
  "%sharedPath%watchedDirectories/workFlowDecisions/extr
  actPackagesChoice/rbsc_arc_1135_19_08-2c1ef660-7377-
  460a-ab6f-
  8d9e7e448b0c/objects/rbsc_arc_1135_19_08_001.pdf"
  "048e9d46-448b-42ec-90db-3bc6a52255dd" "2c1ef660-7377-
  460a-ab6f-8d9e7e448b0c"
  ```

  If you look at the end of the file path (it looks strange because of the UUIDs added to the package name), you will see the file rbsc_arc_1135_19_08_001.pdf referenced. Check this file in the original transfer package. This commonly happens with PDFs which were created using ABBY several years ago. The solution is to open the PDF with Acrobat and simply re-save. Occasionally this may happen with TIFF images as well, again the simplest solution is to open the problem file with PhotoShop and re-save. You may also 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.
- **Normalization fails:** If normalization micro-service gets a failure, it is most likely a problem with the access directory. The *access* directory may be empty or may have files that do not match a file in the *objects* directory.
- **Normalization report errors:** When checking the normalization report you may notice missing/incorrect digital files or errors. Enter "Reject" for normalization, which will stop processing on the package. Check the directory structure for the package. Common problems include no PDF in the access directory, no TIFs in the objects directory, or a zero-byte file in the directory (errors generated when creating the transfer packages). Once the issue is fixed, re-do the transfer.