

## Introduction

This tutorial will walk you through the basics of submitting data to CRDC using the DataHub submission portal (URL HERE). If you have additional questions that aren't answered here, please contact either the DataHub Data Team member assigned to your project or contact the CRDC Helpdesk (EMAIL HERE).

## Prerequisites

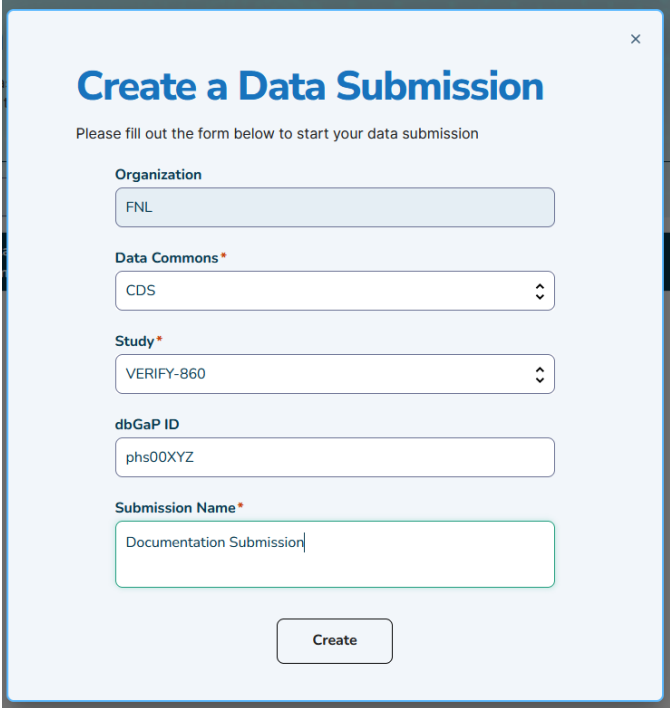
Before starting your data submission, please be sure to have completed all of the pre-requisites:

- Approval from the CRDC Data Governance Board to submit your data (application to submit can be found here URL HERE).
- A Login.gov or NIH identity. For those using Login.gov, using an identifier associated with your institution or company is preferred.
- If the study contains controlled access data, the study should be registered at dbGaP. URL HERE

Additionally, be aware that DataHub relies on CRDC standard Common Data Elements (CDEs) and submissions are expected to use these CDEs and comply with their permissible values. More information about the CRDC CDEs can be found here: URL HERE.

## Starting the submission

To start a new submission, first log in, then click on "Data Submissions" in the menu bar. The system will then display a table of all the submissions that have been started (and the table will be empty if this is the first submission). Clicking on the Create a Data Submission button in the upper right will bring up a dialog box that requests information to start submissions for the project.



**Create a Data Submission**

Please fill out the form below to start your data submission

**Organization**

FNL

**Data Commons \***

CDS

**Study \***

VERIFY-860

**dbGaP ID**

phs00XYZ

**Submission Name \***

Documentation Submission


Create

Figure 1: Creating a new data submission

The Organization box should already be populated with your organization and the Data Commons drop-down is currently restricted to CDS submissions (if you're trying to submit to another data commons, please contact the DataHub helpdesk). The Study drop-down should contain a list of the projects that you're approved to submit to. If this list is in error, please contact the data team member assigned to your project or the DataHub helpdesk. You will also need to know the dbGaP phs number if your project contains controlled data.

### Continuing an existing submission

If this is not a new submission, all of the submissions available to be worked on are visible on the table shown when you've logged in and navigated to the Data Submissions page. Simply click on the Submission Name for the project you want to work on.

 **NATIONAL CANCER INSTITUTE**  
Cancer Research Data Commons

[Return to CRDC](#) [Submission Requests](#) [Data Submissions](#) [Model Navigator](#) TODD

## Data Submission List

Below is a list of data submissions that are associated with your account. Please click on any of the data submissions to review or continue work.

[Create a Data Submission](#)

Rows per page: 10 1-1 of 1

Figure 2: Table with current active submissions

### Obtaining Submission forms

Submission to CRDC requires users to put their data into the submissions sheets that are used to validate the data. To get to the submission sheet, click on "Model Navigator" in the menu bar, and then "CDS Data Model".

## CDS Model

### Data Submission List

Below is a list of data submissions that are associated with your account.  
Please click on any of the data submissions to review or continue work.

[Create a Data Submission](#)

Organization		FNL	Status		All					
Submission Name	Submitter	Data Commons	DM Version	Organization	Study	dbGaP ID	Status	Primary Contact	Created Date	Last Updated
There are no data submissions associated with your account										
Rows per page: 10 0-0 of 0										

Figure 3: Use the Menu bar to navigate to the Data Model Viewer

This will open up the Data Model Viewer, which is where the data model can be viewed in detail and the submission forms downloaded.

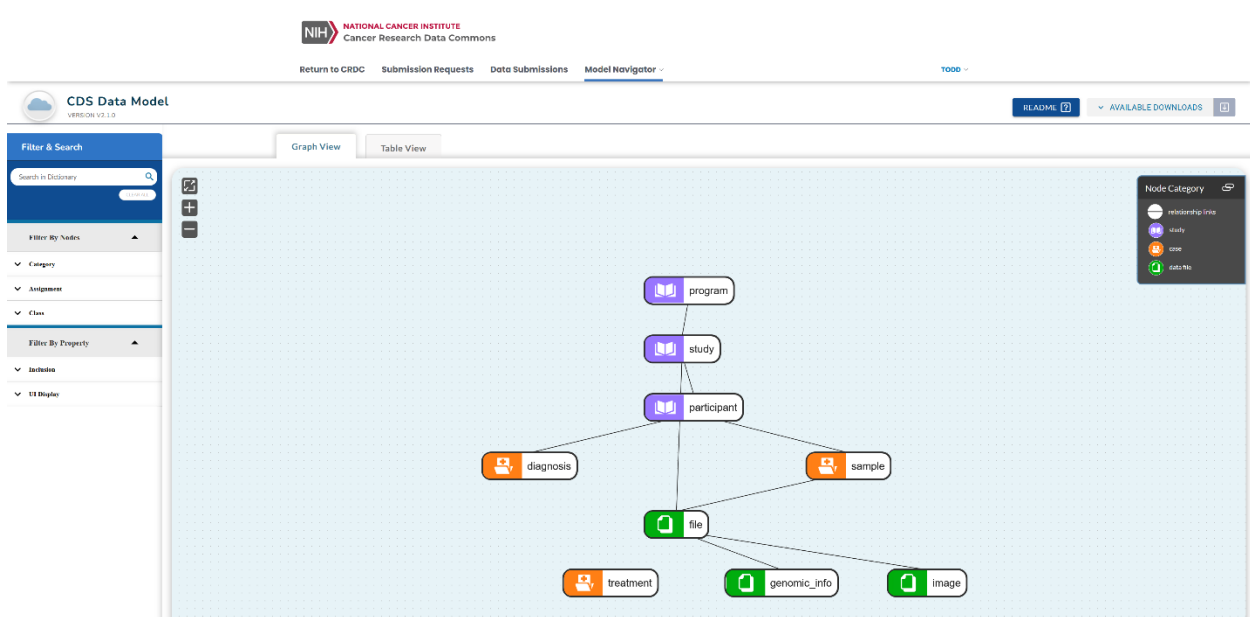


Figure 4 The Data Model Viewer Graph View

The Model Viewer can be used to explore the kinds of data that can be, or is required to be, included in a submission. Clicking on a node brings up a summary of the node and clicking on the View Properties

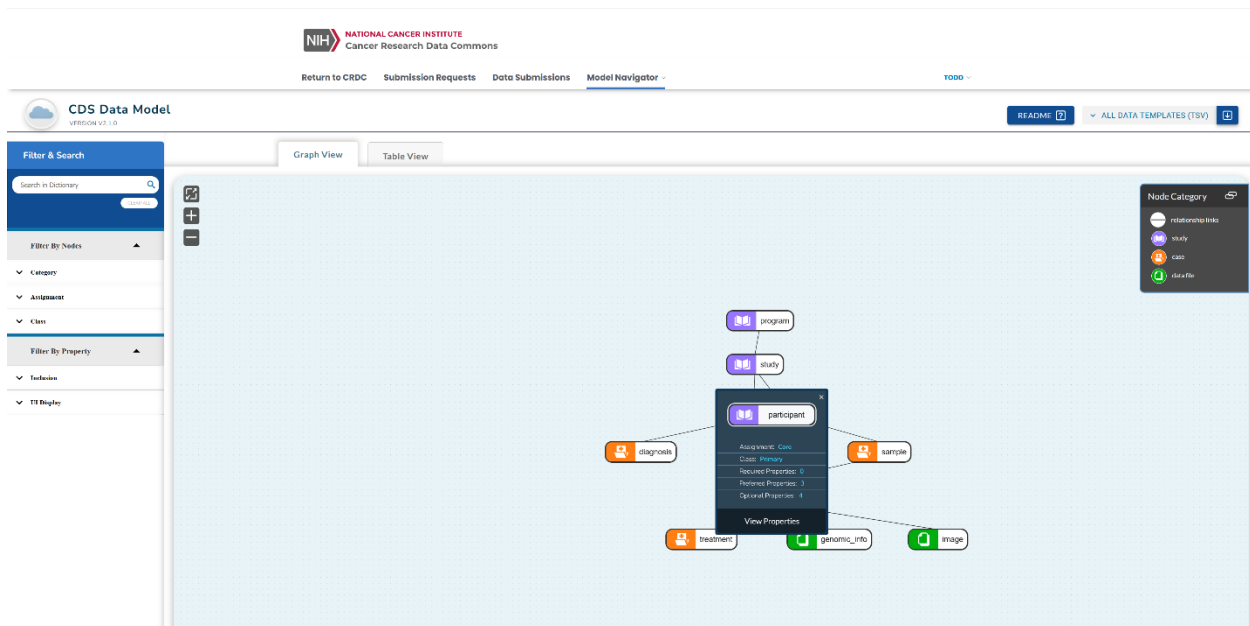


Figure 5: Clicking on a node will bring up a summary and the ability to open a table view

menu will bring up a table view of the node, including the kind of data that's expected in that field (strings, integers, etc.) and which fields are required.

Property	Type	Required	Description	Source
study_participant_id	"string"	Optional	The property study_participant_id is a compound property, combining the property participant_id and the parent property study_plus_accession. It is the ID property for the node participant. The reason why we are doing that is because in some cases, there are some participant id in different studies represent different participants.	
participant_id	"string"	Optional	A number or a string that may contain metadata information, for a participant who has taken part in the investigation or study.	
race	Acceptable Values: • White • American Indian or Alaska Native • Black or African American • Asian • Native Hawaiian or Other Pacific Islander • Unknown • Not Reported • Not Allowed to Collect	Preferred	OMB Race designator	
gender	Acceptable Values: • Female • Male • Unknown • Unspecified • Not Reported	Optional	Biological gender at birth	
ethnicity	Acceptable Values: • Hispanic or Latino • Not Hispanic or Latino • Unknown • Not Reported • Not Allowed to Collect	Preferred	OMB Ethnicity designator	
dxCaP_subject_id	"string"	Preferred	Identifier for the participant as assigned by dxCaP	
erdc_id	"string"	Optional	The erdc_id is a unique identifier that is generated by Data Hub	

Figure 6: The table view of a node

To download the submission files, click on the “Available Downloads” menu in the upper right and select “All Data Templates (TSV)”. Then click on the download arrow to start the download. Note that you can also download the full data dictionary in PDF format as well as vocabularies in either TSV or JSON format. Also, examples of completed submission templates can be downloaded by selecting the “Loading File Example”. These can be useful to understand what each of the columns in the template is supposed to contain.

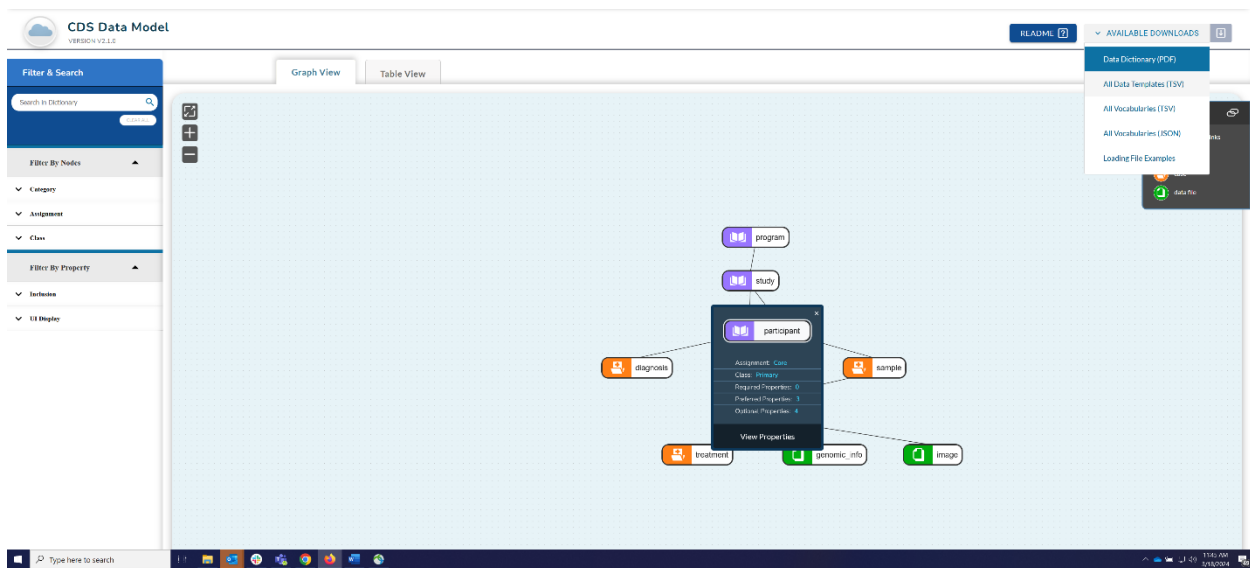


Figure 7: Using the download menu

The files will be downloaded as a .zip archive. Since these are tab-separated text files, they can be viewed in any text editor or a spreadsheet program like Microsoft Excel or OpenOffice Calc. There should be several different files in the zip archive as shown in Fig X, however the exact files may change as the data model and submission requirements change.









Name	Type
 CDS_Data_Loading_Template-diagnosis 2024-03-18 11-45-40	TSV File
 CDS_Data_Loading_Template-genomic_info 2024-03-18 11-45-40	TSV File
 CDS_Data_Loading_Template-image 2024-03-18 11-45-40	TSV File
 CDS_Data_Loading_Template-participant 2024-03-18 11-45-40	TSV File
 CDS_Data_Loading_Template-program 2024-03-18 11-45-40	TSV File
 CDS_Data_Loading_Template-sample 2024-03-18 11-45-40	TSV File
 CDS_Data_Loading_Template-study 2024-03-18 11-45-40	TSV File
 CDS_File_Transfer_Manifest 2024-03-18 11-45-40	TSV File

Figure 8: Submission templates downloaded from the DataHub Model Viewer

Each of the submission templates covers a different part of the data being submitted. Not all templates are required, only those necessary to cover the data being submitted. Let's look at the individual files to help decide which ones are needed:

- **Diagnosis** – This file should contain information related to the participant's diagnosis, including the disease(s) the participant has been diagnosed with, tumor stage information, and where the tumor was found.
- **Genomic info** – This template is used to describe the sequencing experiments done by provide information such as the reference genome, the library strategy, and the sequencing platform. This form can be ignored if the submission does not include sequencing information.
- **Image**– This is used to describe images that are included in the submission. As with the genomic information template, this should only be used when images are part of the submission and can be excluded when they are not.
- **Participant** – This template contains basic information about the participants in the submission, including identifiers used in the study
- **Sample** - This template allows for a description of the samples and an indication of which patient they are associated with. In some cases, studies may not have participants, in which case both the participant sheet and the participant ID column can be ignored.
- **Program** – While one of the shortest templates (frequently one line is sufficient), this template can be difficult as it asks for information about the program that the submitted data are associated with. Note that for CRDC purposes, a program is a high-level structure such as the Human Tumor Atlas Network (HTAN), or Childhood Cancer Data Initiative (CCDI).
- **Study** – This is for information about the study including information such as the dbGaP phs accession number. Note that studies are a child of programs.

- File – This template describes the files that are being submitted to CRDC and their relationship to the samples used in the study.

## Uploading Files and Manifests

There are two ways to move files from their local environment to DataHub for submission:

- CLI Upload Tool – This command line interface is used to transfer large data files like genomic sequence files or images into the submission area on DataHub
- Graphical interface – The DataHub interface can upload small files such as the data loading metadata templates

## CLI Upload Tool

### Introduction

DataHub provides a command line interface for uploading datasets to DataHub. It can be used on any system capable of running Python 3.6 or higher. Note that there are detailed instructions on downloading, installing, and running the CLI Tool in the README file in the GitHub repository (<https://github.com/CBIIT/crdc-datahub-cli-uploader>)

### Downloading the CLI Upload Tool

The CLI Upload tools can be downloaded either directly from the DataHub site or by cloning the GitHub repository.

#### *Download from DataHub*

Clicking on your login name in the upper right will bring up a menu that includes the ability to download the CLI Upload tool.

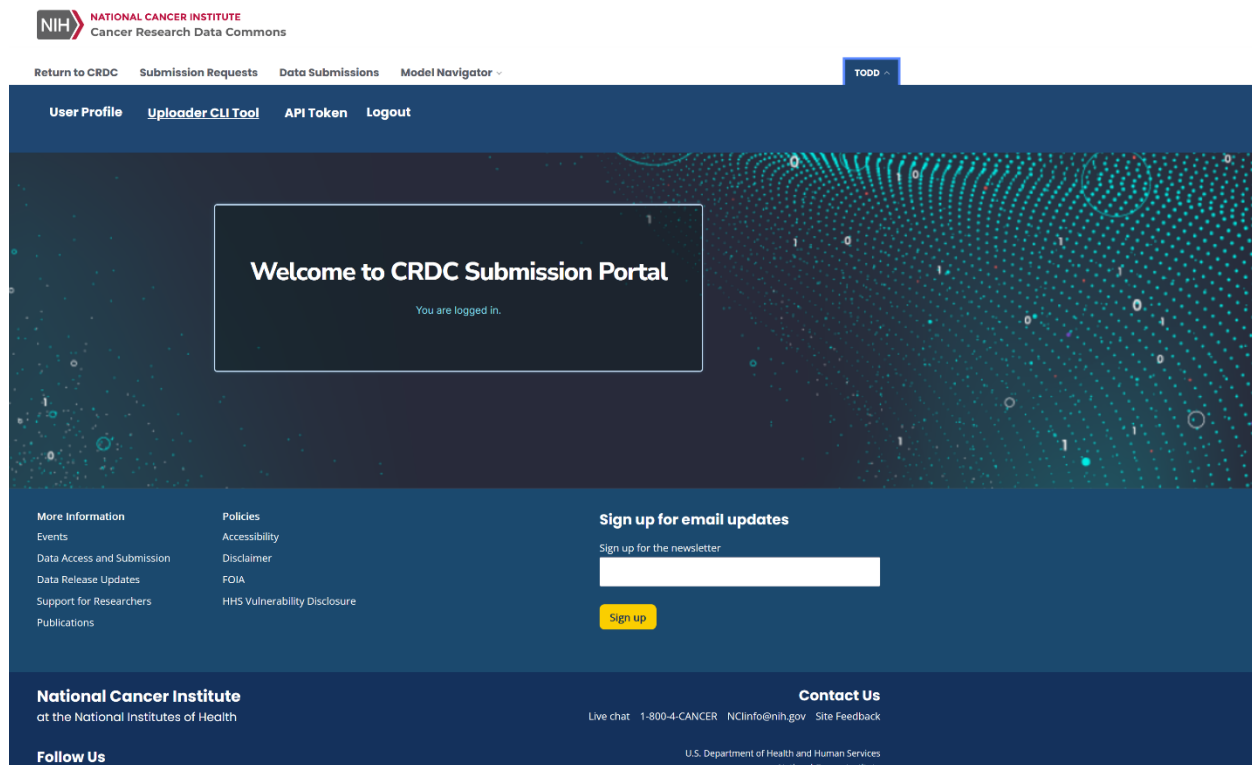
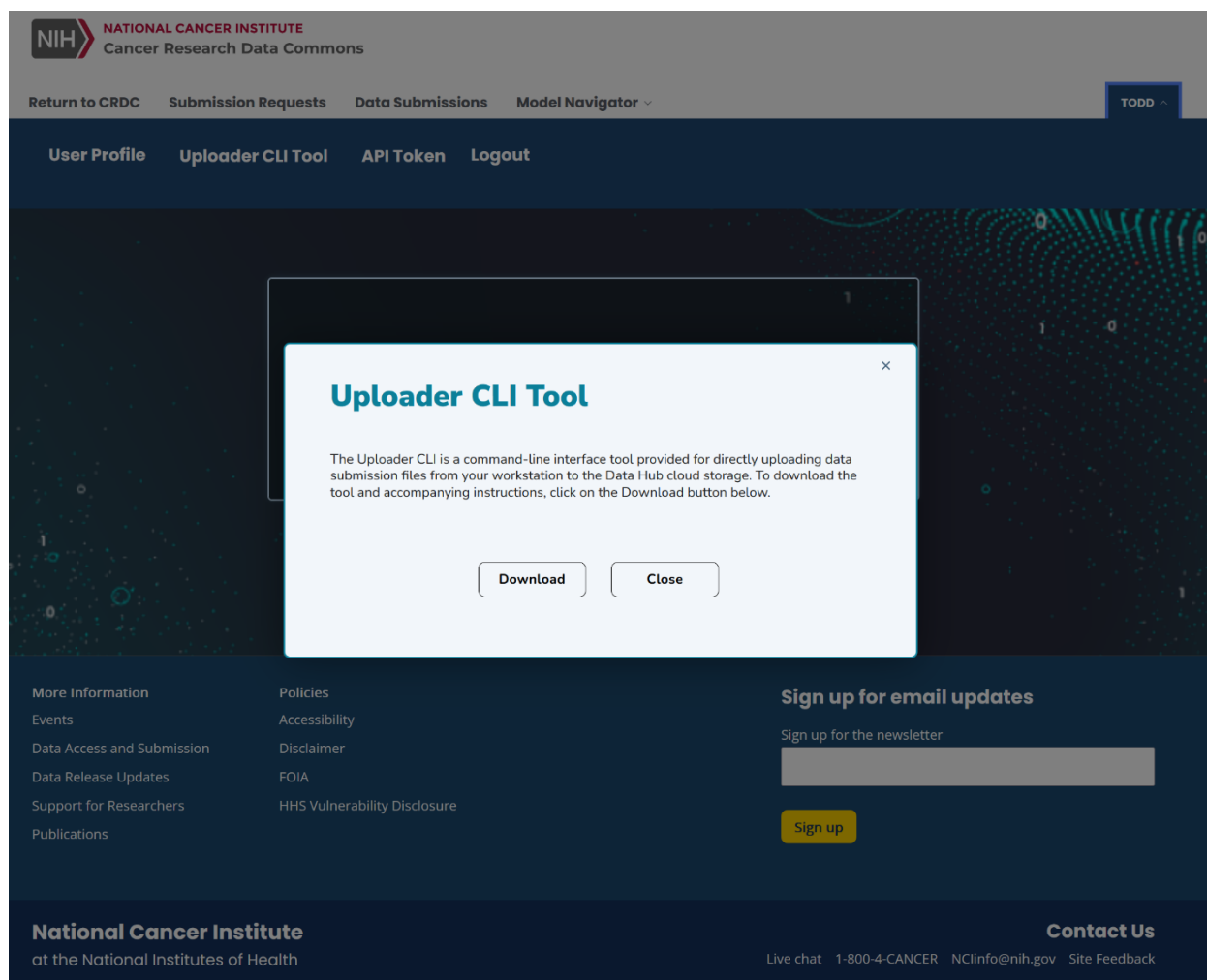


Figure 9: Menu with the CLI Tool download option

Clicking on the “Uploader CLI Tool” menu option will bring up a dialog box, simply click on Download button to download a zip archive to your local machine.





### *Cloning GitHub*

The CLI Uploader tool can also be cloned from the DataHub GitHub repository (<https://github.com/CBIIT/crdc-datahub-cli-uploader/tree/master>). Using `git clone https://github.com/CBIIT/crdc-datahub-cli-uploader.git` should clone the repository to your local machine.

### *Use of the CLI Upload Tool*

#### *CLI tool configuration file*

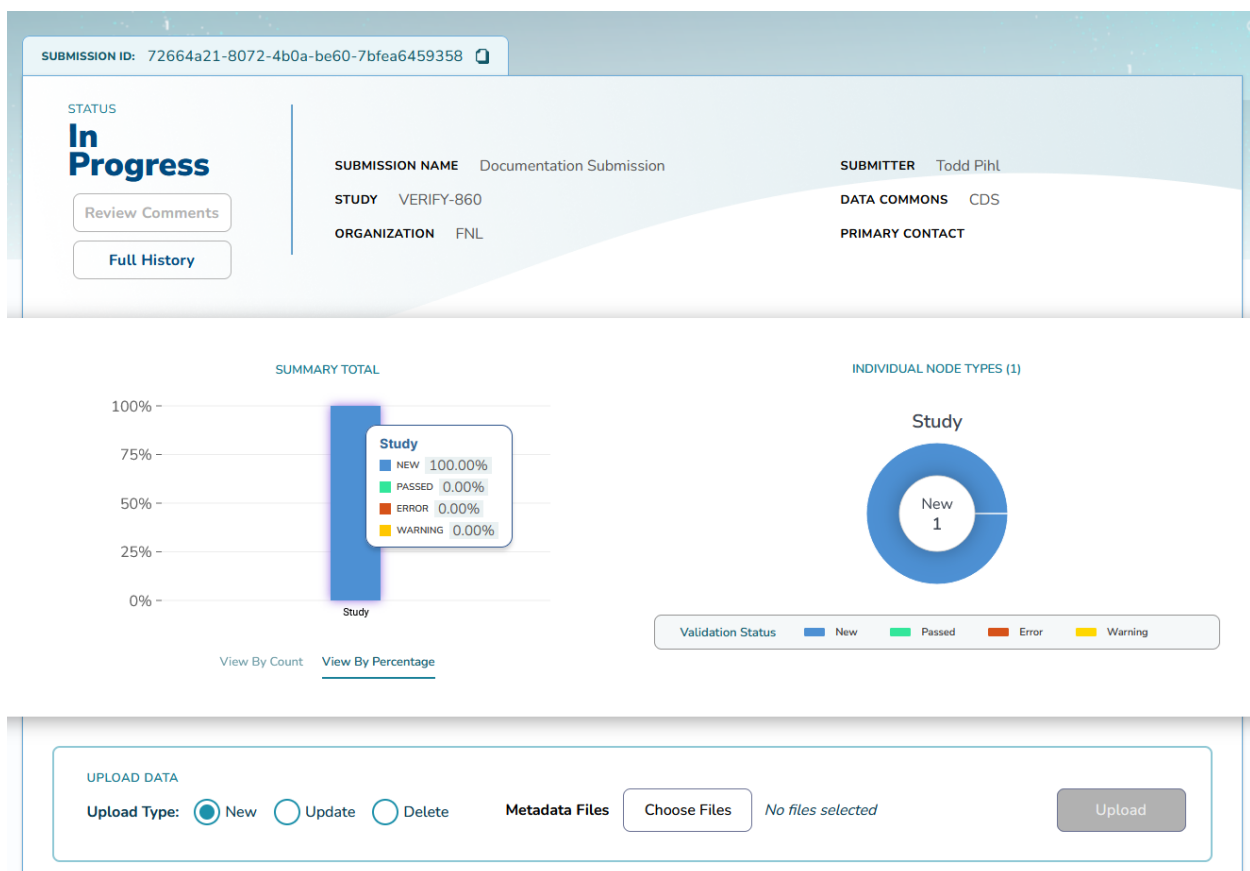
The behavior of the CLI Upload Tools is controlled by configuration files. Examples of these files can be found in the configs directory of either the extracted zip file or the GitHub cloned directories. The examples provided are the same configuration file edited for the two upload types:

- Uploader-metadata-config.example – This file is an example of using the CLI Uploader Tool to upload metadata submission templates rather than submitting them via the DataHub graphical interface.

- **Uploader-file-config.example** – This is an example of a configuration file for uploading large primary data files such as .bam files. Files uploaded this way will go through the File validation system rather than the metadata validation system.

Please note that these files are in YAML format and that the CLI Upload Tool will fail if the file is not valid YAML. It is suggested that editing these files is done using an editor that preserves YAML formatting. The fields in this file are as follows:

- **api-url**: This field provides the CLI Uploader Tools with the DataHub URL used for communication and upload.
- **token**: This is the API access token that is obtained from the DataHub interface. To obtain an API token, log into the DataHub interface, click on your username to bring up the user menu, then select “API Token”. This brings up a dialog box that allows you to create and copy an API token to your clipboard.
- **submission**: This is the Submission ID that identifies which project that the uploaded files will be associated with. The Submission ID can be found in the DataHub graphical interface. Log into the system and navigate to the submission that the upload will be associated with. The Submission ID can be copied from the upper left corner of the interface.



- **type**: This tells the system if this is a metadata upload or a data file upload. Enter the term “metadata” if the upload contains submission templates and “file” if the upload contains data files.

- **data:** This is the local path to the directory that contains the files to be uploaded.
- **manifest** (*Data file upload only*): This is the local path to the manifest file.
- **name-field:** column name in the manifest file that contains file names.
- **size-field** (*Data file upload only*): column name in the manifest file that contains file sizes.
- **md5-field** (*Data file upload only*): column name in the manifest file that contains file MD5 checksums
- **intention** (*Metadata uploads only*): Valid values are **new**, **update**, and **delete**. When to use these values is described below in the section on using the DataHub interface.
- **retries:** number of retries the CLI Upload Tool will perform after a failed upload
- **overwrite:** if set to “true”, CLI Upload Tool will overwrite the file with same name that already exists in the Data Hub target storage. If set to “false”, the CLI Upload tool will not upload if a file with the same name exists in the Data Hub target storage.
- **dryrun:** if set to “true”, CLI will not upload any files to the Data Hub target storage. If set to “false”, CLI will upload files to the Data Hub target storage.

### Starting the upload process

Once the configuration file has been edited, the upload script can be started. The only required parameter is “--config” which should provide the path to the proper configuration file. The command should look something like the following, though the exact details may vary depending on how the tool (and Python) were installed:

```
$ python3 src/uploader.py --config configs/metadata-upload.yml
```

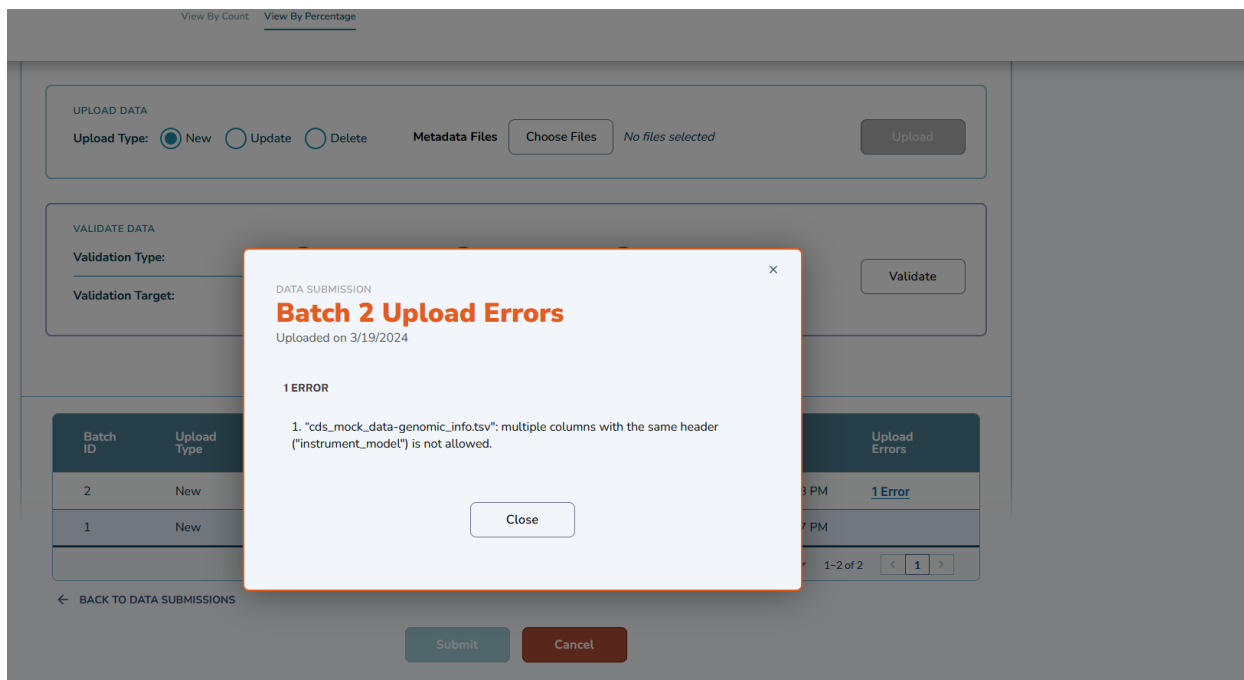
### Using the DataHub interface to upload metadata sheets

The upload feature in the DataHub interface is intended for submitting the completed metadata submission templates to DataHub for submission. To start the process, select what kind of upload this will be:

- **New** – Use this if this is the first time the submission templates have been uploaded.
- **Update** – This should be used if the submission templates have been uploaded before and this template contains corrections to the previous upload. This is used when correcting errors.
- **Delete** – This function is used when data needs to be removed from the submission, but no replacement data will be provided. An example of when to use this function would be when a participant was mistakenly included in an earlier upload and needs to be removed.

Submitters should keep their new information and any subsequent updates separated. For example, if a study has 100 participants, the submitted template could either contain all 100, or it could contain a subset of that 100, with the remainder submitted in later uploads. As long as there is no overlap in participants between the different uploads, each upload would be a **New** upload. However, mixing new data from previously uploaded participants with new participants will result in an error as the system knows about the previously uploaded participants. If corrections need to be made, the **Update** submission should only contain previously submitted participants, even if the data associated with them is new.





## Running validations

Validations can be run at any point in the submission process, there are no restrictions on when, or how often, validations can be run. All validations are run by selection options in the Validation Panel and clicking on the Validate button.

VALIDATE DATA

Validation Type: ☒ Validate Metadata ☐ Validate Data Files ☐ Both

Validation Target: ☒ New Uploaded Data ☐ All Uploaded Data

Validate

The first step will be to select what files are to be validated. The “Validate Metadata” option will run validations only on the submission metadata templates, and not on any of the uploaded data files. The Validate Data Files option will do the reverse and check all of the uploaded data files. The Both option will validate both.

By default, only newly uploaded files will be validated. This can be a significant time saver for large submissions as some validations can take considerable time and the system has a record of any previously submitted files that have already passed validation. However, if there is a need to check the entire submission, regardless of previous validation runs, the All Uploaded Data option will check everything that has been uploaded so far.

## Reviewing validation results

After validations are run, the graphics on the page are updated to give a summary of the results

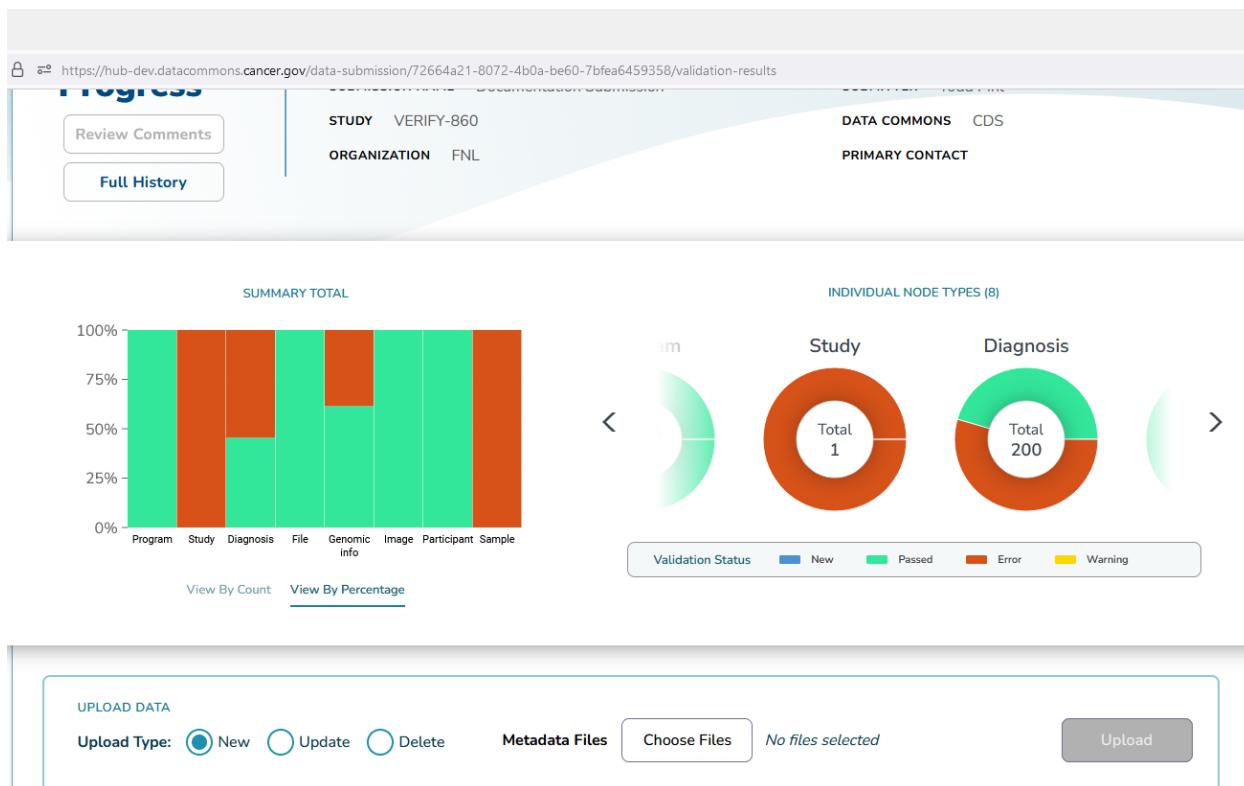


Figure 11: Validation Summary

The left graphics displays a list of the nodes that have been validated and how much of the submitted data has either passed (green) or failed (red). Hovering over each bar will generate a more detailed summary for that node. The graphs on the right are a node-by-node description of the results with the left and right arrows moving between the nodes that have been submitted to date.

All errors and warnings are detailed on the Validation Results tab of the table.

validation target:

New Uploaded Data

All Uploaded Data

Data Activity

Validation Results

Node Type

All

Batch ID

All

Severity

All

Batch ID	Node Type	Submitted Identifier	Severity	Validated Date	Issues
7	Sample	CDS1011_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted <a href="#">See details</a>
7	Sample	CDS1008_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted <a href="#">See details</a>
7	Sample	CDS1010_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted <a href="#">See details</a>
7	Sample	CDS1014_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted <a href="#">See details</a>
7	Sample	CDS1004_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted <a href="#">See details</a>
7	Sample	CDS1013_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted <a href="#">See details</a>
7	Sample	CDS1009_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted <a href="#">See details</a>

This table documents all of the errors that were found after the validations were run. The information in the columns can be interpreted as follows:

- Batch ID – This correlates with the Batch ID shown on the Data Activity tab and indicates which specific upload the error is associated with. This helps to identify which files may be involved
- Node Type – This correlates to the different metadata submission sheets. In the example above, the Node Type of Sample indicates that the error lies in the sample metadata template.
- Submitted Identifier – This is the identifier supplied by the project, it is not a DataHub identifier. Again, this should specifically identify what object is causing the error.
- Severity – Severity will either be Error (which must be corrected before the submission can be finalized) or Warning (which should be fixed, but are not required to be fixed)
- Validated Date – The date that the validation was run
- Issues – This gives a brief description of the error and a link to bring up a dialog box with more details about the error.

The screenshot shows a web interface with a 'Validation Results' tab. A modal window displays a validation error for a specific node. The error message states that 'Archer Fusion' is not a permissible value for the 'library\_strategy' property. The background table shows a list of nodes, with the last row indicating an error for the same reason.

Batch ID	Node Type	Node ID	Severity	Timestamp	Issue
3	Genomic_info				
3	Genomic_info				
3	Genomic_info				
3	Genomic_info	dg.4DFC/840c7e9c-8900-4d1d-96667ce99c1e_142	Error	03-20-2024 at 11:25 AM	Value not permitted <a href="#">See details</a>

## Correcting errors

Errors should be corrected by addressing the issues in local files, re-uploading the corrected file, and running the validation again. This process should be repeated until all errors have been addressed and the validation returns no Errors.

Anything marked as an Error in the Severity table has to be fixed before the dataset can be formally submitted. Anything marked as a Warning will not block the final submission, however submitters are **strongly encouraged** to fix Warnings as well.

## Submitting your final dataset

When a dataset has passed all validations with no outstanding Errors, the Submit button at the bottom of the page will be activated. Clicking on this Submit button locks the submission and passes control to the DataHub Data Team for a final check. No further changes will be allowed. Should the Submit button be clicked in error, please contact the DataHub team and they can reject the submission and return it to your control.

## What to expect after submission

Once the final dataset has been submitted, the DataHub data team will perform some final checks to make sure everything is as needed to pass onto the destination Data Commons. If those checks pass, the submission will be released to the appropriate Data Commons and you will receive notification that the release has taken place. After the release, the Data Commons will be responsible for the timing of the release into their systems.

If the final checks reveal some unexpected issues, the DataHub data team will reach out with additional questions and may re-open the submission to allow additional corrections.