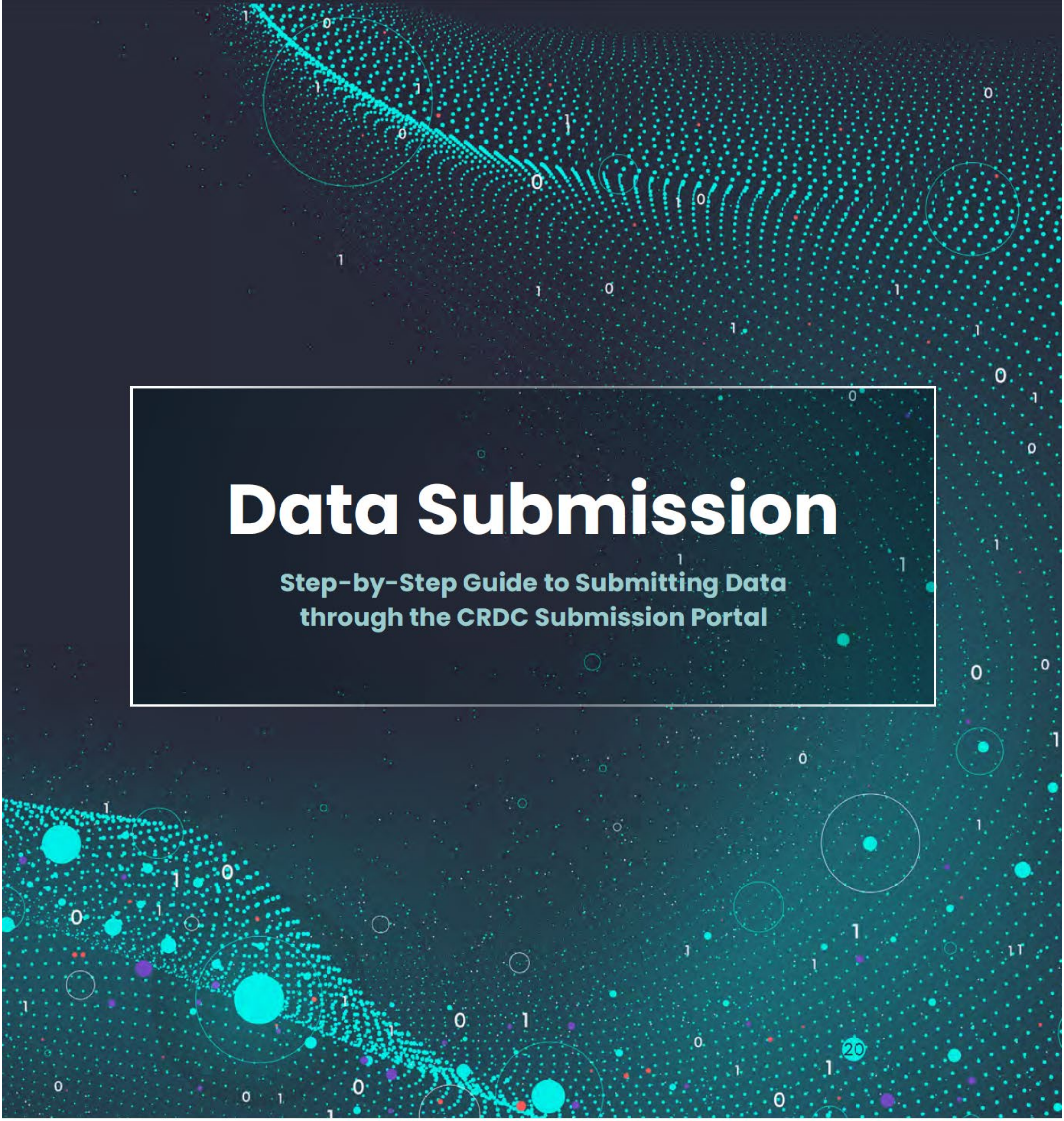


Data Submission

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

This tutorial walks you through the basics of submitting data to CRDC through 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 email the CRDC Helpdesk (NCICRDC@mail.nih.gov).

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 or in an email sent when the request gets approved.
- Create a Login.gov account. It is strongly recommended that the Login.gov identity be associated with your company or institution. Note: NIH staff can login using their PIV card.
- If the study contains controlled access data, you must register the study at dbGaP.
- Additionally, be aware that the CRDC Submission Portal 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. CRDC DATA MODELS

CRDC and its various Data Commons use data models to organize data in a consistent and structured manner, ensuring accuracy and facilitating reusability. CRDC 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. For example, the Study node is a child of the Program node and a parent of the Participant and File nodes.

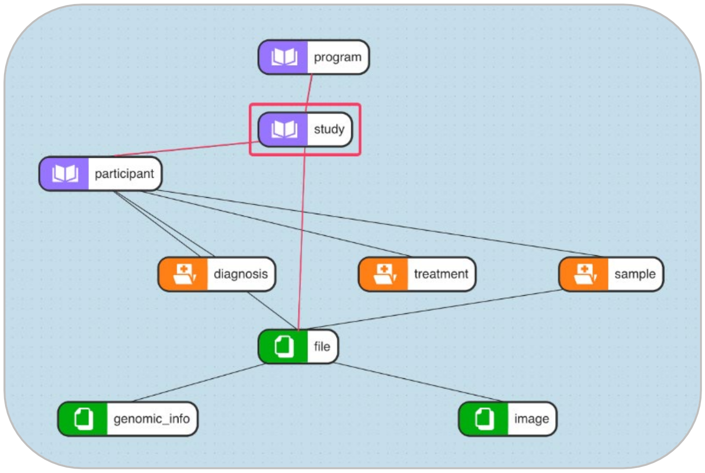


Figure 1. Study Node Relationships in the CDS Data Model

IV. DOCUMENTATION

Users can find the step-by-step instructions for submitting data on the CRDC Submission Portal under the **Documentation** tab. Additionally, instructions for submitting data using APIs are also available.

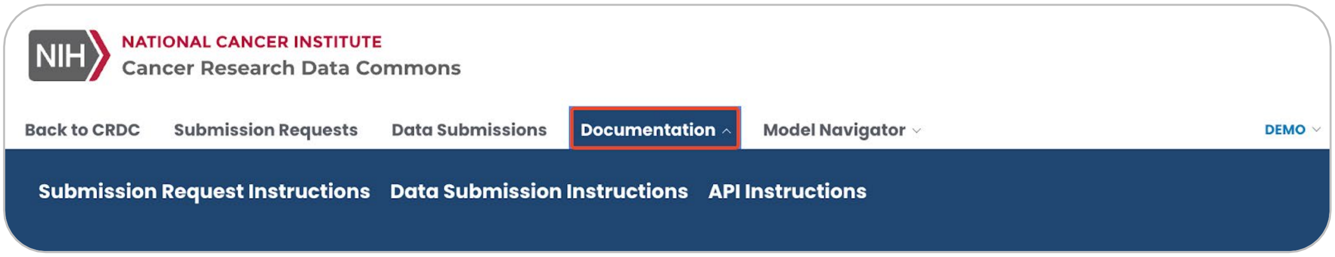

NIH
NATIONAL CANCER INSTITUTE
Cancer Research Data Commons
Back to CRDC
Submission Requests
Data Submissions
Documentation
Model Navigator
DEMO
Submission Request Instructions
Data Submission Instructions
API Instructions

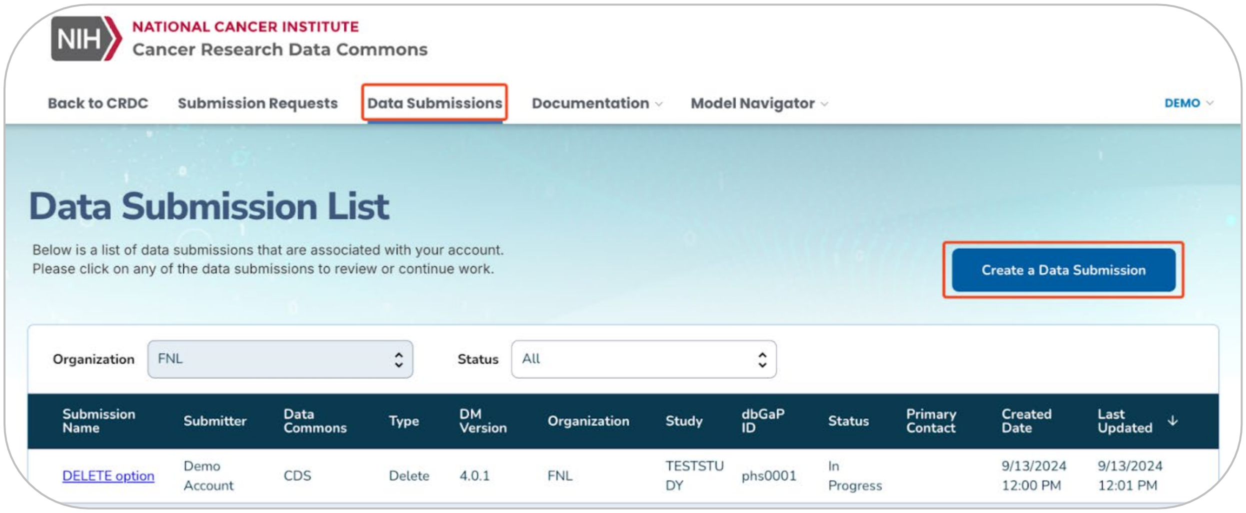
Figure 2. Documentation Menu Showing Available Documents

V. STARTING A NEW SUBMISSION

Do the following to start a new submission:

1. Log in to the CRDC Submission Portal.
2. Email the CRDC Helpdesk (NCICRDC@mail.nih.gov) with the following details: The email address used to log in the CRDC Submission Portal and the Study Title approved in the Submission Request Form.

   Note: You will be assigned the Submitter Role. The CRDC Helpdesk will notify you that you can start submitting data to the CRDC Submission Portal.
3. Click the **Data Submissions** menu. A table appears showing all your submissions to CRDC. The table is empty if this is your first submission.
4. Click the **Create a Data Submission** button.



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.

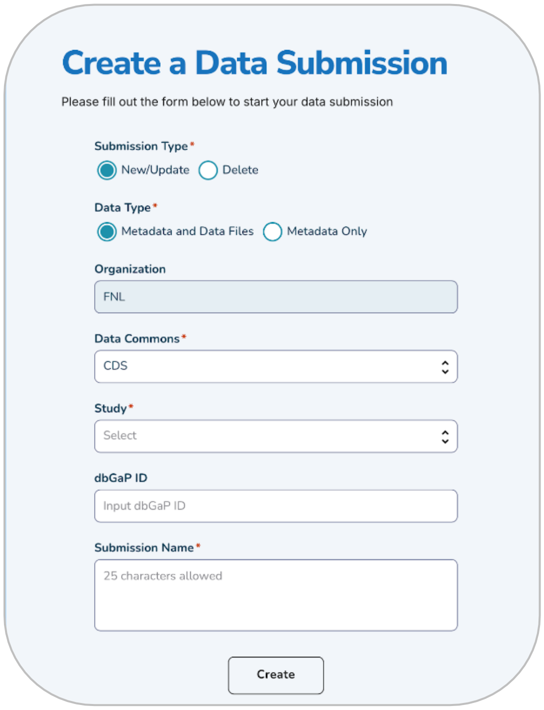
Organization: FNL

Status: All

Submission Name	Submitter	Data Commons	Type	DM Version	Organization	Study	dbGaP ID	Status	Primary Contact	Created Date	Last Updated
DELETE option	Demo Account	CDS	Delete	4.0.1	FNL	TESTSTUDY	phs0001	In Progress		9/13/2024 12:00 PM	9/13/2024 12:01 PM

Figure 3. Create a Data Submission

The Create a Data Submission dialog box appears. Fill out all fields.



Create a Data Submission

Please fill out the form below to start your data submission

Submission Type*
● New/Update ○ Delete

Data Type*
● Metadata and Data Files ○ Metadata Only

Organization
FNL

Data Commons*
CDS

Study*
Select

dbGaP ID
Input dbGaP ID

Submission Name*
25 characters allowed

Figure 4. Data Submission Dialog Box

1. Choose the Submission Type: select **New/Update** to create a new data submission or update an existing one. Select **Delete** option to submit a deletion request, to remove data from a previously submitted study already released publicly by the Data Commons.

2. Under the Data Type, indicate whether you are submitting both metadata and data files or only metadata files, selecting the **Metadata and Data Files**- or **Metadata Only** option, respectively.

   Note: The **Organization** box should already be populated with your organization.

3. Select the **Data Commons** that was approved by the Submission Review Committee (SRC) for your submission. If you are trying to submit to another CRDC Data Commons not already listed, email the CRDC Helpdesk at NCICRDC@mail.nih.gov.

4. The **Study** drop-down menu displays the Study Abbreviation or the Study Title you previously shared through the Submission Request form. If you notice an error in this list, email the CRDC Helpdesk.

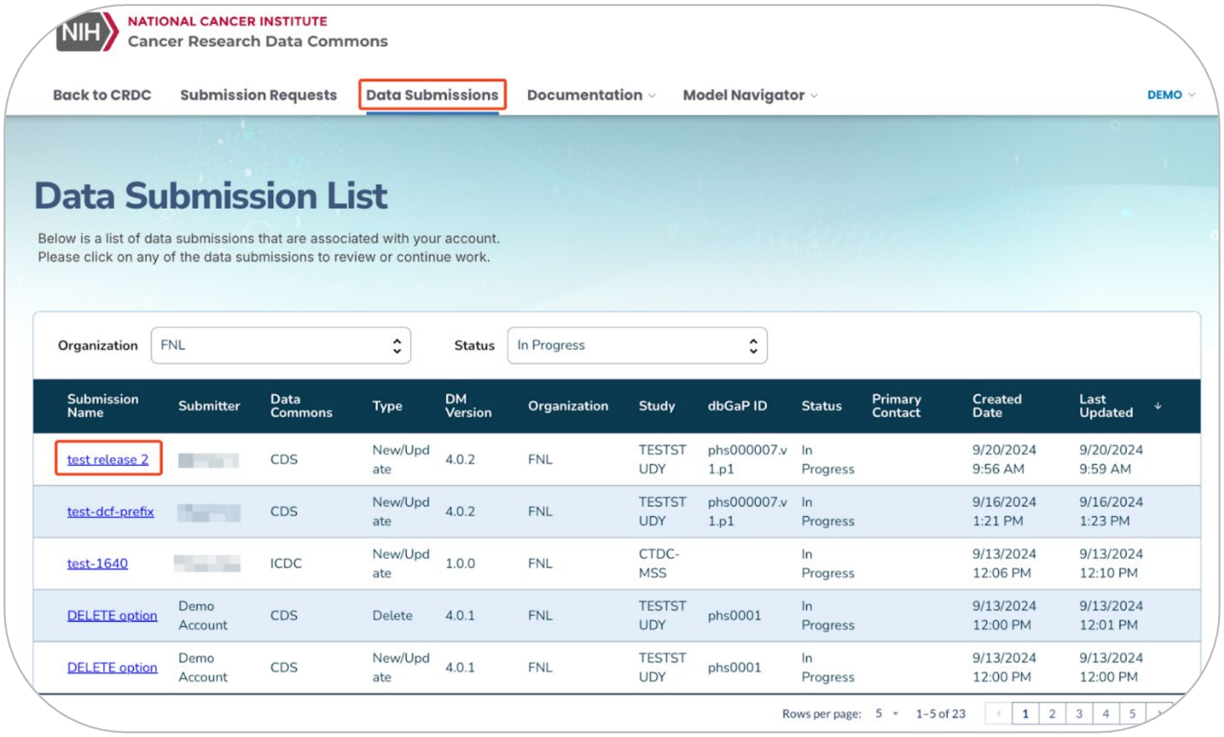
5. If your project contains controlled access data, enter the **dbGaP ID** (the database of Genotypes and Phenotypes accession number), which should be the same that was entered on the approved Submission Request form.

6. Optionally, enter a **Submission Name**, which is a free-text field available to users to label their submissions. This label will appear on the Submissions List table, which is accessible when you click the **Create** button.

When the new page is loaded, you can find the email address of the Data Submission team member assigned to your submission at the top of the page. From this stage on, please direct all questions related to the data submissions to the appointed Data Submission team member.

VI. CONTINUING AN EXISTING SUBMISSION

To access an existing submission and update it, go to the **Data Submission** tab. A table listing all the existing submissions appears on that page. Under the **Submission Name** column, select the submission you want to continue.



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: In Progress

Submission Name	Submitter	Data Commons	Type	DM Version	Organization	Study	dbGaP ID	Status	Primary Contact	Created Date	Last Updated
test release 2	[redacted]	CDS	New/Update	4.0.2	FNL	TESTSTUDY	phs000007.v1.p1	In Progress		9/20/2024 9:56 AM	9/20/2024 9:59 AM
test-dcf-prefix	[redacted]	CDS	New/Update	4.0.2	FNL	TESTSTUDY	phs000007.v1.p1	In Progress		9/16/2024 1:21 PM	9/16/2024 1:23 PM
test-1640	[redacted]	ICDC	New/Update	1.0.0	FNL	CTDC-MSS		In Progress		9/13/2024 12:06 PM	9/13/2024 12:10 PM
DELETE option	Demo Account	CDS	Delete	4.0.1	FNL	TESTSTUDY	phs0001	In Progress		9/13/2024 12:00 PM	9/13/2024 12:01 PM
DELETE option	Demo Account	CDS	New/Update	4.0.1	FNL	TESTSTUDY	phs0001	In Progress		9/13/2024 12:00 PM	9/13/2024 12:00 PM

Rows per page: 5   1–5 of 23

Figure 5. Active Data Submissions List

1. Obtaining Submission Templates

Submitting data to CRDC requires you to put your metadata into the submission templates. The metadata is then used to validate the information about the raw data. For instance, the file size of each uploaded raw data file will be compared with the file size specified in the metadata manifest. To get to the submission templates, click **Model Navigator** in the menu bar (see Figure 6) and then select the Data Model respective to the Data Commons (DC) that the Submission Review Committee (SRC) has approved for you. You can obtain submission templates for various Data Commons, including Cancer Data Services (CDS) Model, Clinical and Translational Data Commons (CTDC) Model or Integrated Canine Data Commons (ICDC) Model.

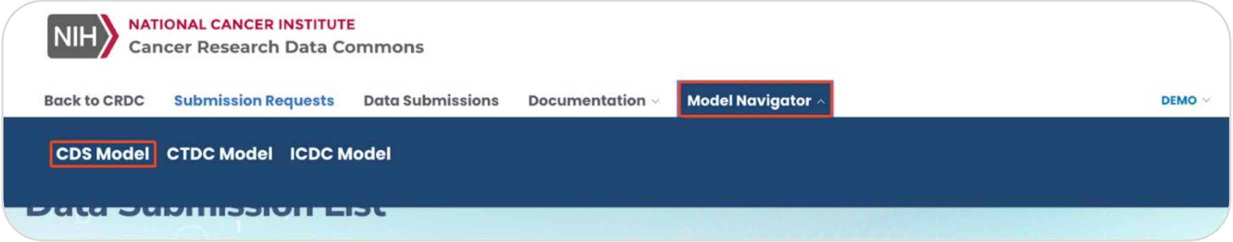


Figure 6. Use the Menu Bar to Navigate to the Data Model Viewer

Once you select the Data Model, the Data Model Viewer appears in the next screen, as seen in Figure 7. On this page, you can view the data model in detail and download the submission templates.

Note: Figure 7 shows the CDS Data Model as an example.

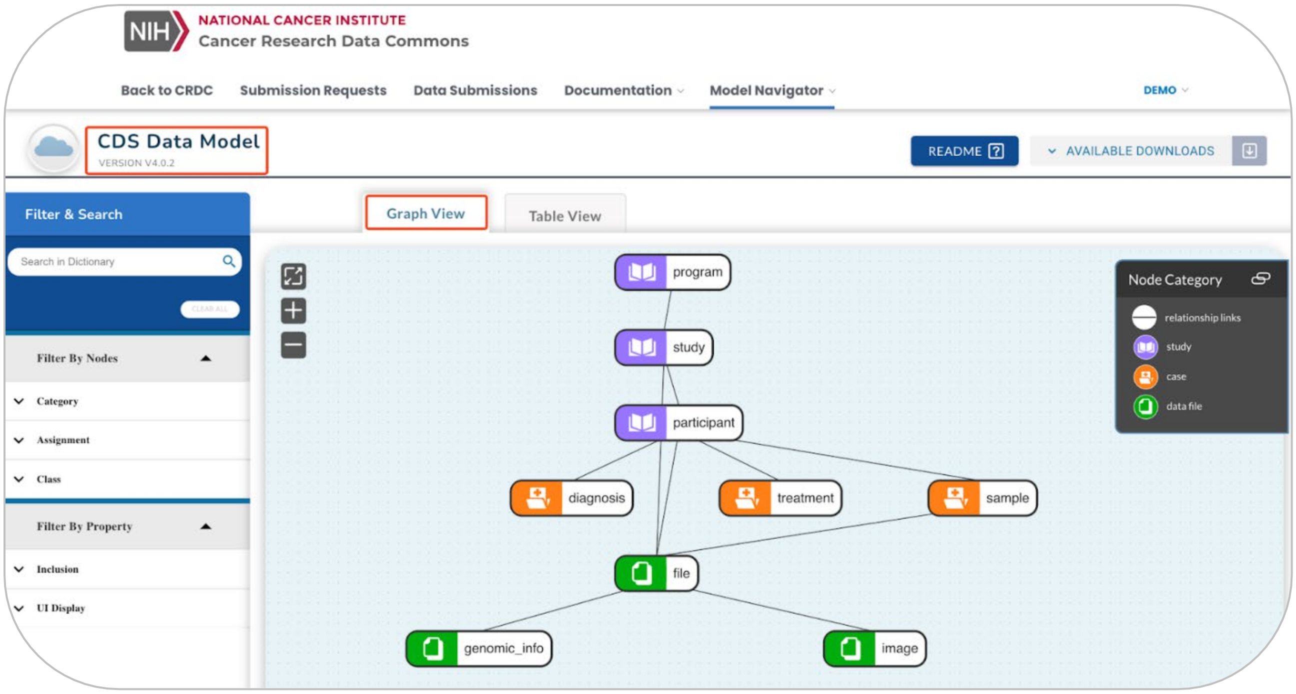
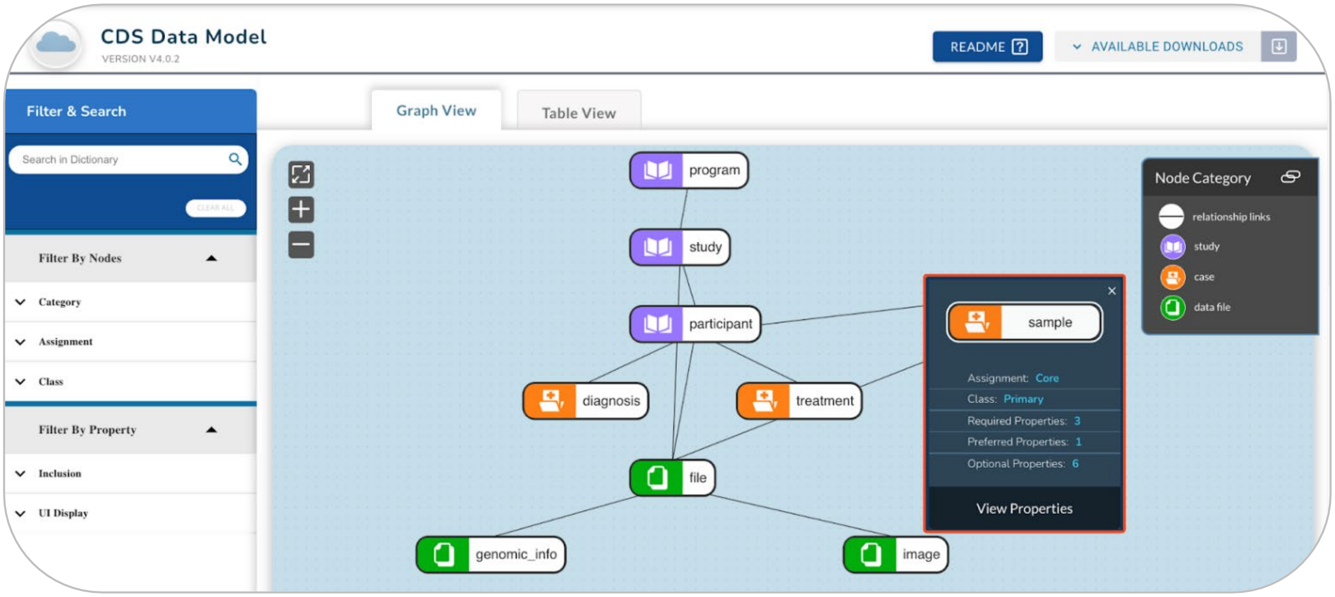


Figure 7. Data Model Viewer Graph View

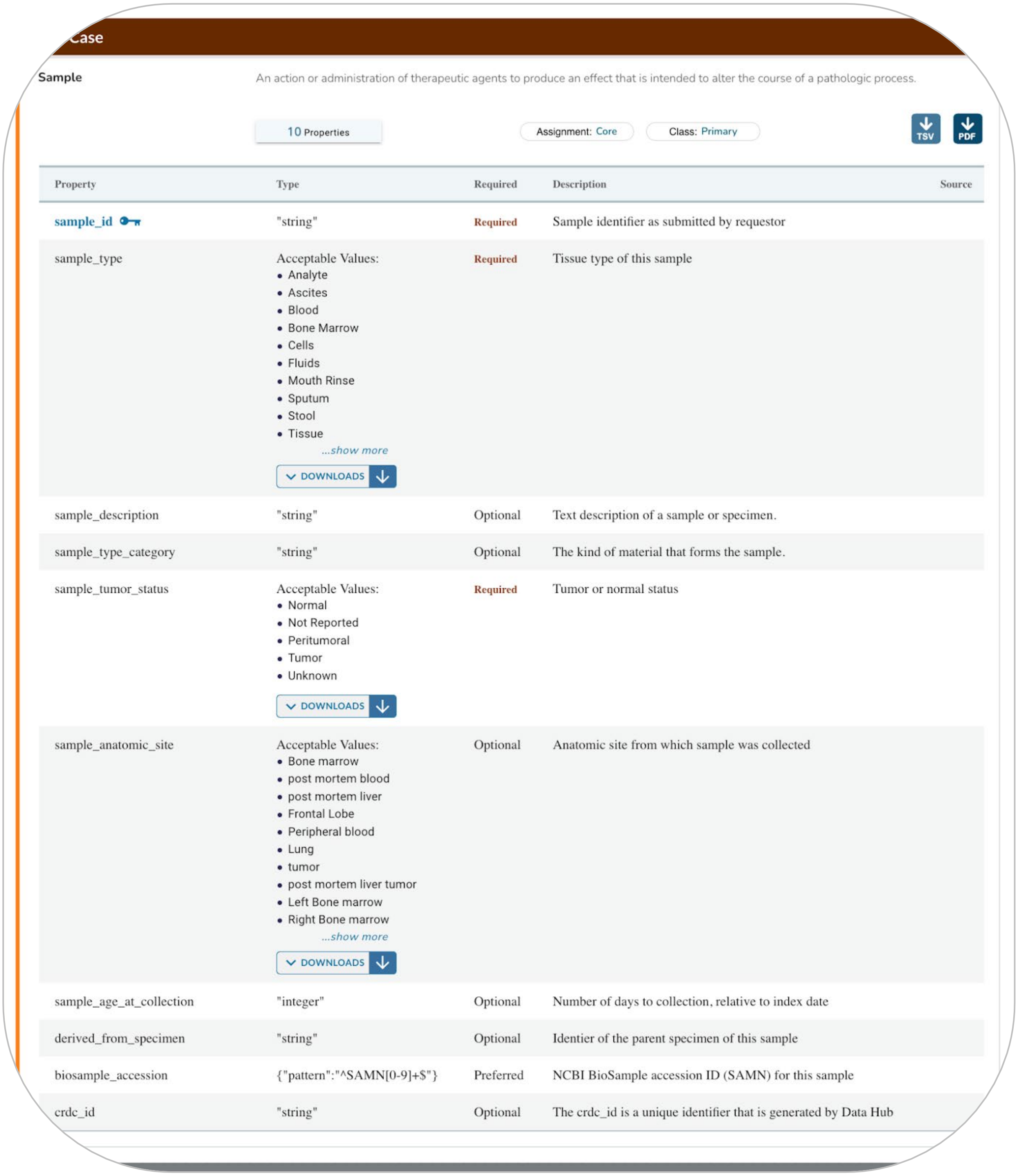
Use the Data Model Viewer to explore the data elements that submissions require or can accept. Click a node in the graph to open its summary. At the bottom of the node summary, click the **View Properties** menu to open a table view of the selected node.



Assignment: Core
Class: Primary
Required Properties: 3
Preferred Properties: 1
Optional Properties: 6

Figure 8. Click a Node to View a Summary and Open the Table View

The table view includes the description of data elements that are expected in that field (such as strings, integers, etc.) and which fields are required.



Case

Sample

An action or administration of therapeutic agents to produce an effect that is intended to alter the course of a pathologic process.

Property	Type	Required	Description	Source
sample_id	"string"	Required	Sample identifier as submitted by requestor	
sample_type	Acceptable Values: • Analyte • Ascites • Blood • Bone Marrow • Cells • Fluids • Mouth Rinse • Sputum • Stool • Tissue ...show more DOWNLOADS	Required	Tissue type of this sample	
sample_description	"string"	Optional	Text description of a sample or specimen.	
sample_type_category	"string"	Optional	The kind of material that forms the sample.	
sample_tumor_status	Acceptable Values: • Normal • Not Reported • Peritumoral • Tumor • Unknown DOWNLOADS	Required	Tumor or normal status	
sample_anatomic_site	Acceptable Values: • Bone marrow • post mortem blood • post mortem liver • Frontal Lobe • Peripheral blood • Lung • tumor • post mortem liver tumor • Left Bone marrow • Right Bone marrow ...show more DOWNLOADS	Optional	Anatomic site from which sample was collected	
sample_age_at_collection	"integer"	Optional	Number of days to collection, relative to index date	
derived_from_specimen	"string"	Optional	Identier of the parent specimen of this sample	
biosample_accession	{"pattern":"^SAMN[0-9]+$"}	Preferred	NCBI BioSample accession ID (SAMN) for this sample	
crdc_id	"string"	Optional	The crdc_id is a unique identifier that is generated by Data Hub	

Figure 9. Table View of a Sample Node

2. Downloading Submission Templates

To download the submission templates, select **Available Downloads**, as shown in Figure 10. Then select one of the following files from the drop-down menu and click the download arrow to start the download.

- **Data Dictionary (PDF)**
- **All Data Templates (TSV)**
- **All Vocabularies (TSV)**
- **All Vocabularies (JSON)**
- **Loading File Examples (TSV)**

Users can use the **All Data Templates** files to upload metadata on the CRDC Submission Portal. These templates must be in TSV format.

The **Data Dictionary** document provides detailed information about the required, preferred and optional data elements for all nodes within the selected Data Model. The **All Vocabularies** document contains the acceptable values for the data elements. The **Loading File Examples** are examples of completed submission templates designed to guide users in preparing the metadata manifest for their data**.** These can be useful to understand what each of the columns in the template is supposed to contain.

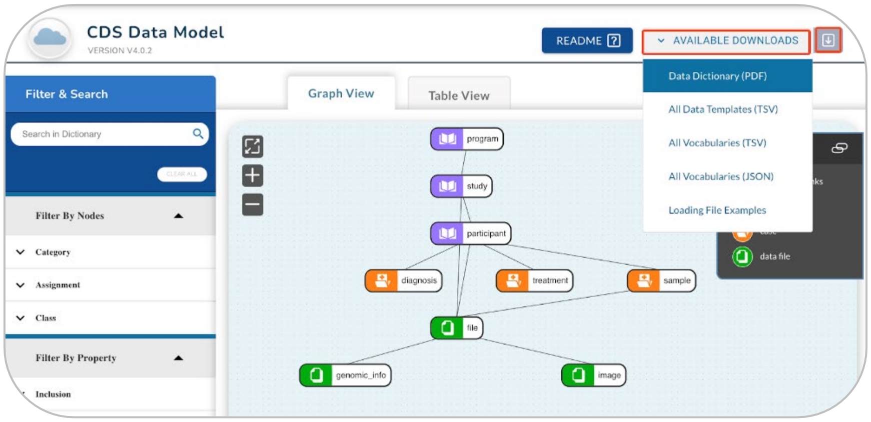
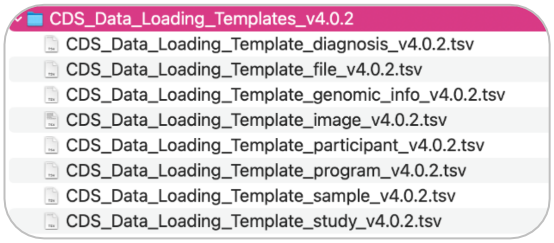


Figure 10. Using the Download Menu

The files are provided as a zip archive for download. Since these are tab-separated text files, they can be viewed in any text editor or spreadsheet program like Microsoft Excel or OpenOffice Calc. Several files should appear within the zip archive as shown in Figure 10.

Note: The exact content of files differs depending on the selected data model and associated submission process requirements.



CDS_Data_Loading_Templates_v4.0.2
CDS_Data_Loading_Template_diagnosis_v4.0.2.tsv
CDS_Data_Loading_Template_file_v4.0.2.tsv
CDS_Data_Loading_Template_genomic_info_v4.0.2.tsv
CDS_Data_Loading_Template_image_v4.0.2.tsv
CDS_Data_Loading_Template_participant_v4.0.2.tsv
CDS_Data_Loading_Template_program_v4.0.2.tsv
CDS_Data_Loading_Template_sample_v4.0.2.tsv
CDS_Data_Loading_Template_study_v4.0.2.tsv

Figure 11. Submission Templates Downloaded from the CRDC Submission Portal Model Viewer

2.1 Submission Templates and Fields

Each of the submission templates covers information relevant to the template; for example, the image template collects imaging data related information. Not all templates are required, rather only those templates relevant to the data being submitted are required. For instance, if the submission does not include imaging data, the user should not use the imaging submission template, CDS_Data_Loading_Template_image_v4.0.2.tsv.

For every template that will be submitted, review the Data Dictionary (accessible through the **Available Downloads** menu) to understand what fields are required, as each individual template has required fields, as well as preferred and optional data elements in the particular node. Note that you should not edit the first row of each template as it contains the property names and other special columns (explained below).

2.1.1 Special Columns

Every template has two types of special columns, also called parent mapping columns: “type” and “relationship.”

2.1.2 Type Column

A “type” column contains the name of the node type (such as study or genomic_info) and is required by all template files. In the downloaded template, the second row is prefilled with the correct node name for that specific template (e.g., study for the CDS_Data_Loading_Template_study_v4.0.2.tsv). All rows should contain the same node name in the “type” column. Mixing multiple node types in one file is not supported. For example, the node type Sample should not be mixed with the node type File and so on.

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, for the study node, the “program.program_acronym” column indicates that the study node has “program” node as its parent node, and the property used to identify the program node is “program_acronym.” Each value in the “program.program_acronym” column is an acronym used for a program, such as HTAN.

Below are brief descriptions of the templates that can be used:

- **Diagnosis** – This template should contain information related to the participant’s diagnosis. This includes 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 experiments such as the library strategy and the sequencing platform. You can ignore this form 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; it can be excluded when they are not.

- **Participant** – This template contains basic information about the participants in the submission, including participant identifiers used in the study.

- **Sample** – This template allows for a description of the samples and an indication of which participant they are associated with. In some cases, studies may not have participants, in which case you can ignore both the participant sheet and the participant ID column.

- **Program** – This template asks for information about the program with which the submitted data are associated; one line is usually sufficient. Note that for CRDC purposes, certain programs are defined at an NCI-level, such as the Human Tumor Atlas Network (HTAN), or Childhood Cancer Data Initiative (CCDI).

- **Study** – This template is for information about the study and includes information such as the study name, study description, and dbGaP ID. Note that the Study is a child node of the parent node Program.

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

3. Uploading Data Files and Metadata Manifests

You can move files from their local environment to the CRDC through the Submission Portal in the following two ways:

- **Uploader CLI Tool** – This command-line interface is used to transfer primary data files like genomic sequence files or imaging data files into the CRDC Submission Portal.

- **Graphical interface** – The graphical interface can be used to upload metadata files such as the Submission Templates.

- **Note:** Submit primary data files using the Uploader CLI Tool only. Do not attempt to upload data files using the CRDC Submission Portal's graphical interface.

3.1 Uploader CLI Tool

3.1.1 Introduction

The CRDC Submission Portal provides a command-line interface for uploading datasets to its temporary target storage. You can install and use it on any system capable of running Python 3.6 or higher.

Notes:

- There are detailed instructions on downloading, installing, and running the Uploader CLI Tool in the README file of the GitHub repository.

- 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 CRDC 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

You can download the Uploader CLI Tool either directly from the CRDC Submission Portal or by cloning the GitHub repository.

3.2.1 Download the Uploader CLI Tool from the CRDC Submission Portal

Click on your user profile name, found in the upper-right corner of the Data Submission page, to open the menu. Select **Uploader CLI Tool** from the menu.

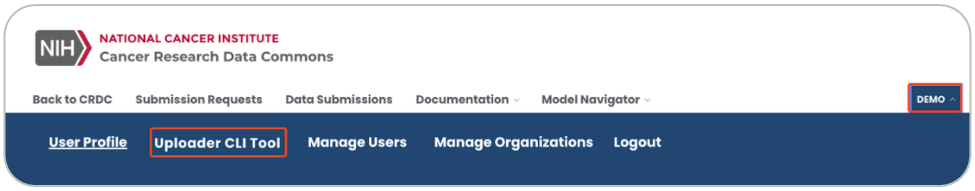
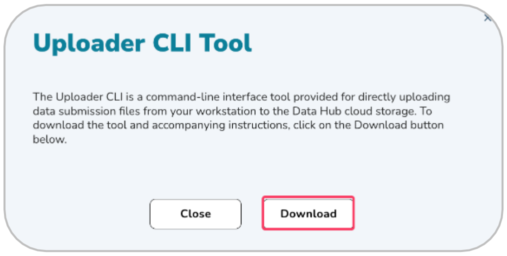


Figure 12. Menu with the Uploader CLI Tool Download Option

Click the **Download** button from the new window that pops up. A zip archive downloads to your local machine.


Uploader CLI Tool

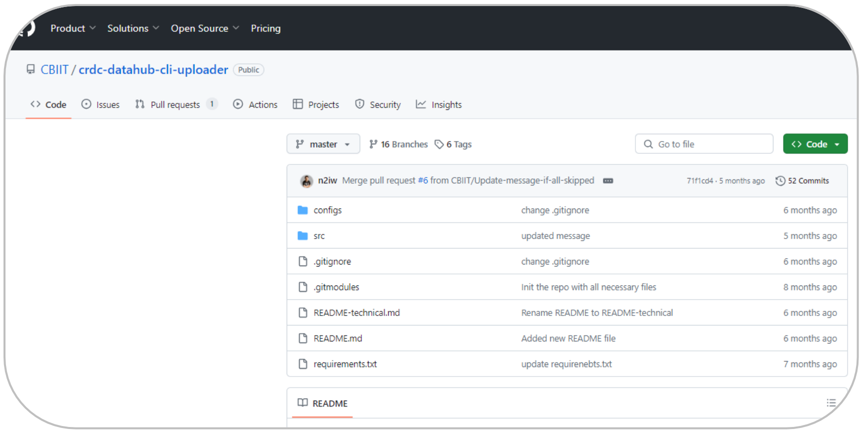
The Uploader CLI is a command-line interface tool provided for directly uploading data submission files from your workstation to the Data Hub cloud storage. To download the tool and accompanying instructions, click on the Download button below.

Figure 13. 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 Data Hub GitHub repository. 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 71f1cd4 · 5 months ago 52 Commits

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 14. Uploader CLI Tool as it Appears in GitHub

3.3 Setting Up the Python Environment

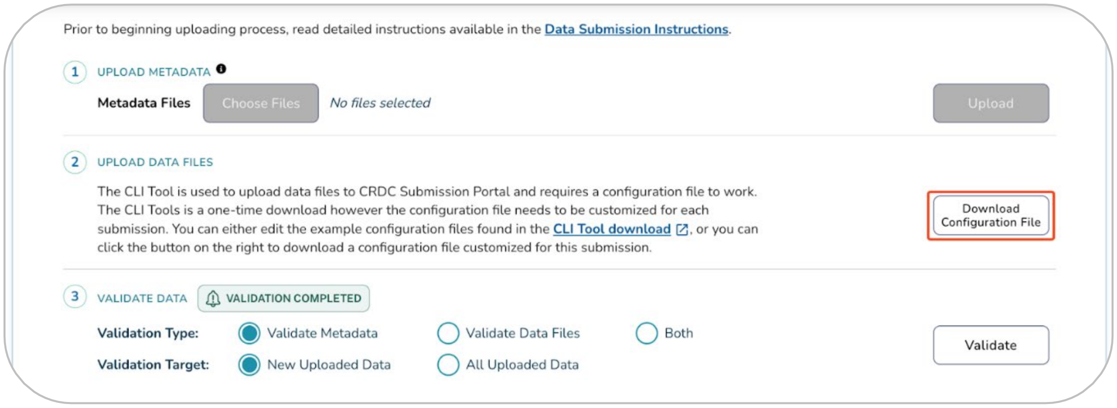
The Uploader CLI Tool has Python library dependencies that you must install before running the CLI. These dependencies can be installed by running the command `pip3 install -r requirements.txt`. If you want to install the dependencies individually, install the following libraries with pip3:

- 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 the configuration file. You can directly download the configuration file from the CRDC Submission Portal interface to upload the Data Files. Click the **Download** Configuration **File** button to open a pop-up window. Enter the path to the local folder containing your Data Files and the local path to the Manifest File (explained in Section 3.4.2) in the designated fields. Then, click the **Download** button to download the configuration file in YML format to your computer.



Prior to beginning uploading process, read detailed instructions available in the Data Submission Instructions.

1 UPLOAD METADATA

Metadata Files: Choose Files — No files selected — Upload

2 UPLOAD DATA FILES

The CLI Tool is used to upload data files to CRDC Submission Portal and requires a configuration file to work. The CLI Tools is a one-time download however the configuration file needs to be customized for each submission. You can either edit the example configuration files found in the CLI Tool download, or you can click the button on the right to download a configuration file customized for this submission.

Download Configuration File

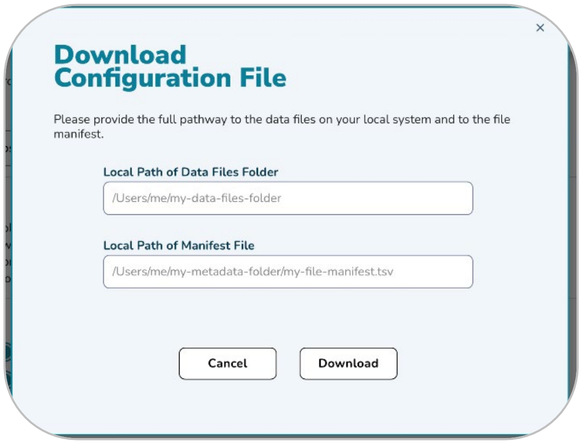
3 VALIDATE DATA — VALIDATION COMPLETED

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

Validation Target: ● New Uploaded Data ○ All Uploaded Data

Validate

Figure 15. Download Configuration File



Download Configuration File

Please provide the full pathway to the data files on your local system and to the file manifest.

Local Path of Data Files Folder: /Users/me/my-data-files-folder

Local Path of Manifest File: /Users/me/my-metadata-folder/my-file-manifest.tsv

Cancel — Download

Figure 16. Dependencies to Download Configuration File

Additionally, if you need to populate the configuration files manually, you can find examples in the *configs* directory of either the extracted zip file or the cloned GitHub repository. The examples provided are the same configuration files modified for the two different upload types:

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

These files are in YAML format and the Uploader CLI Tool will fail if the file is not a 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 CRDC storage used for API communications and upload.
- **token** – This is the API access token that is obtained from the CRDC Submission Portal's graphical interface. To obtain an API token, log into the CRDC 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. You can copy the Submission ID from the upper-left corner of the interface by clicking the icon to the right of the Submission ID number.
  Note: A project consists of one or more submissions (often many more), with each submission ID linked to the parent project. A single user working on multiple projects must carefully track which submission IDs they are uploading to ensure the data is associated with the correct project.
- **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.
- **file manifest (*Data file upload only*)** – This is the local path to the manifest file.
- **id-field (*Data file upload only*)** – This is the column name in the manifest file that contains file IDs (Keys). Please refer to the data model regarding which property is the ID/Key property.
- **omit-DCF-prefix (*Data file upload only*)** – For most data commons, this should be set to "false." One exception is ICDC, which should be set to "true."
- **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 CRDC 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 CRDC 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 CRD Submission Portal target storage.
- **dryrun** – If this is set to *true*, CLI does not upload any files to the CRDC Submission Portal target storage. If set to *false*, CLI uploads files to the CRDC Submission Portal target storage.

3.4.2 File Manifest

The Uploader CLI Tool uses a file manifest to upload the files to the CRDC 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 available 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.

File.tsv template downloaded from Data Model viewer doesn't contain file Keys/IDs column, because the system will generate it for you. The generated final manifest will be saved in the same place the file manifest is and has a "-final" suffix. For example, file-final.tsv. Uploader CLI Tool will upload this final manifest for you once all data files have been successfully uploaded. So that you don't have to upload file manifest in the CRDC submission portal. However, if you need to update the content of the file manifest, please make sure to edit the **final manifes**t and upload it to the CRDC submission portal. You can also use the final manifest in the Uploader CLI Tool, if you want to upload the data files again for any reason. In that case, the Uploader CLI Tool will use the file IDs/Keys provided in the file manifest instead of generating new ones.

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

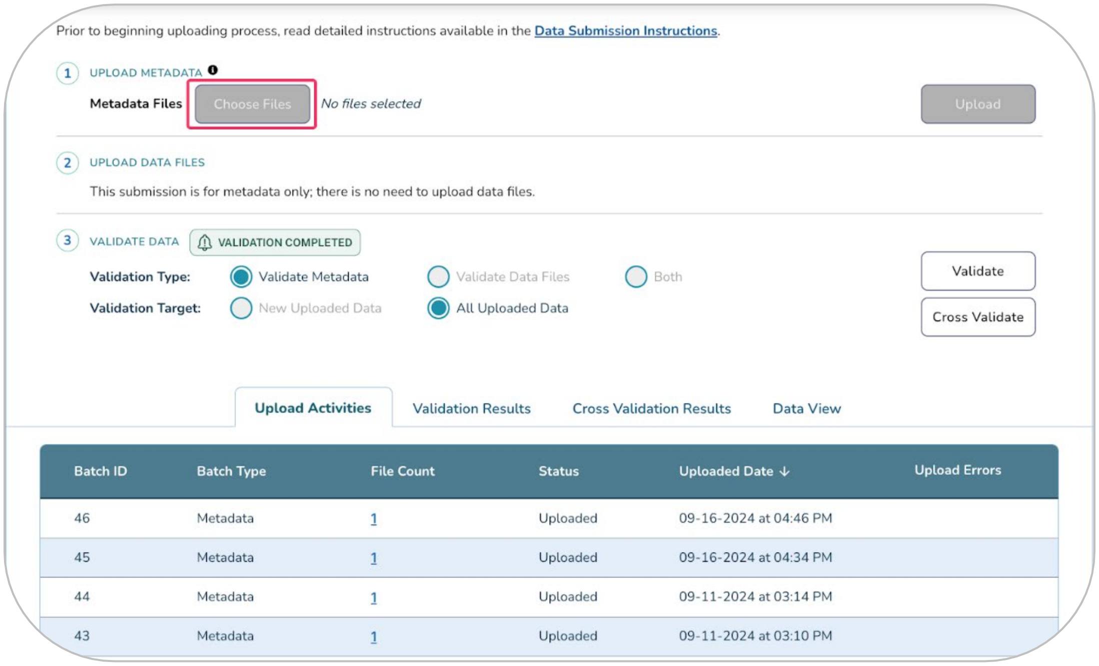
command should look something like the following, though the exact details may be customized depending on how the tool (and Python) were installed:

```
$ python3 path/to/uploader.py --config path/to/metadata-upload.yml
```

3.6 Using the CRDC Submission Portal's Graphical Interface to Upload Metadata Submission Templates

The Upload Metadata feature in the CRDC Submission Portal's graphical interface is intended for submitting completed metadata templates. Users can access tooltips and a link to the data submission user guide directly from the interface. The **Upload Activities** table tracks the uploading process of data and metadata files and provides details on any errors related to failed uploads. After the data upload is complete, the system automatically runs the basic validation process, and the results are shown in the **Validation Results** table. You can delete the specific data or metadata files within the submission in the **Data View** table. Refer to Figure 16.

In case you intend to upload the metadata in phases, you should keep your new information and any subsequent updates separated. For example, if a study has 100 participants, the submitted template could either contain all 100 or a subset of that 100, with the remainder submitted in later uploads. If there is no overlap in participants between the different uploads, the system will not flag an error. 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. To make corrections, select the file you want to delete in the Data View table and click the **Delete** button.



Prior to beginning uploading process, read detailed instructions available in the Data Submission Instructions.

1 UPLOAD METADATA

Metadata Files Choose Files *No files selected* Upload

2 UPLOAD DATA FILES

This submission is for metadata only; there is no need to upload data files.

3 VALIDATE DATA VALIDATION COMPLETED

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

Validation Target: ○ New Uploaded Data ● All Uploaded Data

Validate

Cross Validate

Upload Activities | Validation Results | Cross Validation Results | Data View

Batch ID	Batch Type	File Count	Status	Uploaded Date ↓	Upload Errors
46	Metadata	1	Uploaded	09-16-2024 at 04:46 PM	
45	Metadata	1	Uploaded	09-16-2024 at 04:34 PM	
44	Metadata	1	Uploaded	09-11-2024 at 03:14 PM	
43	Metadata	1	Uploaded	09-11-2024 at 03:10 PM	

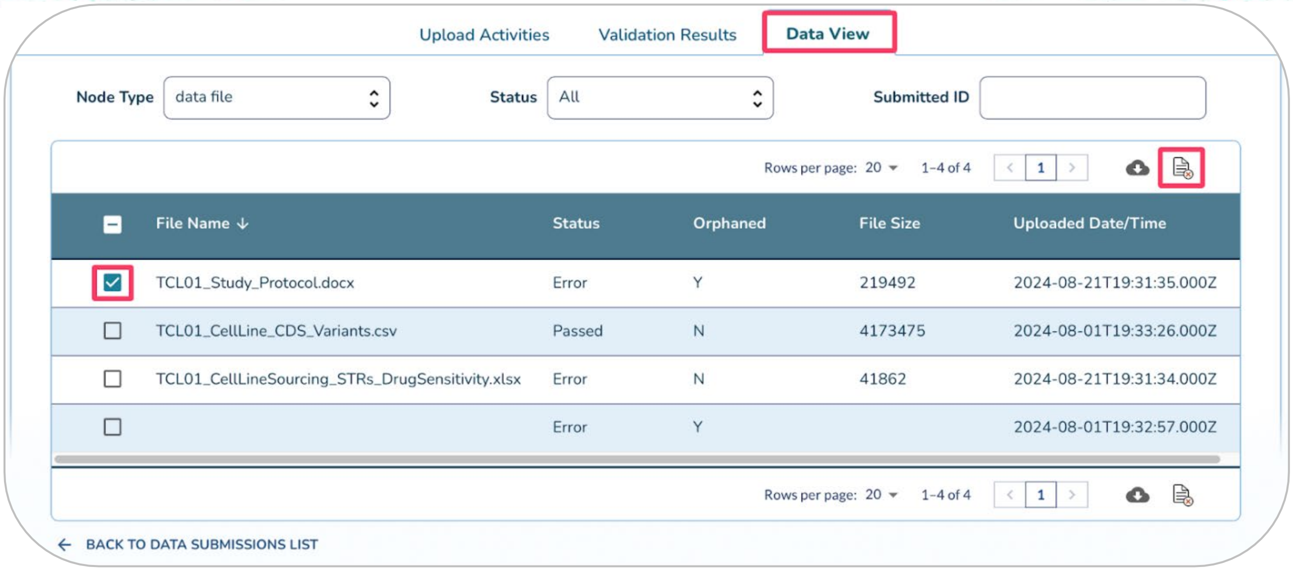
Figure 17. Upload and Validate Data and Metadata

As depicted in Figure 17, to start the upload process, click the ***Choose Files*** button, and select the metadata Submission Templates you want to submit. The total number of files that you select appears. If that number is correct, click the ***Upload*** button to start the upload. The Status column in the Upload Activities table displays “Uploading” until the upload and primary validation are completed. Once you have selected and uploaded the files, the CRDC Submission Portal automatically performs basic validations on the files and reports the results in the Validation Results table. 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 is possible to repeat these validations by selecting ***Validate Metadata*** in the Validate Data portion of the page and clicking the ***Validate*** button.

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.

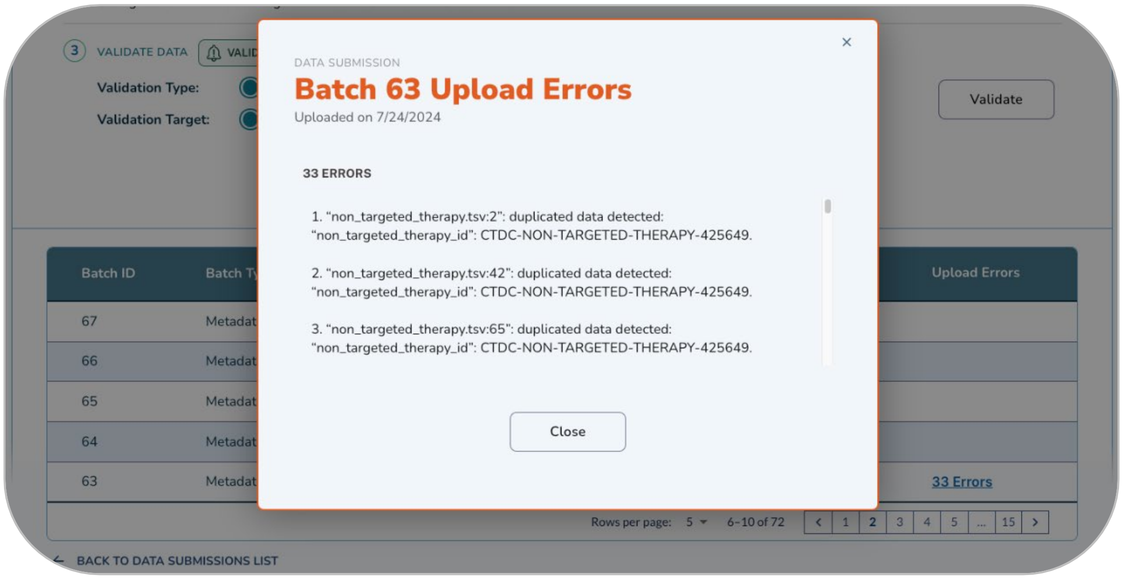
You can remove specific metadata previously uploaded in the system in the Data View table, as depicted in Figure 18. For instance, if you want to remove metadata for a participant, select the file and click the ***Delete*** icon to remove the specific file from your submission. You can also download the selected metadata files in the Data View table by selecting the file(s) and clicking on the cloud shaped ***download*** icon.



☐	File Name ↓	Status	Orphaned	File Size	Uploaded Date/Time
☑	TCL01_Study_Protocol.docx	Error	Y	219492	2024-08-21T19:31:35.000Z
☐	TCL01_CellLine_CDS_Variants.csv	Passed	N	4173475	2024-08-01T19:33:26.000Z
☐	TCL01_CellLineSourcing_STRs_DrugSensitivity.xlsx	Error	N	41862	2024-08-21T19:31:34.000Z
☐		Error	Y		2024-08-01T19:32:57.000Z

Figure 18. Delete and Download Metadata File(s)

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. An example of batch upload error message is presented in Figure 19.



DATA SUBMISSION

Batch 63 Upload Errors

Uploaded on 7/24/2024

33 ERRORS

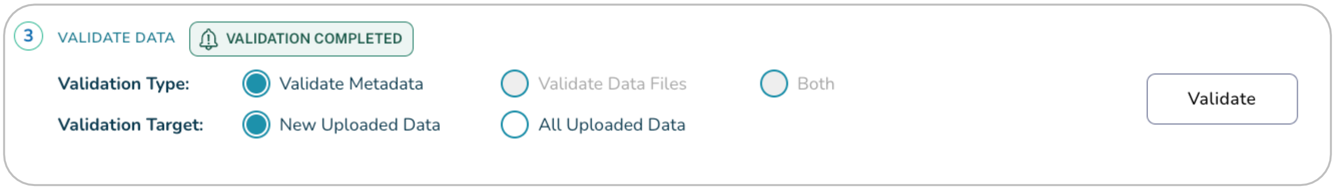
1. "non_targeted_therapy.tsv:2": duplicated data detected: "non_targeted_therapy_id": CTDC-NON-TARGETED-THERAPY-425649.
2. "non_targeted_therapy.tsv:42": duplicated data detected: "non_targeted_therapy_id": CTDC-NON-TARGETED-THERAPY-425649.
3. "non_targeted_therapy.tsv:65": duplicated data detected: "non_targeted_therapy_id": CTDC-NON-TARGETED-THERAPY-425649.

Close

Figure 19. 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. To run a validation, select options in the Validate Data panel and click the ***Validate*** button. Refer to Figure 20.



3 VALIDATE DATA — VALIDATION COMPLETED

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

Validation Target: ● New Uploaded Data ○ All Uploaded Data

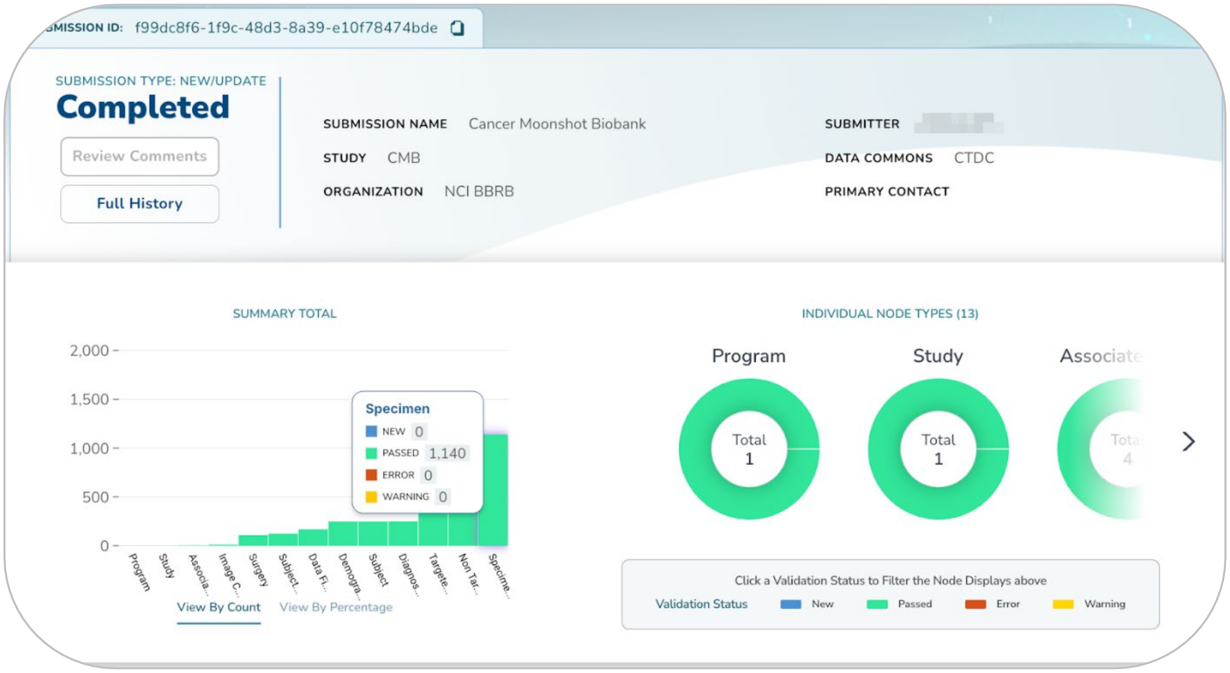
Figure 20. Validate Data Options

The first step is selecting which files to validate. The ***Validate Metadata*** option runs validations only on the submission metadata templates, 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 as depicted in Figure 21.



SMISSION ID: f99dc8f6-1f9c-48d3-8a39-e10f78474bde

SUBMISSION TYPE: NEW/UPDATE

Completed

SUBMISSION NAME Cancer Moonshot Biobank

STUDY CMB

ORGANIZATION NCI BBRB

SUBMITTER [redacted]

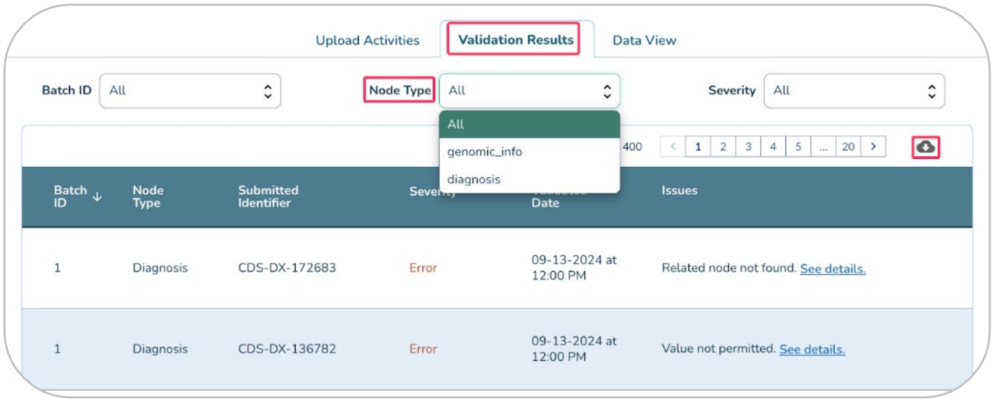
DATA COMMONS CTDC

PRIMARY CONTACT

Figure 21. Validation Summary

The left graphic in Figure 21 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 in the **Validation Results** table. Users can select to view errors and warnings from either a file or data file using the Node Type dropdown menu. Choosing "All" from this menu displays a list of all files containing errors or warnings. The Node Type *file* pertains to metadata manifest templates, while the *data file* refers to raw data, such as sequencing data.



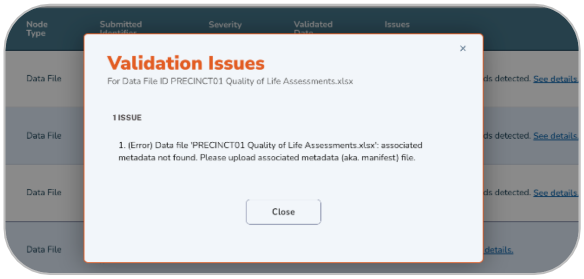
Batch ID: All
Node Type: All
Severity: All

Batch ID	Node Type	Submitted Identifier	Sever	Date	Issues
1	Diagnosis	CDS-DX-172683	Error	09-13-2024 at 12:00 PM	Related node not found. See details.
1	Diagnosis	CDS-DX-136782	Error	09-13-2024 at 12:00 PM	Value not permitted. See details.

Figure 22. Validation Results 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 templates. 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 user and is not a CRDC 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. Figure 23 presents the example of the Validation Issue details.



Validation Issues

For Data File ID PRECINCT01 Quality of Life Assessments.xlsx

1 ISSUE

1. (Error) Data file 'PRECINCT01 Quality of Life Assessments.xlsx': associated metadata not found. Please upload associated metadata (aka. manifest) file.

Figure 23. Validation Error Details

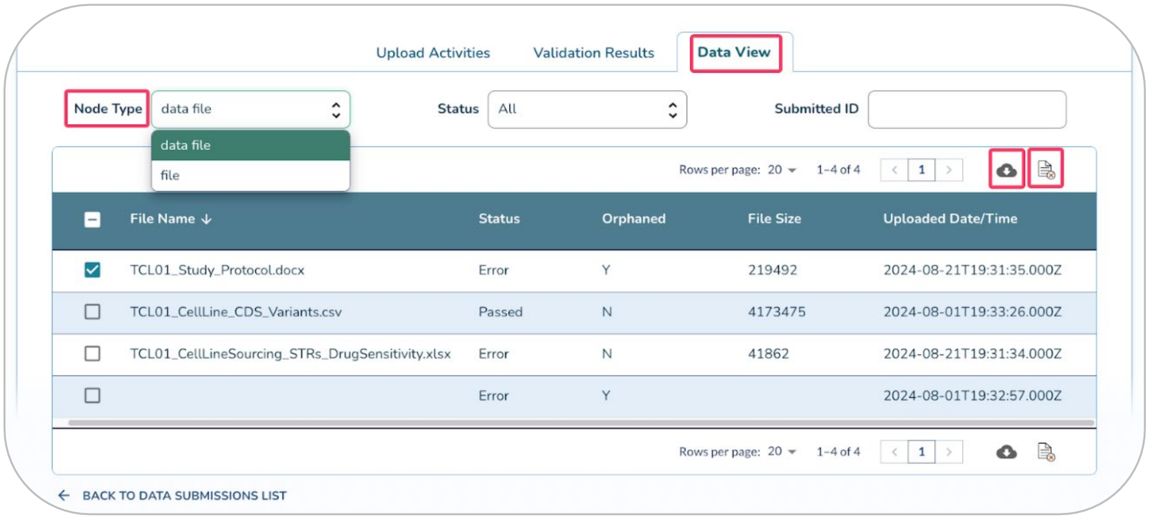
4.2 Correcting Errors

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, users are ***strongly encouraged*** to fix Warnings as well.

4.3 Remove Specific Files

After validating the uploaded data and metadata files, users can view the high-level details of these files in the **Data View** table. To do this, select either "data file" or "file" from the **Node Type** dropdown menu. To remove a specific data file or file ID along with its associated metadata, select the file and click the delete icon. Users can also download the contents displayed in the **Data View** table as a TSV file by choosing the Node Type and clicking the download icon.



Node Type: data file

Status: All

Submitted ID:

Rows per page: 20 1–4 of 4 1

☐	File Name ↓	Status	Orphaned	File Size	Uploaded Date/Time
☑	TCL01_Study_Protocol.docx	Error	Y	219492	2024-08-21T19:31:35.000Z
☐	TCL01_CellLine_CDS_Variants.csv	Passed	N	4173475	2024-08-01T19:33:26.000Z
☐	TCL01_CellLineSourcing_STRs_DrugSensitivity.xlsx	Error	N	41862	2024-08-21T19:31:34.000Z
☐		Error	Y		2024-08-01T19:32:57.000Z

Rows per page: 20 1–4 of 4 1

Figure 24. Removing Specific Files

The contents of the **Data View** table can be interpreted as

- **File Name** - This column displays the name of the uploaded data file.
- **Status** - This indicates the validation status, which can be "Error," "Warning," or "Passed." Details of any errors are available under the Validation Results. Warnings do not halt the submission process.
- **Orphaned** - This column shows whether the data file has associated metadata uploaded in the system. "Y" means the file is orphaned and lacks associated metadata, while "N" means the file has associated metadata uploaded in the system.
- **File Size** - This column lists the file size in bytes.
- **Uploaded Data/Time** - This column records the date and time when the file was uploaded.

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 is 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 assigned member from 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 CRDC Submission team will perform some final checks to make sure everything is as required by the destination Data Commons (for example, CDS, ICDC, CTDC). If those checks pass, the submission will be released to the appropriate CRDC Data Commons and you will receive notification that the Data Commons is now responsible for the next steps. The respective Data Commons will be responsible for indexing and releasing the files that will be made available and accessible on their portal.

If the final checks reveal some unexpected issues, the CRDC Submission team will reach out with additional questions and may reopen the submission to allow additional corrections.