DRAG AND DROP PROCESSING

Sightline Help Documentation

Consilio®

Together. Stronger.

Table of Contents

Drag and Drop Processing

As a Review Manager User, you can use drag-and-drop processing to upload files to a project in Sightline, process the data, index text for searching, run analytics, and make them available for review.

The "Datasets" menu option provides the landing page for the drag-and-drop processing. This menu option is presented to only Review Managers and Project Administrators. You may also restrict access to this feature from certain Review Manager Users is needed as either a Project Administrator or Domain Administrator through the Manage->Users page.

From the "Datasets" landing page, users can manage the datasets in a project, initiate uploads for processing, monitor the progress of a dataset being processed and loaded, and review the summary and exception reports of a previously published dataset.

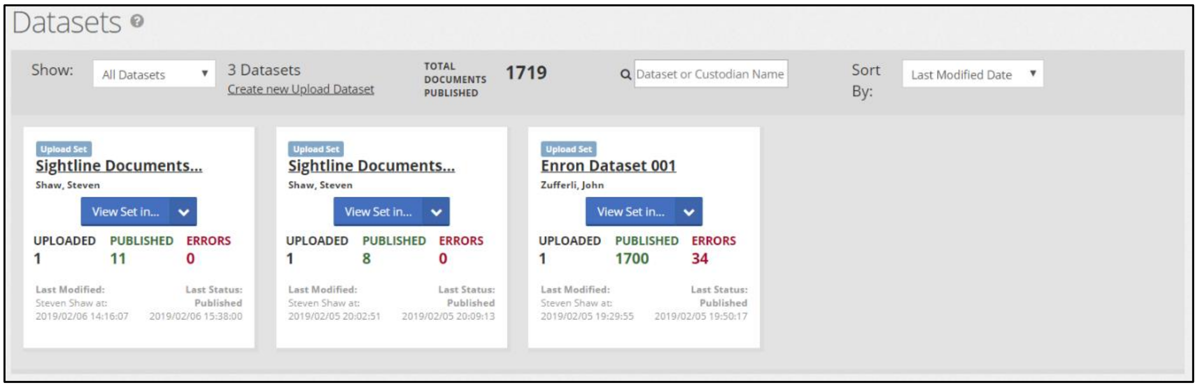
A dataset can be:

- An Upload Dataset: A dataset that has been uploaded and processed through the self-service drag-and-drop user experience in Sightline.
- A Mapped Dataset: A dataset that has been processed externally and has come into Sightline via the Ingestions module

Datasets Main Page

On the Dataset main page, users can:

- Filter the Dataset tiles: Users can filter the Dataset tiles to only show Datasets of certain types, or of certain Custodians or with certain Dataset names.
- Sort the Datasets: Users can sort the presentation of the Dataset tiles by the Last Modified Datetime, the Last Status, the Last Modified User who modified the Dataset or by Custodian Name attribute.
- The Datasets main page also shows a readout of the count of unique DocIDs that were published in a project.



Datasets

Show: All Datasets | 3 Datasets | Create new Upload Dataset | TOTAL DOCUMENTS PUBLISHED 1719 | Dataset or Custodian Name | Sort By: Last Modified Date

Upload Set	Upload Set	Upload Set
Sightline Documents...	Sightline Documents...	Enron Dataset 001
Shaw, Steven	Shaw, Steven	Zufferli, John
UPLOADED 1	UPLOADED 1	UPLOADED 1
PUBLISHED 11	PUBLISHED 8	PUBLISHED 1700
ERRORS 0	ERRORS 0	ERRORS 34
Last Modified: Steven Shaw at: 2019/02/06 14:16:07	Last Modified: Steven Shaw at: 2019/02/05 20:02:51	Last Modified: Steven Shaw at: 2019/02/05 19:29:55
Last Status: Published 2019/02/06 15:38:00	Last Status: Published 2019/02/05 20:09:13	Last Status: Published 2019/02/05 19:50:17

User Actions

Users can perform the following actions from the Datasets page:

- Create a new Upload Dataset
- View all datasets in the project with the current state of each.
- Monitor the progress of a Dataset being processed and ingested
- View the documents of a Dataset in DocList, DocView or Tally
- View the Dataset summary report
- View the Dataset exception report

Dataset Tiles

Each tile in the Datasets main page represents a discrete Dataset. Users will be able to take different actions based on the current status of the Dataset as shown in the tile.

- Draft: Users can click on 'Upload/Manage' to add files to the Upload Dataset before initiating processing. Users can also delete any Dataset that is in Draft status (only).
- In Progress: Users can click on the Dataset name to go the Dataset Progress Page.
- Published: Users can click on the Dataset name and see the Dataset Summary Report Page. Users can also click on the 'View Set In...' button and can navigate to either DocView, DocList or Tally to view all records in that Dataset.

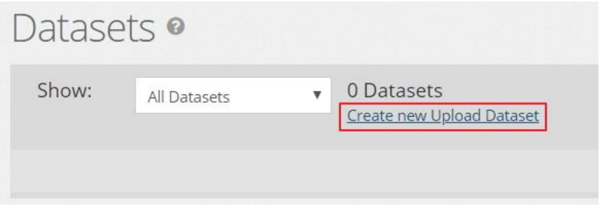
Each Dataset tile presents useful information related to the Dataset, including:

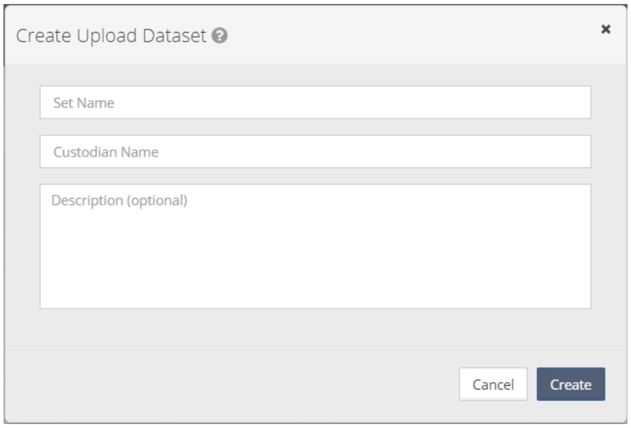
- Uploaded Count: The count of unique files that were successfully uploaded for that Dataset
- Published Count: The count of unique DocIDs that were published into Sightline for the Dataset
- Error Count: The aggregate count of errors encountered as the Dataset went through the steps of processing, upload and publishing into Sightline
- Last Status: This is the last successfully completed status for the Dataset and the system datetime stamp of completion.
- Last modified: This is the datetime stamp when the Dataset configuration was last modified, and which User made that modification

Creating a New Dataset

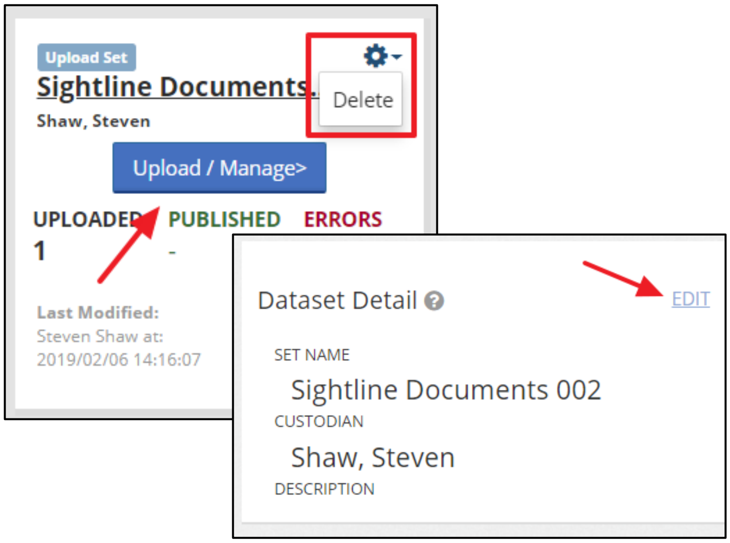
In order to upload data, a Dataset must first be created. A Dataset is any grouping of data with a common Custodian. Multiple files can be uploaded into a Dataset, but Consilio recommends keeping uploads to under 20 GB. Uploads can be PSTs or loose efiles, however, all files should be placed into a compressed container file, like a ZIP.

To create a new Dataset, go to the Datasets page by clicking on the menu option on the left-hand side. Once in the Datasets page, click the link to “Create new Upload Dataset”. This will open the Create Upload Dataset window where you will give the Dataset a name, a Custodian value and an optional Description.





Until the data in the Dataset is processed, the information entered into the Create Upload Dataset window can be edited. New data can be uploaded to an unprocessed Dataset and previously uploaded data can be removed from the dataset. An unprocessed Dataset can also be deleted. Once the Dataset is processed this information is locked.



Dataset Name

While you are free to name the dataset with whatever name you wish, it is Consilio's recommendation to name the dataset with the Custodian's name, the type of data, e.g., email, efiles, etc., and then a numerical counter. Joe Smith's first set of email data could be named: Joe Smith Email 001. Jane Doe's second set of efiles could be named: Jane Doe Efiles 002.

If there is a naming convention that is already in use, please feel free to continue to use that. If you have any questions about how to name a dataset, you can reach out to your project manager.

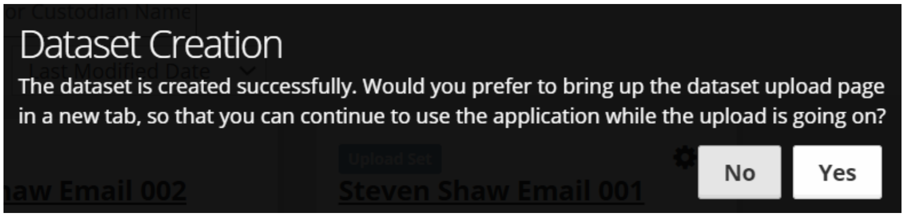
In the event that multiple datasets will be uploaded for numerous custodians, it is important to use a consistent, clear naming convention to be able to identify the data as the case progresses. You can also make use of the Description field to help keep track of your data.

Custodian name

The Custodian name, and manner in which you enter the name, should be considered during the creation of the dataset. This information will be populated in the Custodian field in Sightline and would be used for Production purposes and cannot be changed after the data has been processed. It is important to be careful and consistent. If, for example, you decide to use ***LastName, FirstName*** make sure you do this for all Custodians. Also, be sure to use the same first name where nicknames or shortened names are possible, e.g., Steve Shaw vs Steven Shaw.

Using Multiple Tabs for Upload

When you click the "Create" button on the Create Upload Dataset window you will be presented with an option to open the new Dataset in a new Tab.



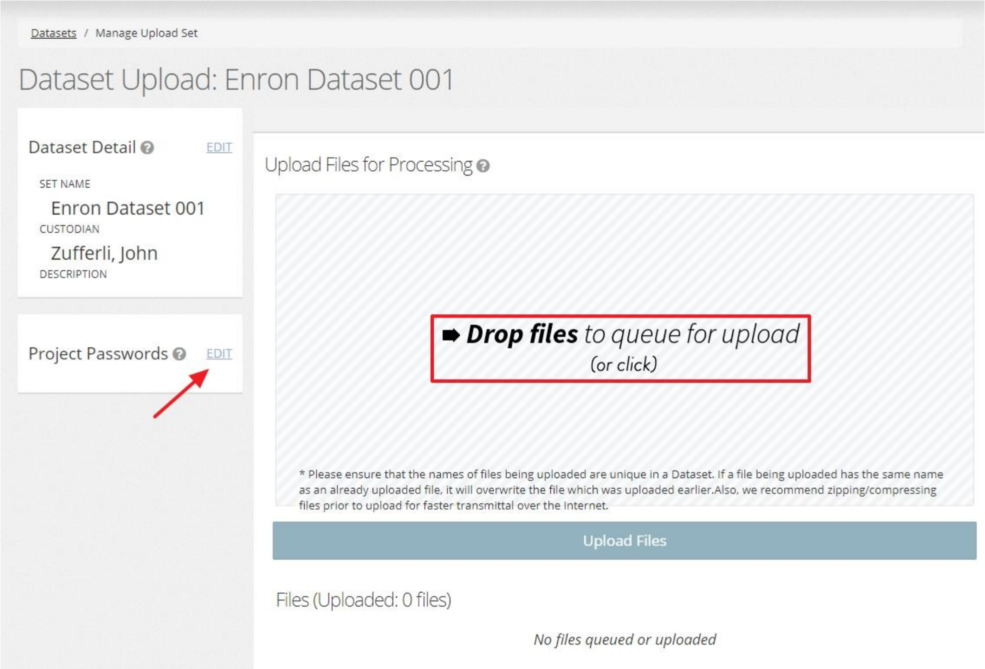
Dataset Creation

The dataset is created successfully. Would you prefer to bring up the dataset upload page in a new tab, so that you can continue to use the application while the upload is going on?

If you choose Yes, Sightline will automatically open the Dataset in a new browser tab. This will allow you to upload the data in one tab, while doing other work – searching, database management, creating and uploading additional data – in the original tab. If an upload active in the secondary tab, the User's session will not be timed out in the primary session. If for some reason the application faces any technical glitch in being able to launch the upload in a new tab upon the user's choice, it will then attempt to continue to bring up the upload in the primary tab itself.

If you choose No at this point, Sightline will open the Dataset in the original tab and you will not be able to navigate away from that page until the data has completed loading.

Uploading Data



Datasets / Manage Upload Set

Dataset Upload: Enron Dataset 001

Dataset Detail

SET NAME: Enron Dataset 001

CUSTODIAN: Zufferli, John

DESCRIPTION

Upload Files for Processing

➡ **Drop files** to queue for upload (or click)

* Please ensure that the names of files being uploaded are unique in a Dataset. If a file being uploaded has the same name as an already uploaded file, it will overwrite the file which was uploaded earlier.Also, we recommend zipping/compressing files prior to upload for faster transmittal over the Internet.

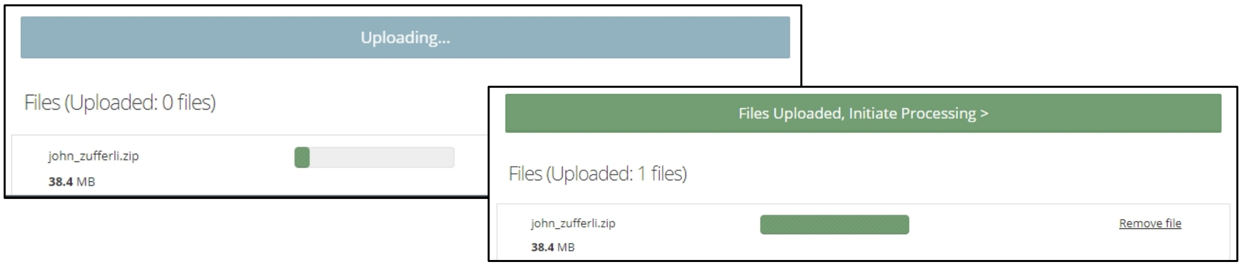
Files (Uploaded: 0 files)

No files queued or uploaded

After the Dataset has been created you can upload data through the interface. To upload, drag files from the local machine to the main area of the window that reads, "Drop files to queue for upload (or click)". This will initiate the upload process. Please ensure that the files are under 20 GB and are in a compressed container file. Files can also be uploaded by clicking in the main window to open a local window to upload. If the files are password protected, you can add passwords to the Project Passwords list by clicking the Edit link and adding one password per line to this file.

Processing Data

Once the data has completed upload, the system will provide the ability to initiate processing. After processing has initiated, files can no longer be added, deleted or edited. Clicking Initiate Processing will display the Dataset Progress Page.



Files (Uploaded: 0 files)

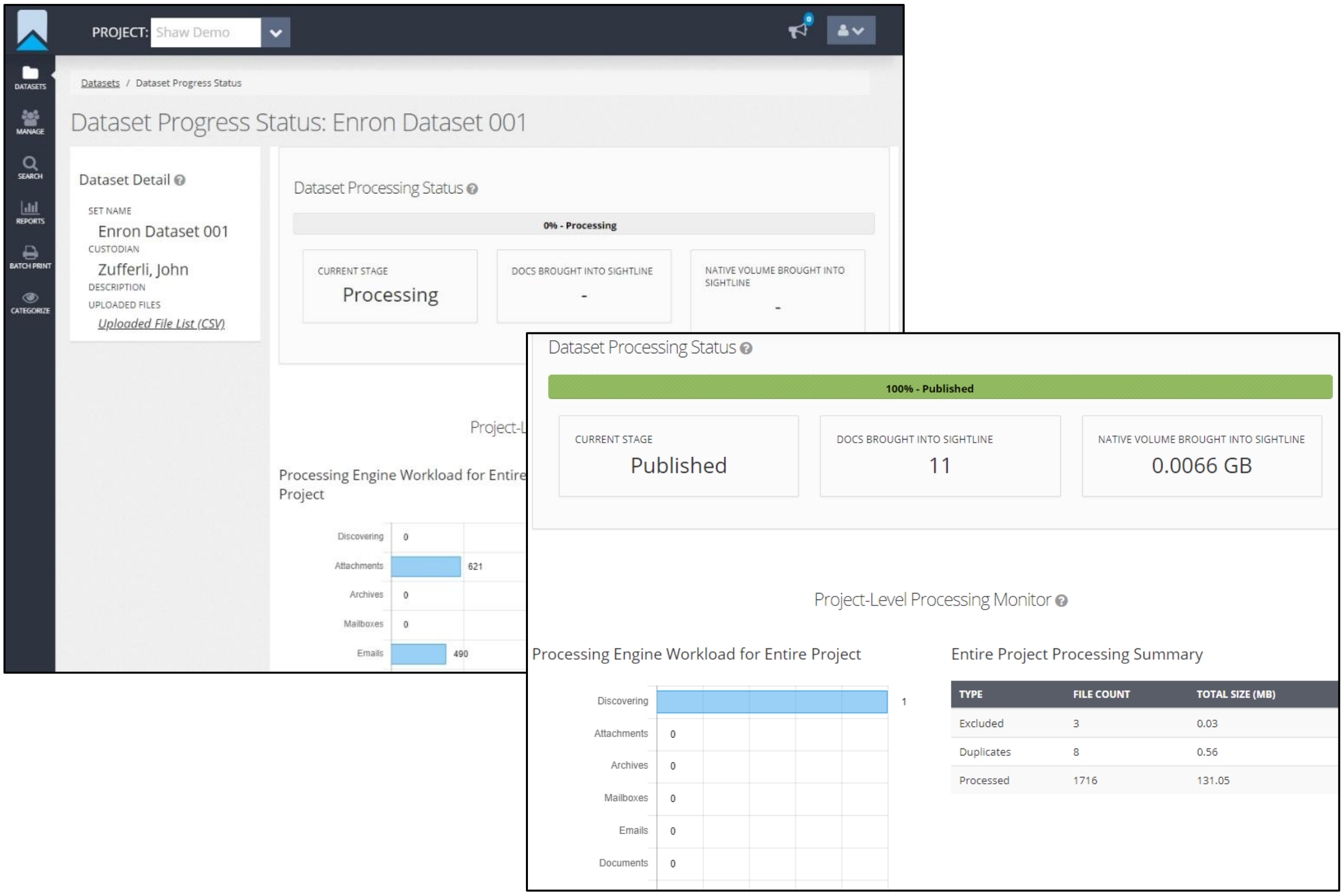
john_zufferli.zip
38.4 MB

Files (Uploaded: 1 files)

john_zufferli.zip
38.4 MB

Dataset Progress Page

The Dataset Progress Page presents real-time progress information for an Uploaded Dataset that is currently being processed and ingested in Sightline. The End-to-end Progress Bar indicates the current state of the Dataset, as well as the percentage of progress to get to completion.



Dataset Progress Status: Enron Dataset 001

Dataset Detail

SET NAME
Enron Dataset 001

CUSTODIAN
Zufferli, John

DESCRIPTION

UPLOADED FILES
Uploaded File List (CSV)

Dataset Processing Status

0% - Processing

CURRENT STAGE
Processing

DOCS BROUGHT INTO SIGHTLINE
-

NATIVE VOLUME BROUGHT INTO SIGHTLINE
-

Dataset Processing Status

100% - Published

CURRENT STAGE
Published

DOCS BROUGHT INTO SIGHTLINE
11

NATIVE VOLUME BROUGHT INTO SIGHTLINE
0.0066 GB

Project-Level Processing Monitor

Processing Engine Workload for Entire Project

Entire Project Processing Summary

TYPE	FILE COUNT	TOTAL SIZE (MB)
Excluded	3	0.03
Duplicates	8	0.56
Processed	1716	131.05

The stages of processing are:

- **Processing**: In this stage, the upload files are being discovered by the processing software, child records (e.g. attachments to emails, embedded records within files, etc.) are being extracted, metadata (e.g. created datetime, created by user, etc.) is being extracted and fielded in the database, text (e.g. the body of a document) is being extracted and fielded, OCR (optical character recognition) of image files is happening, and more.
- **Post-processing**: In this stage, the processed records are being hashed with a unique fingerprint, and are being deduplicated at the family level, against all other records that have been processed into the database.
- **Enriching Data**: In this stage, records are being stored in Sightline in efficiently-designed table views to ensure downstream speed of the software, unique files are getting identified to minimize storage on the network, file investigator is analyzing the file types and adding data to records for proper downstream operations, metadata (e.g. EmailRecipientCount, etc.) is being added to assist with more textured searching, and image records are being converted to PDF for efficient presentation in DocView and to speed up downstream printing and productions.
- **Indexing**: A searchable index is being added to with the records from the Dataset. By default, Sightline indexes all body text as well as select, commonly searched metadata fields.
- **Analytics Running**: In this step, the surviving records from the Dataset are being analyzed for (a) textual near duplicates, (b) email threading/conversation analysis, (c) language composition analysis and (d) concept - which is then exposed in many different areas of the application. Analytics-derived metadata attributes are being stored in Sightline's database and are indexed for searching.
- **Publishing**: During this stage, all records that completed all upstream stages are being set to 'active' in the project database, and the 'working copy' tables and indexes are being set to the active ones for the project.
- **Releasing**: If the Dataset was initiated by a Review Manager, there is an additional stage called 'Releasing'. In this stage, the records that were Published will be Released to the same Security Group that the Review Manager was operating in.

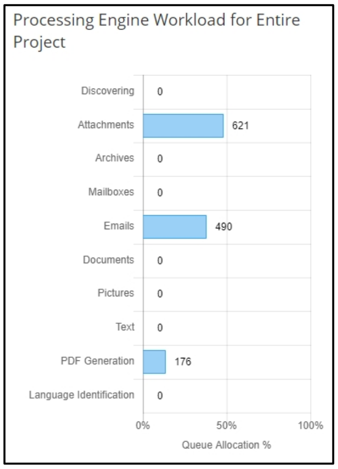
Also presented on this page are the following attributes:

- **Current Stage**: This corresponds to the stages above.
- **Docs Brought Into Sightline**: The count of unique DocIDs that made it through processing and ingestion and were brought into Sightline from this Dataset.
- **Native Volume Brought Into Sightline**: The aggregate native file sizes of all Docs Brought Into Sightline.

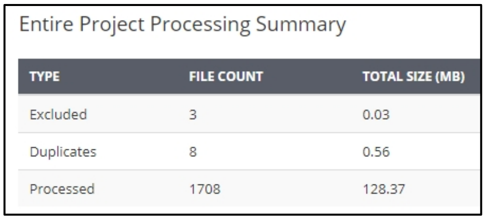
If there are any errors with the processing and ingestion of the Dataset, an error message will be presented above the progress/status bar. You may need to reach out to your administrator if you encounter such errors.

Processing Engine Workload for Entire Project

On the bottom left of the Progress Page is the Processing Engine Workload for Entire Project. This displays the real-time progress of the underlying agent server and queue activity for the entirety of the project - not specific to the Dataset you are viewing. This shows how the underlying software components are actively employed to process data and bring it into Sightline.



Project-level Processing Monitor



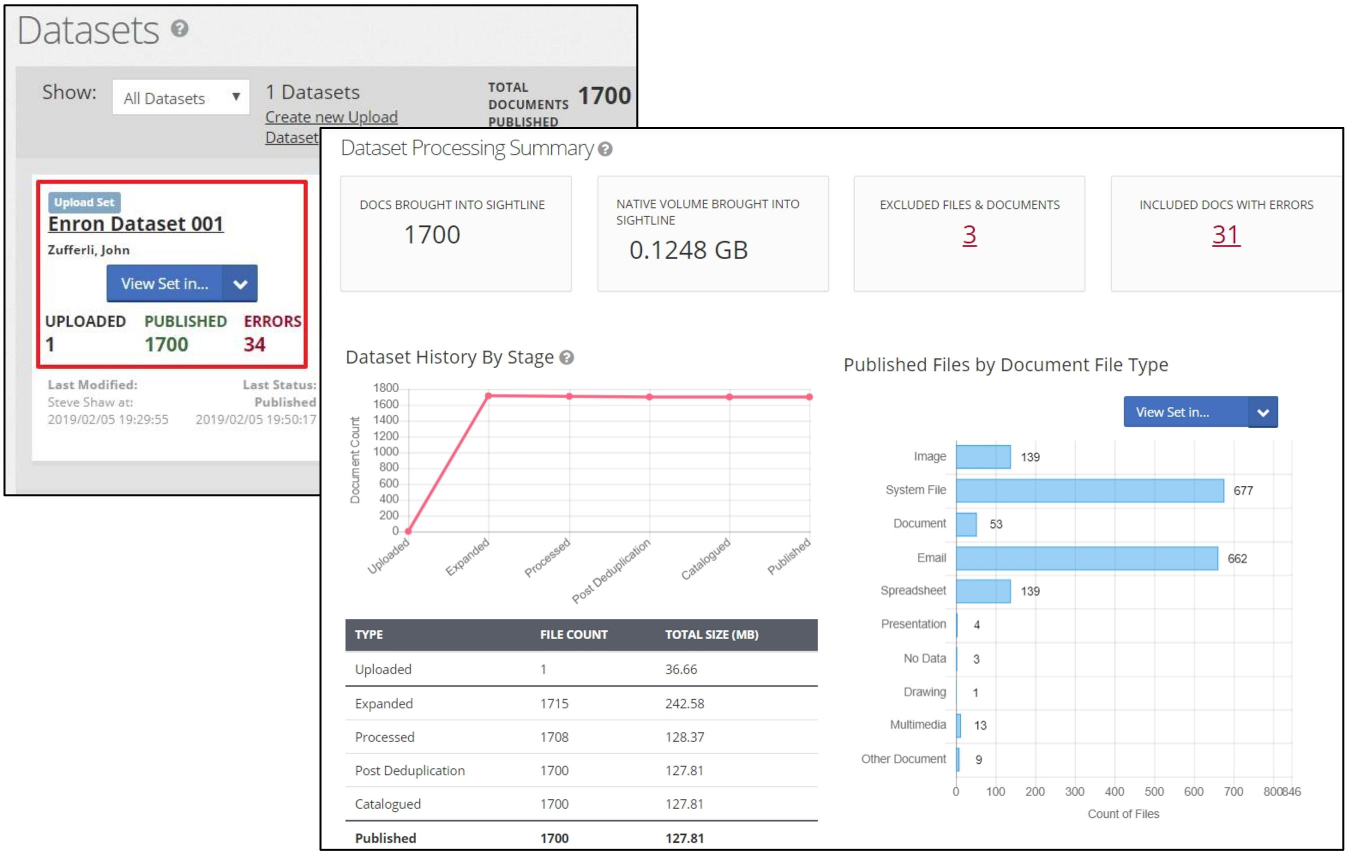
Entire Project Processing Summary

TYPE	FILE COUNT	TOTAL SIZE (MB)
Excluded	3	0.03
Duplicates	8	0.56
Processed	1708	128.37

At the bottom right of this page is a Project-level Processing Monitor which shows Type of files, the File Count of each type and the Total Size in MB for each type. This provides a high-level view of the entire project's processing.

Dataset Summary

After the Dataset has completed processing, going back to the Datasets Page will provide a quick view of the Uploaded files, the Published Files and the Errors. To get a bigger picture, click on the Dataset name to open the Dataset Summary Page.



Datasets

Show: All Datasets

1 Datasets

Create new Upload Dataset

TOTAL DOCUMENTS PUBLISHED 1700

Upload Set

Enron Dataset 001

Zufferli, John

UPLOADED	PUBLISHED	ERRORS
1	1700	34

Last Modified: Steve Shaw at: 2019/02/05 19:29:55

Last Status: Published 2019/02/05 19:50:17

Dataset Processing Summary

DOCS BROUGHT INTO SIGHTLINE: 1700

NATIVE VOLUME BROUGHT INTO SIGHTLINE: 0.1248 GB

EXCLUDED FILES & DOCUMENTS: 3

INCLUDED DOCS WITH ERRORS: 31

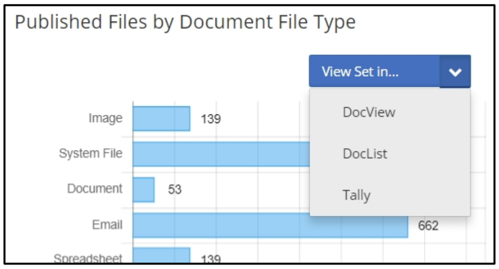
TYPE	FILE COUNT	TOTAL SIZE (MB)
Uploaded	1	36.66
Expanded	1715	242.58
Processed	1708	128.37
Post Deduplication	1700	127.81
Catalogued	1700	127.81
Published	**1700**	**127.81**

Dataset Processing Summary

The Dataset Summary page presents the following counts:

- **Docs Brought Into Sightline**: The count of unique DocIDs that made it through processing and ingestion and were brought into Sightline from this Dataset.
- **Native Volume Brought Into Sightline**: The aggregate native file sizes of all Docs Brought Into Sightline.
- **Excluded Files and Documents**: the count of files that got excluded during the processing and ingestion into Sightline because of some critical issue with the data.
- **Included Docs with Errors**: the count of DocIDs that were published into Sightline, despite having some non-critical errors

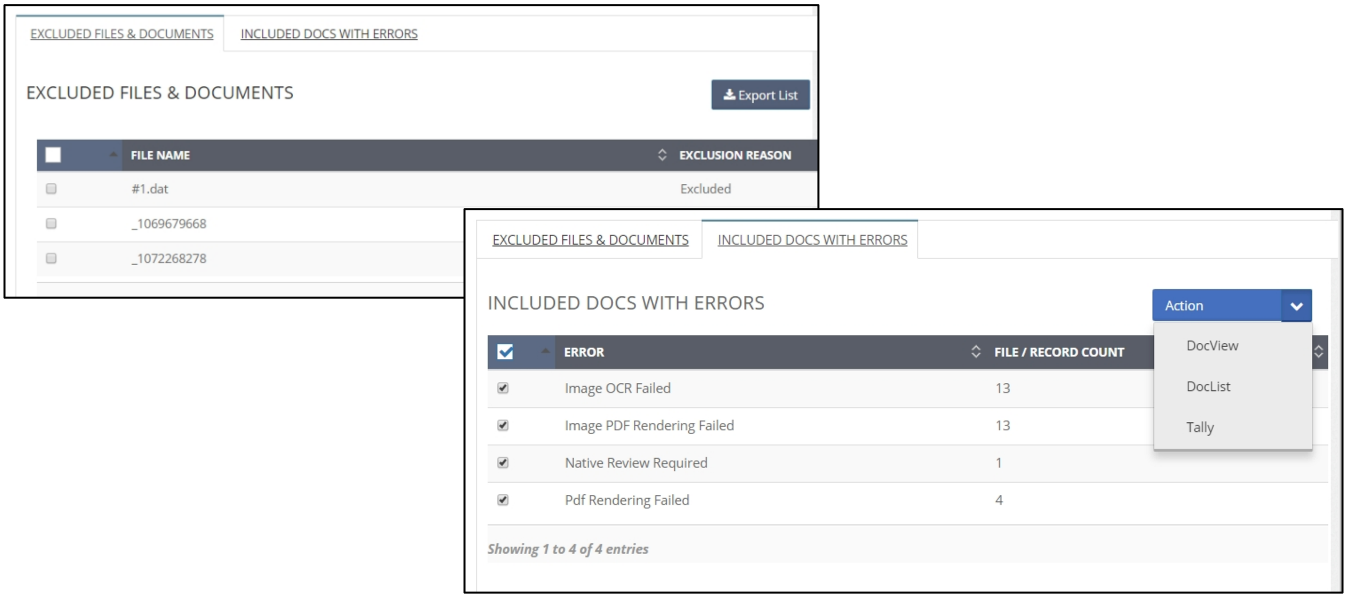
Users can also see the details of the Dataset at the bottom of this page. The lower left grid and chart shows how many files were uploaded as part of the Dataset, and how those files were expanded in processing - when child records/attachments/embedded records were extracted - and then how the full processed set was filtered through post-process filtering (ex. deduplication) and how many records were published. The lower right tally shows a breakdown of all published records in the Dataset by DocFileType attribute.



You can view these files using the "View Set In…" dropdown to view the Dataset in the DocView, the DocList or in the Tally Report. This selection will open the published files in these views and allow the User to interact with the newly loaded data.

Exceptions Page

The user can click on the count in the 'Excluded Files and Documents' to be taken to the Exception Report page to examine those files that were excluded. Similarly, the user can click on the count in the 'Included Docs with Errors' to be taken to the Exceptions Page to view those documents that were published despite encountering errors during processing and/or ingestion.


EXCLUDED FILES & DOCUMENTS
INCLUDED DOCS WITH ERRORS
Export List

EXCLUDED FILES & DOCUMENTS

☐	FILE NAME	EXCLUSION REASON
☐	#1.dat	Excluded
☐	_1069679668	
☐	_1072268278	

EXCLUDED FILES & DOCUMENTS
INCLUDED DOCS WITH ERRORS
Action
DocView
DocList
Tally

INCLUDED DOCS WITH ERRORS

☑	ERROR	FILE / RECORD COUNT
☑	Image OCR Failed	13
☑	Image PDF Rendering Failed	13
☑	Native Review Required	1
☑	Pdf Rendering Failed	4

Showing 1 to 4 of 4 entries

From the Excluded Files & Documents tab you can see a list of the Excluded files with the file name and the Exclusion Reason. Clicking the Export List button will export out this view as a spreadsheet to a local machine. These files have not been loaded into the system and cannot be viewed in Sightline.

The Included Docs with Errors allows you to see the counts by error for documents that have been loaded into Sightline but encountered an error during processing. Errors include OCR failure, PDF rendering failure, etc. Checking the box next to any, or all, of these error groups will allow the viewing of these documents in either the DocView, DocList or Tally report.