Data Submission

Step-by-Step Guide to Submitting Data through the CRDC Submission Portal

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

datacommons.cancer.gov

Table of Contents

I. Introduction . . . 3

II. Prerequisites . . . 3

Iii. Data Submission Portal Data Models . . . 3

IV. Starting a New Submission . . . 4

V. Continuing an Existing Submission . . . 5

1. Obtaining Submission Forms . . . 5

2. Downloading Submission Files . . . 8

2.1 Submission File Templates and Fields . . . 8

2.1.1 Special Columns . . . 9

2.1.2 Type Column . . . 9

2.1.3 Relationship Columns (Parent Mapping Columns) . . . 9

3. Uploading Files and Manifests . . . 10

3.1 Uploader CLI Tool . . . 10

3.1.1 Introduction . . . 10

3.2 Downloading the Uploader CLI Tool . . . 10

3.2.1 Download the Uploader CLI Tool from the Data Submission Portal . . . 10

3.2.2 Cloning the Uploader CLI Tool from GitHub . . . 11

3.3 Setting Up the Python Environment . . . 12

3.4 Using the Uploader CLI Tool . . . 12

3.4.1 Uploader CLI Tool Configuration File . . . 12

3.4.2 File Manifest . . . 12

3.5 Starting the Upload Process . . . 13

3.6 Using the Data Submission Portal's Graphical Interface to Upload Metadata Forms . . . 13

4. Running Validations . . . 16

4.1 Reviewing Validation Results . . . 16

4.2 Correcting Errors . . . 18

5. Submitting Your Final Dataset . . . 18

6. What to Expect After Submission . . . 18

I. Introduction

This tutorial walks you through the basics of submitting data to CRDC using the CRDC Submission Portal. If you have questions that are not answered here, please contact either the Data Submission Team member assigned to your project or the CRDC Helpdesk.

II. Prerequisites

Before starting your data submission, complete the following prerequisites:

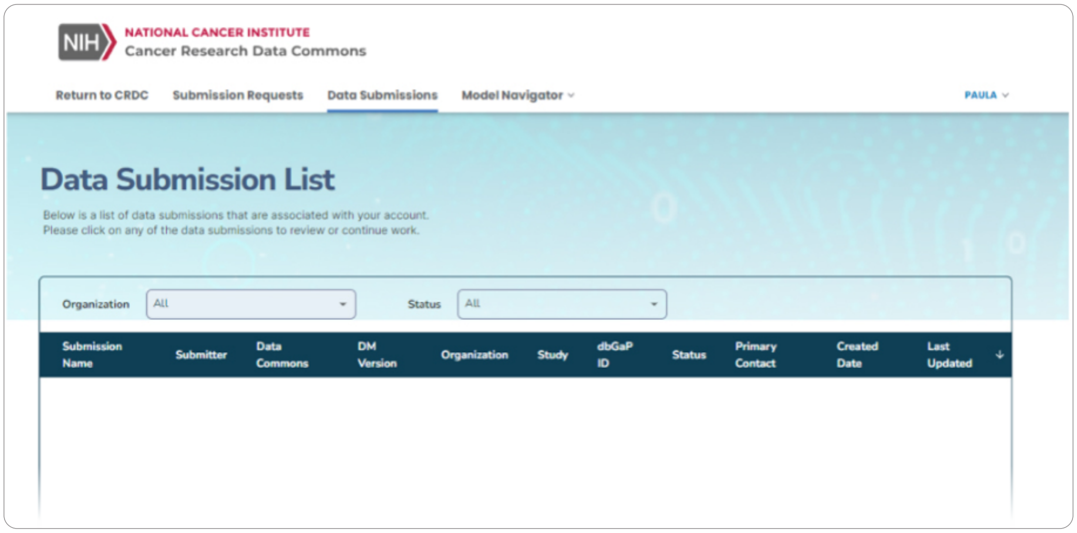
- Secure approval from the CRDC Submission Review Committee to submit your data. Approval notification will be found on the portal under Submission Request.
- Receive a Login.gov or an NIH identity that uses a PIV card (for NIH staff only). For those using Login.gov, using an identifier associated with your institution or company is preferred.
- If the study contains controlled access data, you must register the study at dbGaP.
- Additionally, be aware that the CRDC Submission Review Committee relies on CRDC standard Common Data Elements (CDEs), and submissions are expected to use these CDEs and comply with their permissible values. A comprehensive list of CRDC standard CDEs can be found at caDSR. Click the “CRDC Standard Data Elements” link in the Links to Favorites section or download them from the getCRDCList endpoint of the caDSR API.

III. Data Submission Portal Data Models

Data Submission Portal data models are graph-based and data are organized as nodes and relationships. Nodes contain properties and can have relationships with other nodes. Nodes are equivalent to tables in a relational model, and a property is equivalent to a column in a relational model. Relationships serve a similar purpose as foreign keys in a relational model.

IV. Starting a New Submission

To start a new submission, log in to the CRDC Submission Portal and then click **Data Submissions** in the menu bar. A table of all the submissions that have been started appears. The table will be empty if this is the first submission.



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.

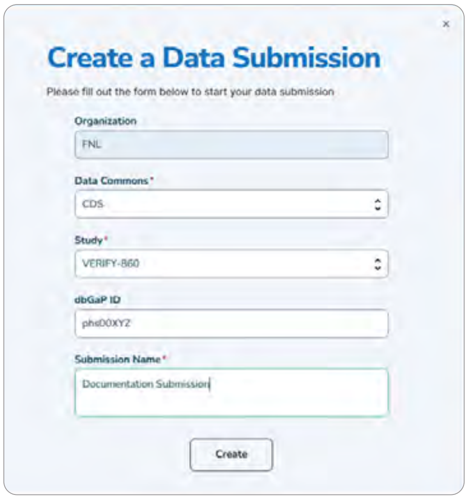
Organization: All

Status: All

Submission Name	Submitter	Data Commons	DM Version	Organization	Study	dbGaP ID	Status	Primary Contact	Created Date	Last Updated

Figure 1. Empty Data Submission List

Click the **Create a Data Submission** button. The Create a Data Submission dialog box appears.



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

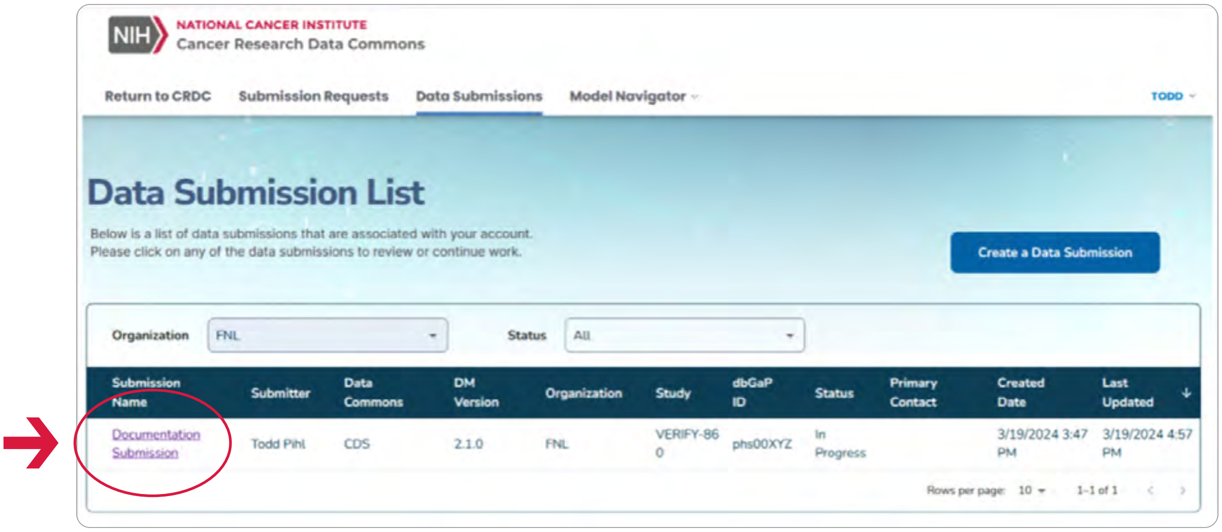
Submission Name*: Documentation Submission

Figure 2. Create a Data Submission Dialog Box

The Organization box should already be populated with your organization. The Data Commons list is currently restricted to what was formerly known as CDS submissions. If you are trying to submit to another CRDC data commons not already listed, contact your assigned team member, or email the CRDC Helpdesk. The Study list contains a list of the projects approved for your submission. If you notice an error in this list, contact the team member assigned to your project or email the CRDC Helpdesk. You will also need to know the dbGaP (database of Genotypes and Phenotypes) accession number if your project contains controlled access data.

V. Continuing an Existing Submission

In the table on the Data Submissions page, click the Submission Name for the project you want to continue working on.



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.

Organization: FNL　Status: All

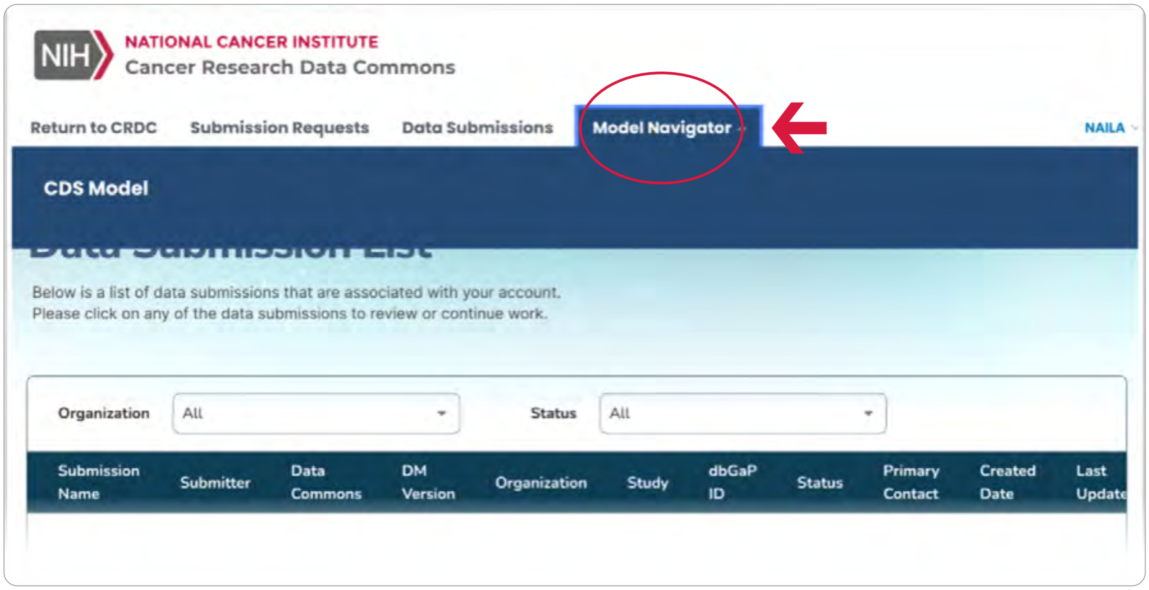
Submission Name	Submitter	Data Commons	DM Version	Organization	Study	dbGaP ID	Status	Primary Contact	Created Date	Last Updated
Documentation Submission	Todd Pihl	CDS	2.1.0	FNL	VERIFY-86 0	phs00XYZ	In Progress		3/19/2024 3:47 PM	3/19/2024 4:57 PM

Rows per page: 10　1–1 of 1

Figure 3: Table with Current Active Submissions

1. OBTAINING SUBMISSION FORMS

Submitting data to CRDC requires users to put their data into the submissions formssheets that are used to validate the data. To get to the submission formsheet, click **Model Navigator** in the menu bar (see Figure 3) and then **CDS Data Model.**



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.

Organization: All　Status: All

Submission Name	Submitter	Data Commons	DM Version	Organization	Study	dbGaP ID	Status	Primary Contact	Created Date	Last Update

Figure 4. Use the Menu Bar to Navigate to the Data Model Viewer

Once you click on Data Model Viewer, the screen below appears. This is where you can view the data model in detail and download the submission forms. in detail and download the submission forms.

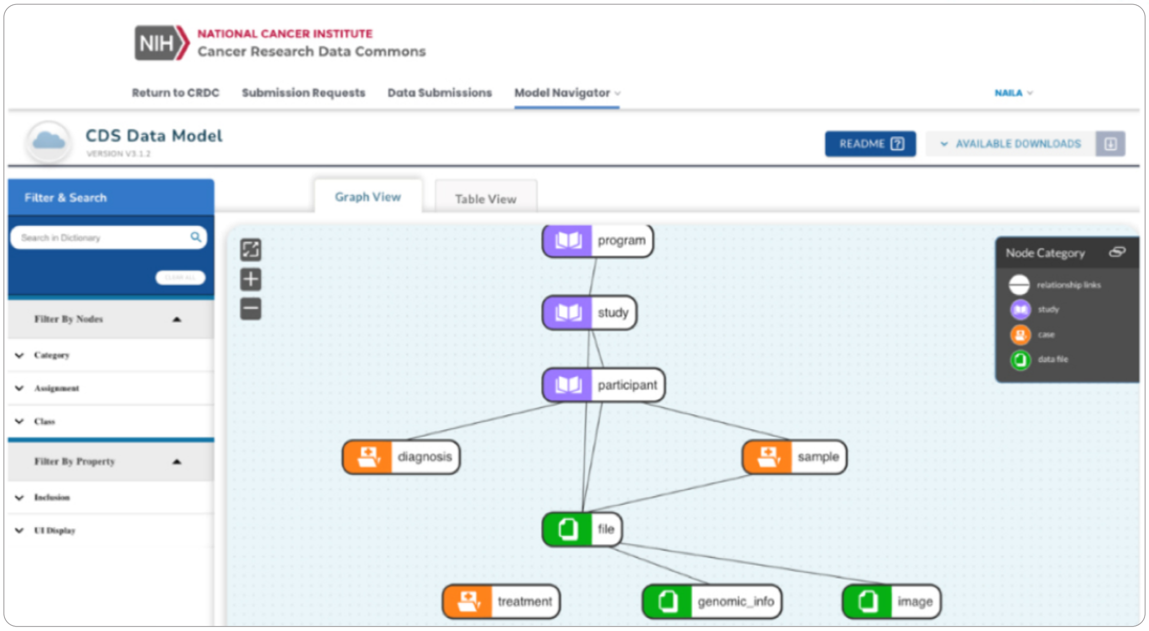
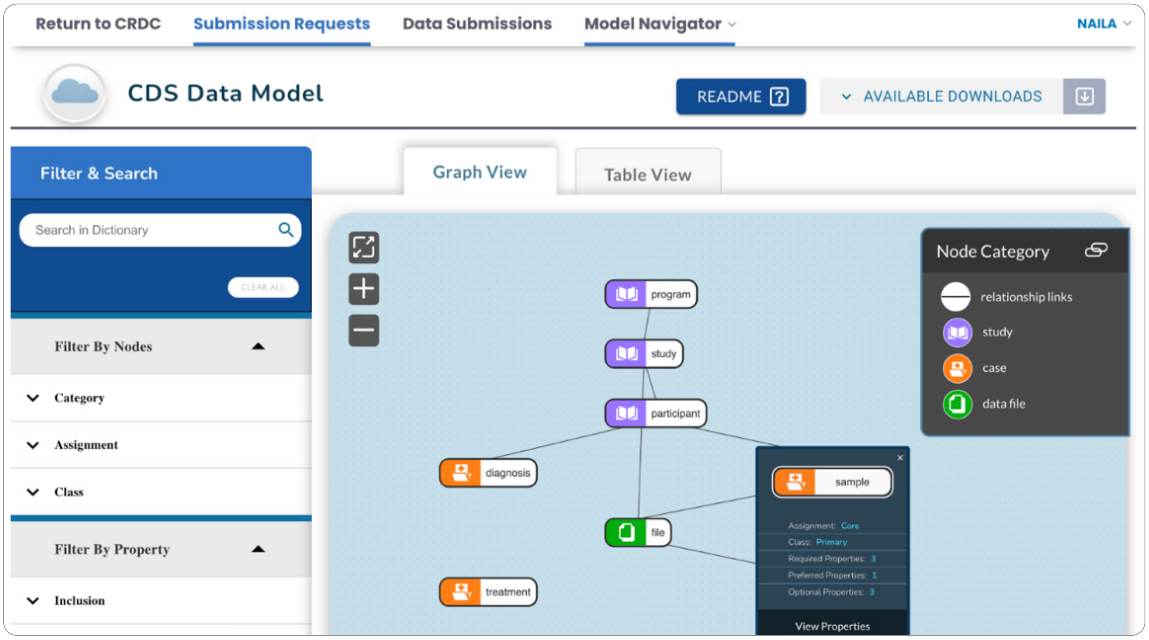


Figure 5. Data Model Viewer Graph View

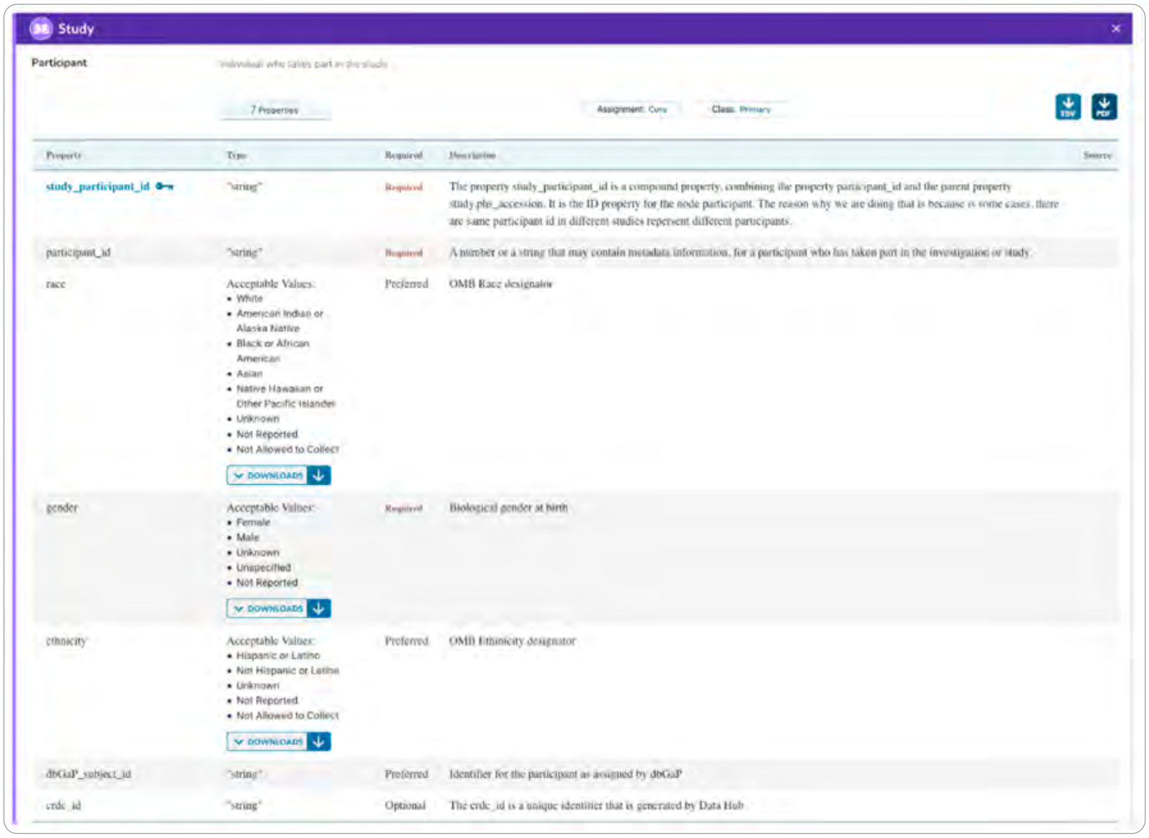
Use the Model Viewer to explore the data that submissions require or can accept. Click a node to open its summary.



Assignment: Core
Class: Primary
Required Properties: 3
Preferred Properties: 1
Optional Properties: 3

Figure 6. Click a Node to View a Summary and Open the Table View

Click the View Properties menu to open a table view of the node, including the kind of data that is expected in that field (such as strings and integers) and which fields are required.



Participant Individual who takes part in the study

Property	Type	Required	Description	Source
study_participant_id	"string"	Required	The property study_participant_id is a compound property, combining the property participant_id and the parent property study.phs_accession. It is the ID property for the node participant. The reason why we are doing that is because in some cases, there are same participant id in different studies repersent different participants.	
participant_id	"string"	Required	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 DOWNLOADS	Preferred	OMB Race designator	
gender	Acceptable Values: • Female • Male • Unknown • Unspecified • Not Reported DOWNLOADS	Required	Biological gender at birth	
ethnicity	Acceptable Values: • Hispanic or Latino • Not Hispanic or Latino • Unknown • Not Reported • Not Allowed to Collect DOWNLOADS	Preferred	OMB Ethnicity designator	
dbGaP_subject_id	"string"	Preferred	Identifier for the participant as assigned by dbGaP	
crdc_id	"string"	Optional	The crdc_id is a unique identifier that is generated by Data Hub	

Figure 7. Table View of a Node

2. DOWNLOADING SUBMISSION FILES

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

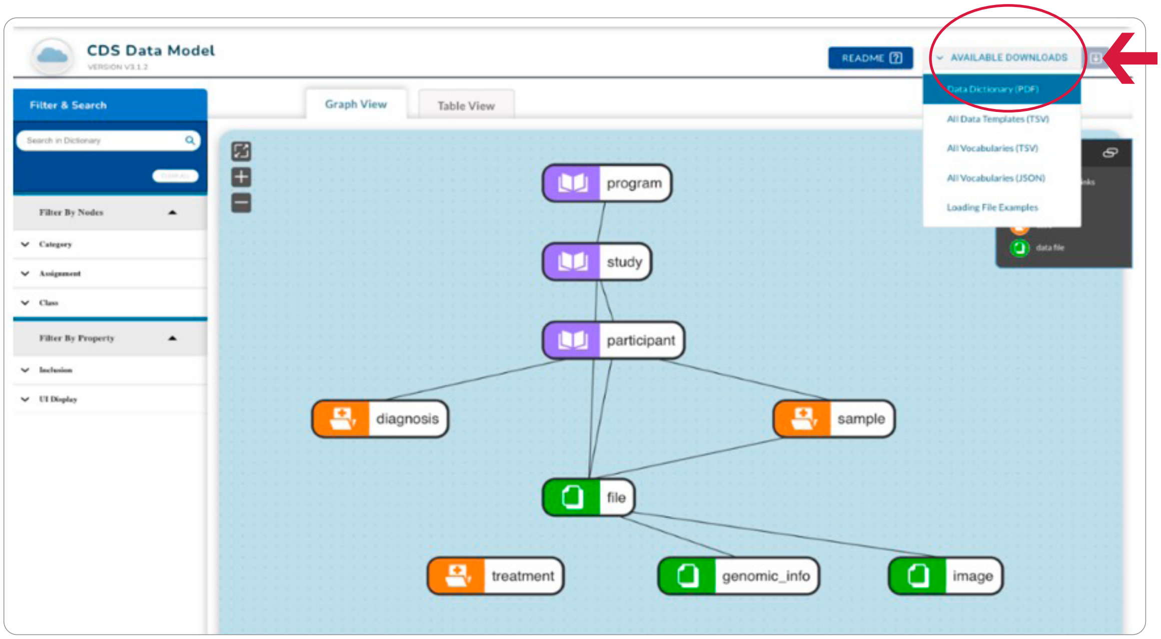
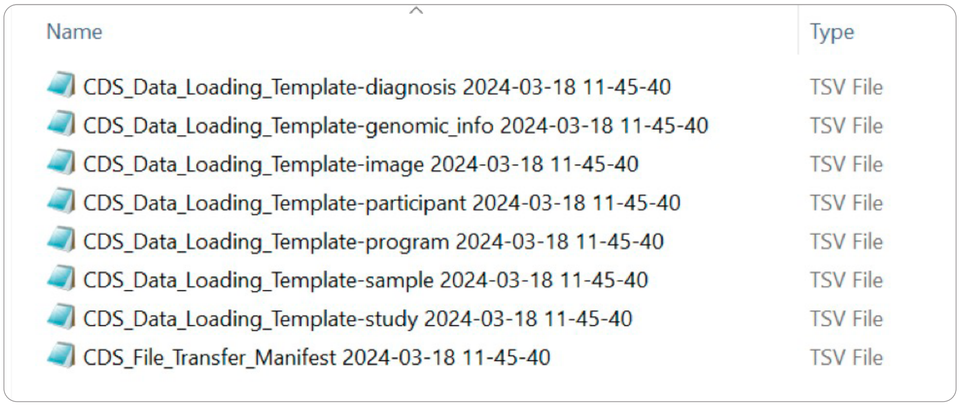


Figure 8. Using the Download Menu

The files are 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. Several files should appear in the zip archive as shown in Figure 8; 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 9. Submission Templates Downloaded from the Data Submission Portal Model Viewer

2.1 Submission File Templates and Fields

Each of the submission templates covers a different part of the data being submitted. Not all templates are required, rather only those necessary to cover the data being submitted. However, each individual sheet has required fields, so for every sheet that will be submitted, it is advisable to review the data dictionary (available from the Model Navigator) to understand what fields are required. Note that the first row of each template should NOT be edited as it contains the property names and other special columns.

2.1.1 Special Columns

Two types of special columns, also called parent mapping columns, exist: "type" and "relationship."

2.1.2 Type Column

A "type" column contains the node's name (aka node type) and is required in all template files. In the downloaded template, the second row is prefilled with the correct node name for that specific template (e.g., study). All rows should contain the same node name in the "type" column. Mixing multiple node types in one file is not supported.

2.1.3 Relationship Columns (Parent Mapping Columns)

A relationship column is used to specify relationships between the current node and its related (parent) nodes. A relationship column has a header in the form of "<parent node name>.<parent ID property name>." Values in the relationship columns are IDs of the related nodes (like a foreign key in a relational model).

For example, "study.study_id" column indicates the current node can have a "study" node as its parent node, and the property used to identify the study node is "study_id". Each value in this column is a study_id that identifies a "study" node.

Below is a brief description of the files that can be used:

- **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 details of the sequencing such as 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 with which the submitted data are associated. Note that for CRDC purposes, certain programs are defined at 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 ID. Note that studies are a "child" (subset) of programs.
- **File** – This template describes the files that are being submitted to CRDC and their relationship to the samples used in the study.

3. UPLOADING FILES AND MANIFESTS

You can move files from their local environment to the DataHub for submission in the following two ways:

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

Note: Send primary data files only using the Uploader CLI Tool. Don't attempt to upload data files using the Data Submission Portal's graphical interface.

3.1 Uploader CLI Tool

3.1.1 Introduction

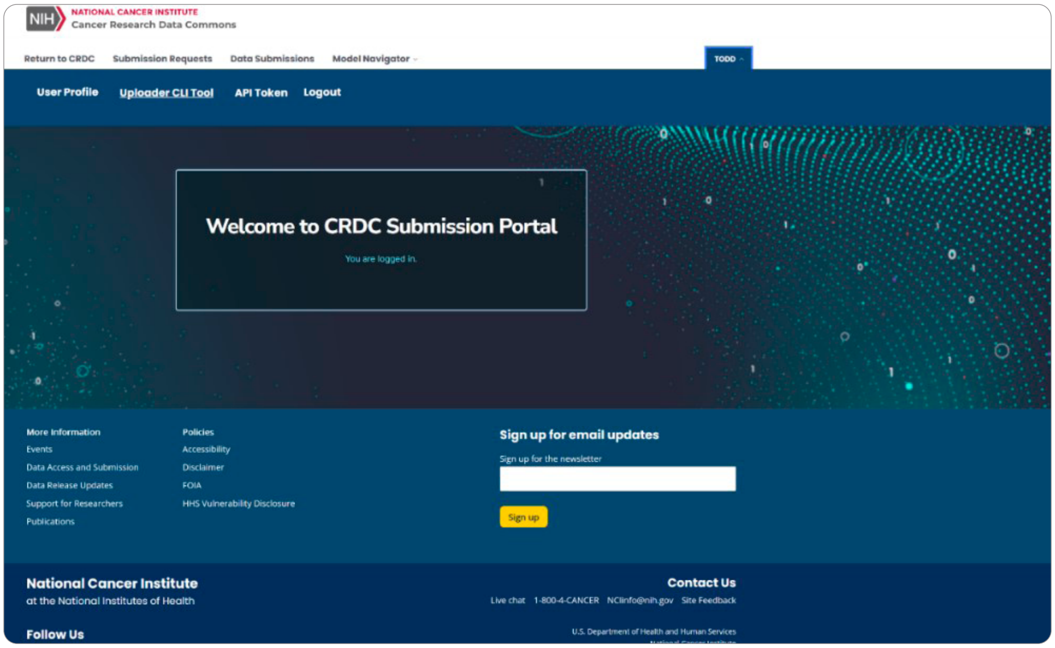
The Data Submission Portal provides a command-line interface for uploading datasets to temporary target storage. It can be used on any system capable of running Python 3.6 or higher. Note that there are detailed instructions on downloading, installing, and running the Uploader CLI Tool in the README file of the GitHub repository. Also note that the Uploader CLI Tool does not have to be downloaded for each submission; this is a Python script that can be used for any upload to the Data Submission Portal. The only aspect that must be tailored to each submission is the configuration file, which is discussed below.

3.2 Downloading the Uploader CLI Tool

The Uploader CLI Tool can be downloaded either directly from the Data Submission Portal or by cloning the GitHub repository.

3.2.1 Download the Uploader CLI Tool from the Data Submission Portal

Click your user profile name in the upper-right and then select **Uploader CLI Tool** from the menu.



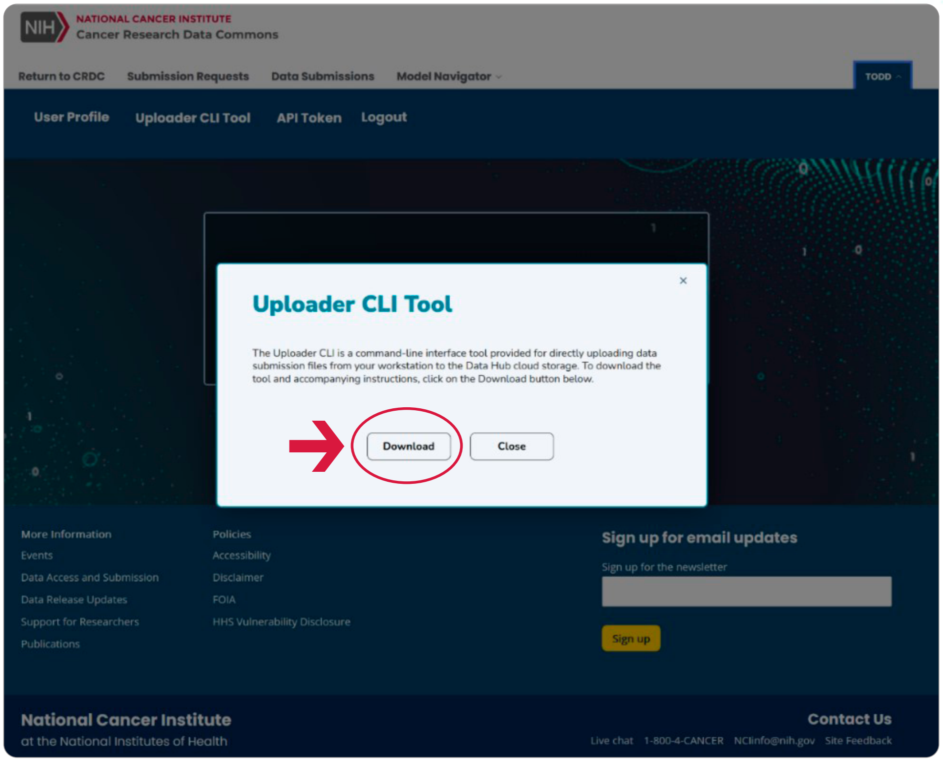
Welcome to CRDC Submission Portal

You are logged in.

More Information
Events
Data Access and Submission
Data Release Updates
Support for Researchers
Publications
Policies
Accessibility
Disclaimer
FOIA
HHS Vulnerability Disclosure
Sign up for email updates
Sign up for the newsletter
Sign up
National Cancer Institute
at the National Institutes of Health
Contact Us
Live chat
1-800-4-CANCER
NCIinfo@nih.gov
Site Feedback
Follow Us
U.S. Department of Health and Human Services

Figure 10. Menu with the Uploader CLI Tool Download Option

Click the **Download** button. A zip archive downloads to your local machine.



Uploader CLI Tool

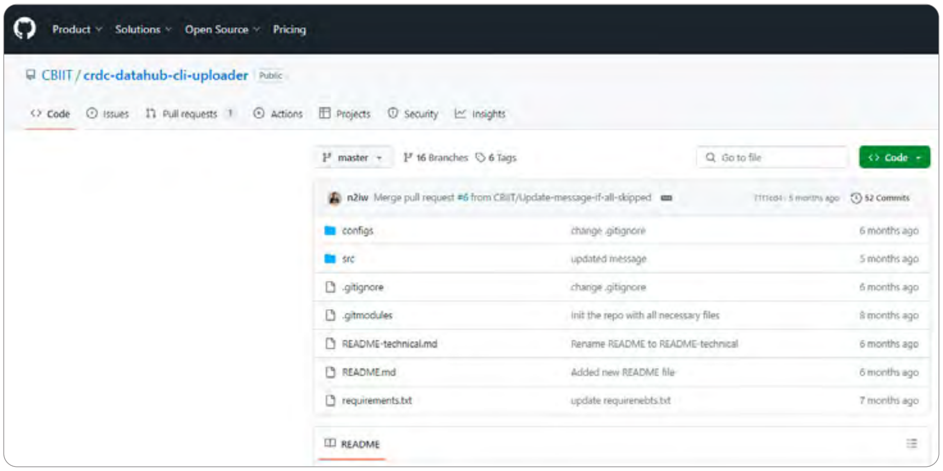
The Uploader CLI is a command-line interface tool provided for directly uploading data submission files from your workstation to the Data Hub cloud storage. To download the tool and accompanying instructions, click on the Download button below.

Figure 11. Download the Uploader CLI Tool

3.2.2 Cloning the Uploader CLI Tool from GitHub

The Uploader CLI Tool can also be cloned from the DataHub GitHub repository (https://github.com/CBIIT/crdc-datahub-cli-uploader/tree/master). To clone the repository to your local machine, use the following command:

```
git clone --recurse-submodules https://github.com/CBIIT/crdc-datahub-cli-uploader.git
```



n2iw Merge pull request #6 from CBIIT/Update-message-if-all-skipped

configs	change .gitignore	6 months ago
src	updated message	5 months ago
.gitignore	change .gitignore	6 months ago
.gitmodules	init the repo with all necessary files	8 months ago
README-technical.md	Rename README to README-technical	6 months ago
README.md	Added new README file	6 months ago
requirements.txt	update requirenebts.txt	7 months ago

Figure 12. The Uploader CLI Tool as it Appears in GitHub

3.3 Setting Up the Python Environment

The Uploader CLI Tool has Python library dependencies that you must install before the program will run. These dependencies can be installed by running the command `pip3 install -r requirements.txt`. If you want to install the dependencies individually, install the following libraries with pip3:

- pyyaml
- boto3
- requests
- requests_aws4auth

3.4 Using the Uploader CLI Tool

3.4.1 Uploader CLI Tool Configuration File

The behavior of the Uploader CLI Tool 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 Uploader CLI 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.

3.4.2 File Manifest

The Uploader CLI Tool uses a file manifest to upload the files to the Data Submission Portal temporary target storage. This is a TSV (tab-separated values) text file that contains columns for the file names, file sizes, and MD5 checksums. Note that the file.tsv template obtained from the Data Model viewer can be used as the file manifest for uploading. This can be convenient as it saves having to create two files with the same information.

These files are in YAML format and the Uploader CLI Tool will fail if the file is not valid YAML. YAML-aware text editors such as Microsoft Visual Studio Code, Sublime Text, or Notepad++ can be extremely helpful in preserving YAML formatting. The fields in this file are as follows:

- **api-url** – This field provides the Uploader CLI Tools with the URL/location of the temporary target storage used for communication and upload.

- **token** – This is the API access token that is obtained from the Data Submission Portal's graphical interface. To obtain an API token, log into the Data Submission Portal graphical interface to bring up the user menu, then select **API Token**. This opens 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. To find the correct submission ID, log into the system and select the project from the Data Submission Lists by clicking on the submission name. The Submission ID can be copied from the upper-left corner of the interface and can be copied to the clipboard by clicking the icon to the right of the Submission ID number.

- **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.

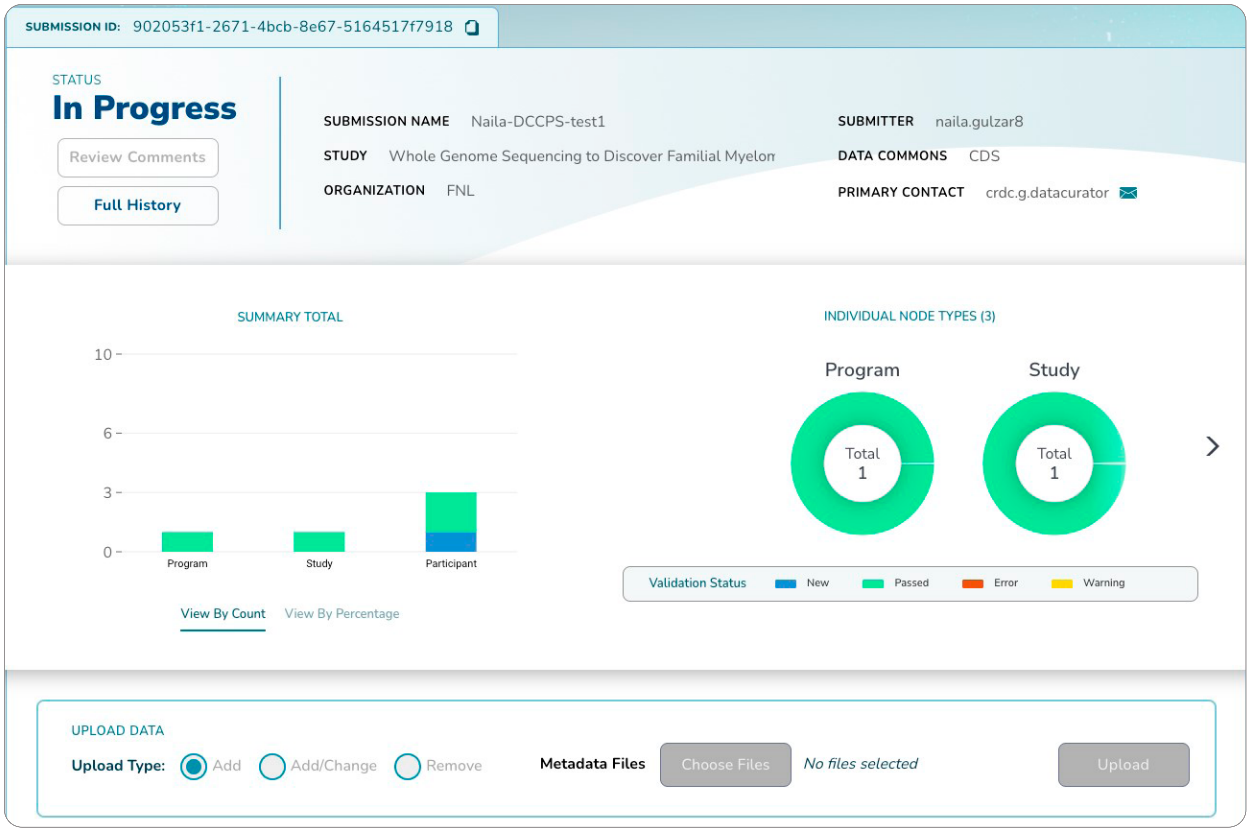
- **data** – This is the local path to the directory that contains the files to be uploaded.

- **manifest *(Data file upload only)*** – This is the local path to the manifest file.
- **name-field** – This is the column name in the manifest file that contains file names.
- **size-field *(Data file upload only)*** – This is the column name in the manifest file that contains file sizes.
- **md5-field *(Data file upload only)*** – This is the 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 Data Submission Portal's graphical interface.
- **retries** – This is the number of retries the Uploader CLI Tool will perform after a failed upload.
- **overwrite** – If this is set to *true*, the Uploader CLI Tool overwrites the file with the same name that already exists in the Data Submission Portal target storage. If set to *false*, the Uploader CLI tool does not upload if a file with the same name exists in the Data Submission Portal target storage.
- **dryrun** – If this is set to *true*, CLI does not upload any files to the Data Submission Portal target storage. If set to *false*, CLI uploads files to the Data Submission Portal target storage.

3.5 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 full path and file name for the completed 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:

3.6 Using the Data Submission Portal's Graphical Interface to Upload Metadata Forms



SUBMISSION ID: 902053f1-2671-4bcb-8e67-5164517f7918

STATUS
In Progress

SUBMISSION NAME Naila-DCCPS-test1

STUDY Whole Genome Sequencing to Discover Familial Myelon

ORGANIZATION FNL

SUBMITTER naila.gulzar8

DATA COMMONS CDS

PRIMARY CONTACT crdc.g.datacurator

UPLOAD DATA

Upload Type: ● Add ○ Add/Change ○ Remove

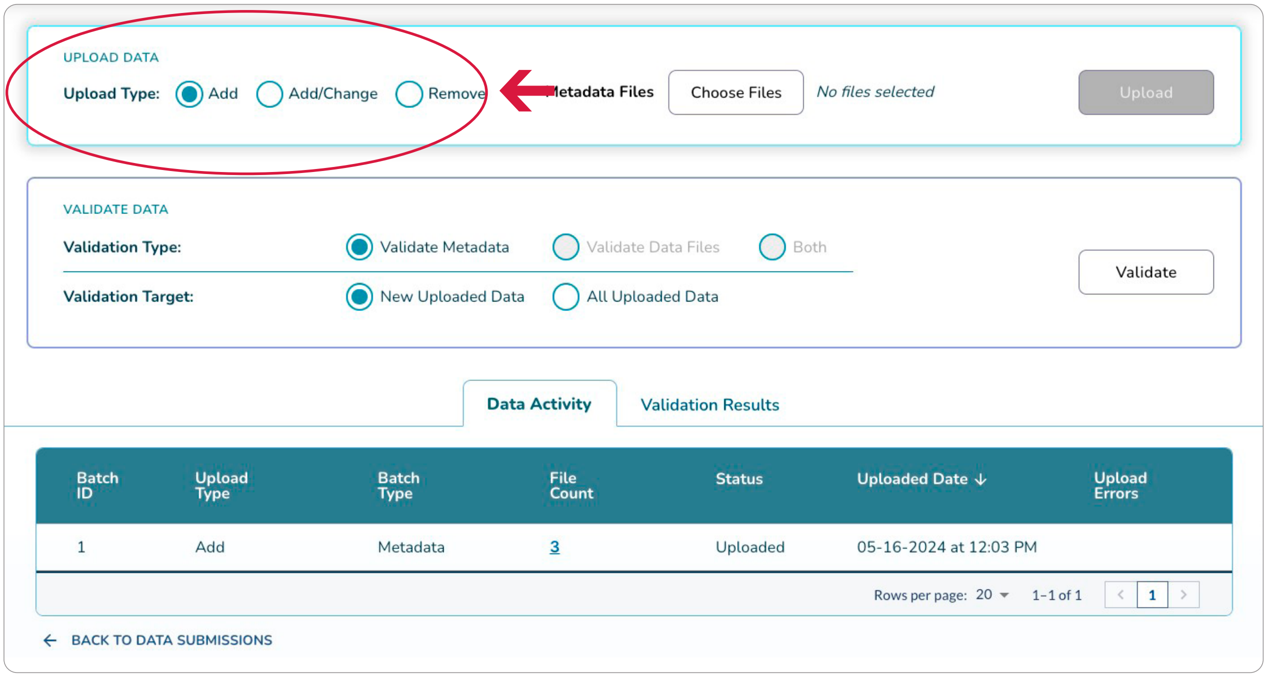
Metadata Files No files selected

Figure 13. Data Submission Status

The upload feature in the Data Submission Portal's graphical interface is intended for submitting completed metadata templates. To start the process, select what kind of upload this will be:

- **Add** – Use this if this is the first time the submission templates have been uploaded.
- **Add/Change** – 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.
- **Remove** – 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. If 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.



UPLOAD DATA

Upload Type: ● Add ○ Add/Change ○ Remove

Metadata Files Choose Files *No files selected* Upload

VALIDATE DATA

Validation Type: ● Validate Metadata ○ Validate Data Files ○ Both

Validation Target: ● New Uploaded Data ○ All Uploaded Data

Validate

Data Activity | Validation Results

Batch ID	Upload Type	Batch Type	File Count	Status	Uploaded Date ↓	Upload Errors
1	Add	Metadata	3	Uploaded	05-16-2024 at 12:03 PM	

Rows per page: 20 1–1 of 1 ‹ 1 ›

← BACK TO DATA SUBMISSIONS

Figure 14. Upload and Validate Data and Metadata

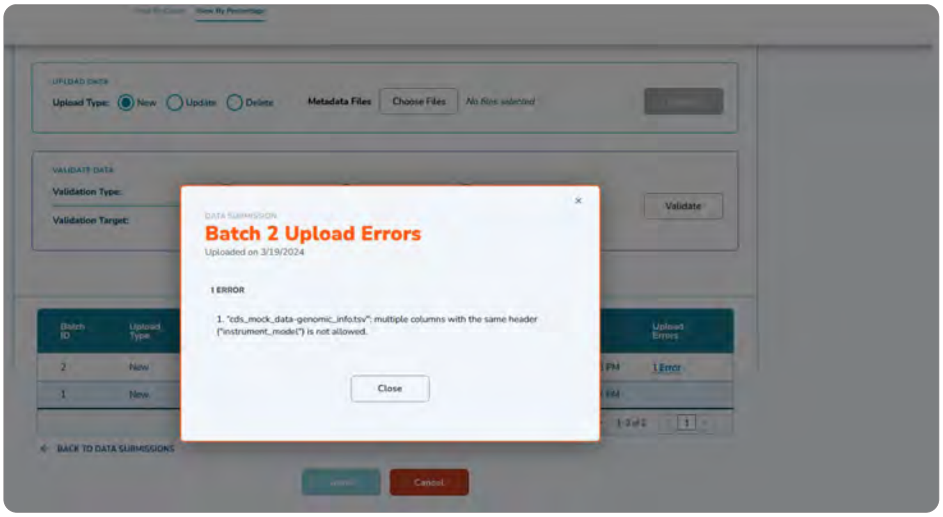
After selecting the **Add** click the **Choose Files** button, and select the metadata template files you want to submit. The quantity of files that you have selected appears. If that number is correct, click the **Upload** button to start the upload. The Status column in the Data Activity table displays "Uploading" until the upload and validation is completed. Once the files have been selected and uploaded, the Data Submission Portal automatically validates the files and reports the results in the table below. Successful files show "Uploaded" in the Status column. If a file fails validation, the status displays as "Failed."

Clicking the File Count button displays a list of all files uploaded in that batch. It's possible to repeat these validations by selecting **Validate Metadata** in the Validate Data portion of the page and clicking the **Validate** button.

Submitters can use the Add/Change and Remove functions to modify or remove specific metadata previously uploaded in the system. For instance, if a user wants to update metadata for a participant, they can click on the **Add/Change** button and upload the file with the desired changes only. Likewise, clicking on the **Remove** button allows users to upload a file containing metadata to be removed from the submission.

If there are errors in the metadata submission templates, the Status column displays "Failed" and the Upload Errors column displays a link to the errors. Clicking that link opens a dialog box that explains what errors have been encountered. Correct all identified errors and resubmit the file. Failed submissions do not count as "submissions," so if a *new* submission fails, subsequent uploads would also be *new* until the submitted file passes validation.

If multiple files are uploaded in a batch, a failure in one of the files fails the entire batch. All files in a failed batch must be resubmitted.



Batch 2 Upload Errors

Uploaded on 3/19/2024

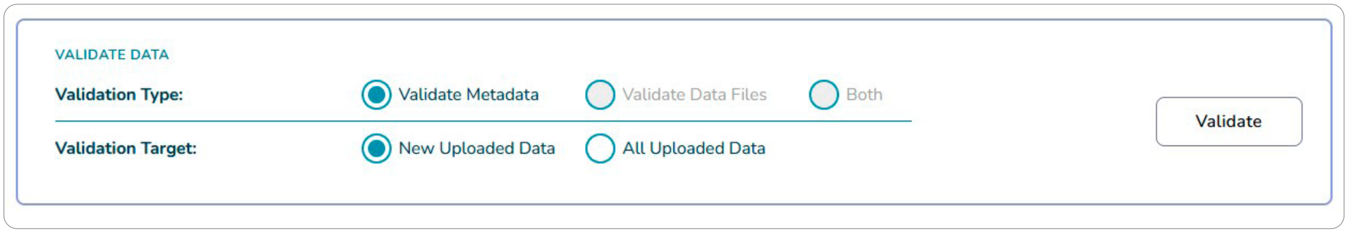
1 ERROR

1. "cds_mock_data-genomic_info.tsv": multiple columns with the same header ("instrument_model") is not allowed.

Figure 15. Batch Upload Errors

4. 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 selecting options in the Validate Data panel and clicking the **Validate** button.



VALIDATE DATA

Validation Type: ● Validate Metadata ○ Validate Data Files ○ Both

Validation Target: ● New Uploaded Data ○ All Uploaded Data

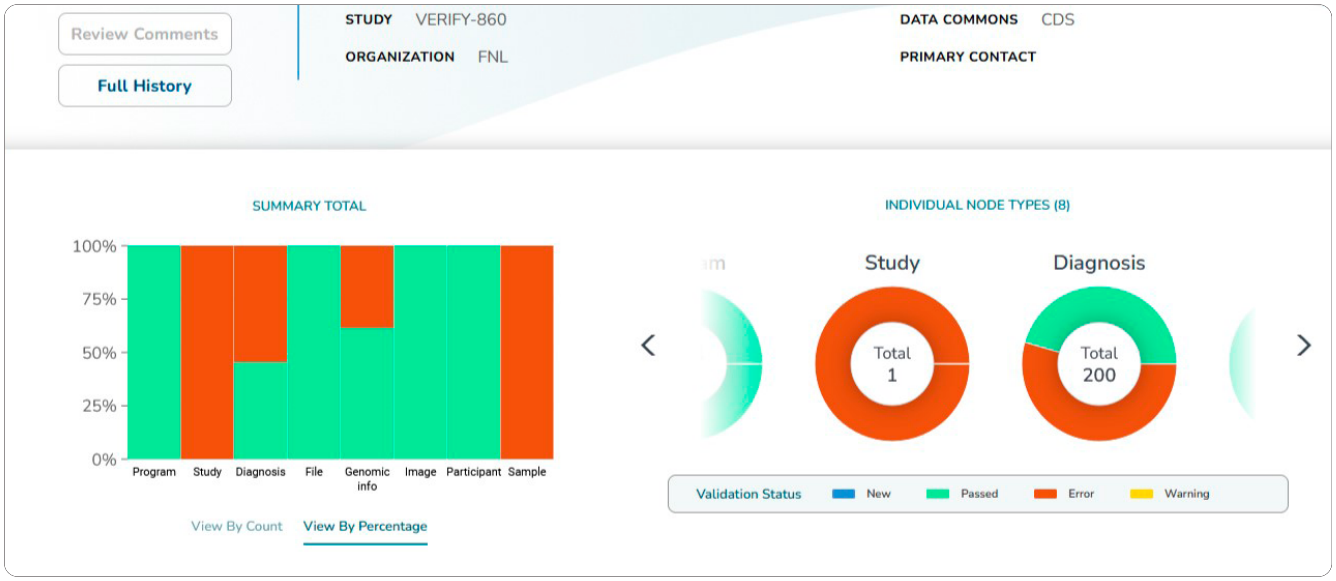
Figure 16. Validate Data Options

The first step is selecting which files to validate. The Validate Metadata option runs validations only on the submission metadata templates, and not on any of the uploaded data files. The Validate Data Files option does the reverse and checks all the uploaded data files. The Both option validates both.

By default, only newly uploaded files are validated. This can be a significant time saver for large submissions as some validations can take considerable time and the system keeps 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 checks everything that has been uploaded so far.

4.1 Reviewing Validation Results

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



STUDY VERIFY-860

ORGANIZATION FNL

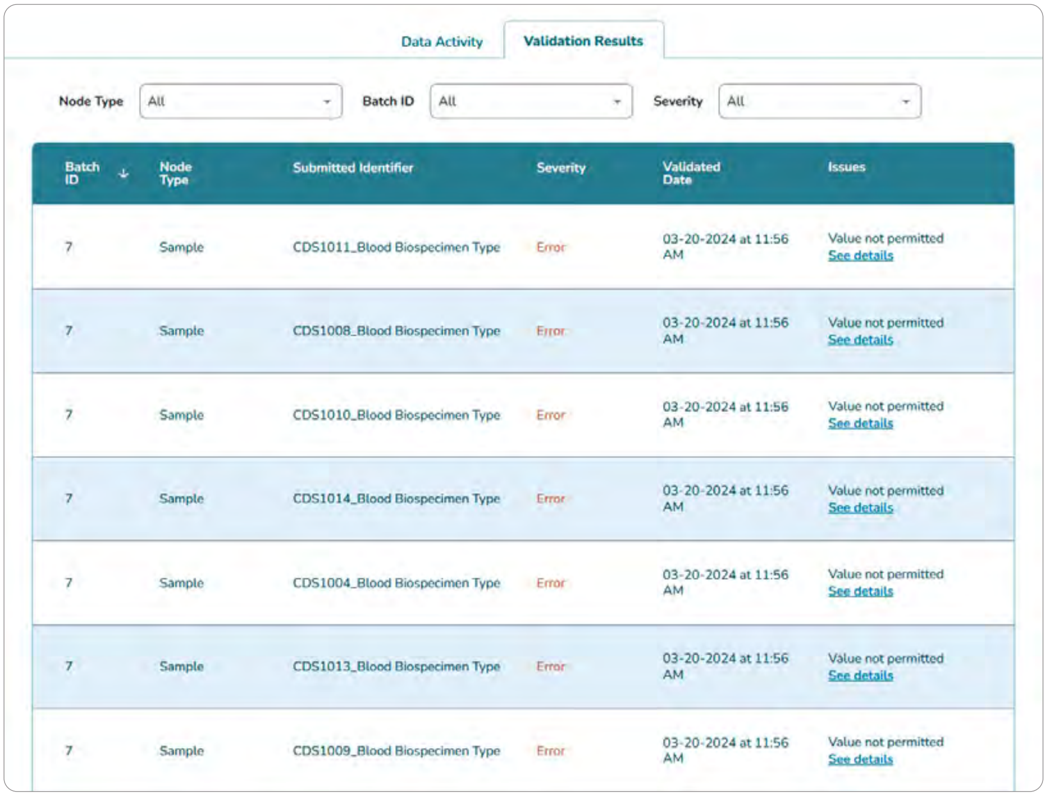
DATA COMMONS CDS

PRIMARY CONTACT

Figure 17. Validation Summary

The left graphic 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 generates 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.



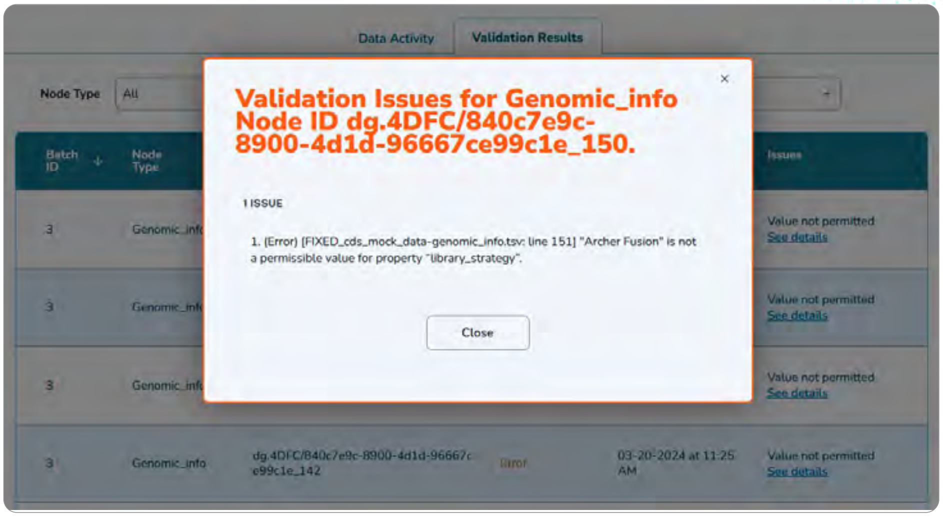
Node Type: All
Batch ID: All
Severity: All

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 See details
7	Sample	CDS1008_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted See details
7	Sample	CDS1010_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted See details
7	Sample	CDS1014_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted See details
7	Sample	CDS1004_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted See details
7	Sample	CDS1013_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted See details
7	Sample	CDS1009_Blood Biospecimen Type	Error	03-20-2024 at 11:56 AM	Value not permitted See details

Figure 18. Validation Errors Table

This table shows all 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** – These correlate 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 and is not a Data Submission Portal 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 is not required to be fixed)
- **Validated Date** – This is 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.



Validation Issues for Genomic_info Node ID dg.4DFC/840c7e9c-8900-4d1d-96667ce99c1e_150.

1 ISSUE

1. (Error) [FIXED_cds_mock_data-genomic_info.tsv: line 151] "Archer Fusion" is not a permissible value for property "library_strategy".

Batch ID	Node Type				Issues
3	Genomic_info				Value not permitted See details
3	Genomic_info				Value not permitted See details
3	Genomic_info				Value not permitted See details
3	Genomic_info	dg.4DFC/840c7e9c-8900-4d1d-96667ce99c1e_142	Error	03-20-2024 at 11:25 AM	Value not permitted See details

Figure 19. Validation Error Details

4.2 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 must 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.

5. 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 the **Submit** button locks the submission and passes control to the Data Submission team for a final check. No further changes will be allowed. Should the Submit button be clicked in error, please contact the Data Submission team and they can reject the submission and return it to your control.

6. WHAT TO EXPECT AFTER SUBMISSION

Once the final dataset has been submitted, the Data Submission team will perform some final checks to make sure everything is as required by 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 Data Submission team will reach out with additional questions and may re-open the submission to allow additional corrections.