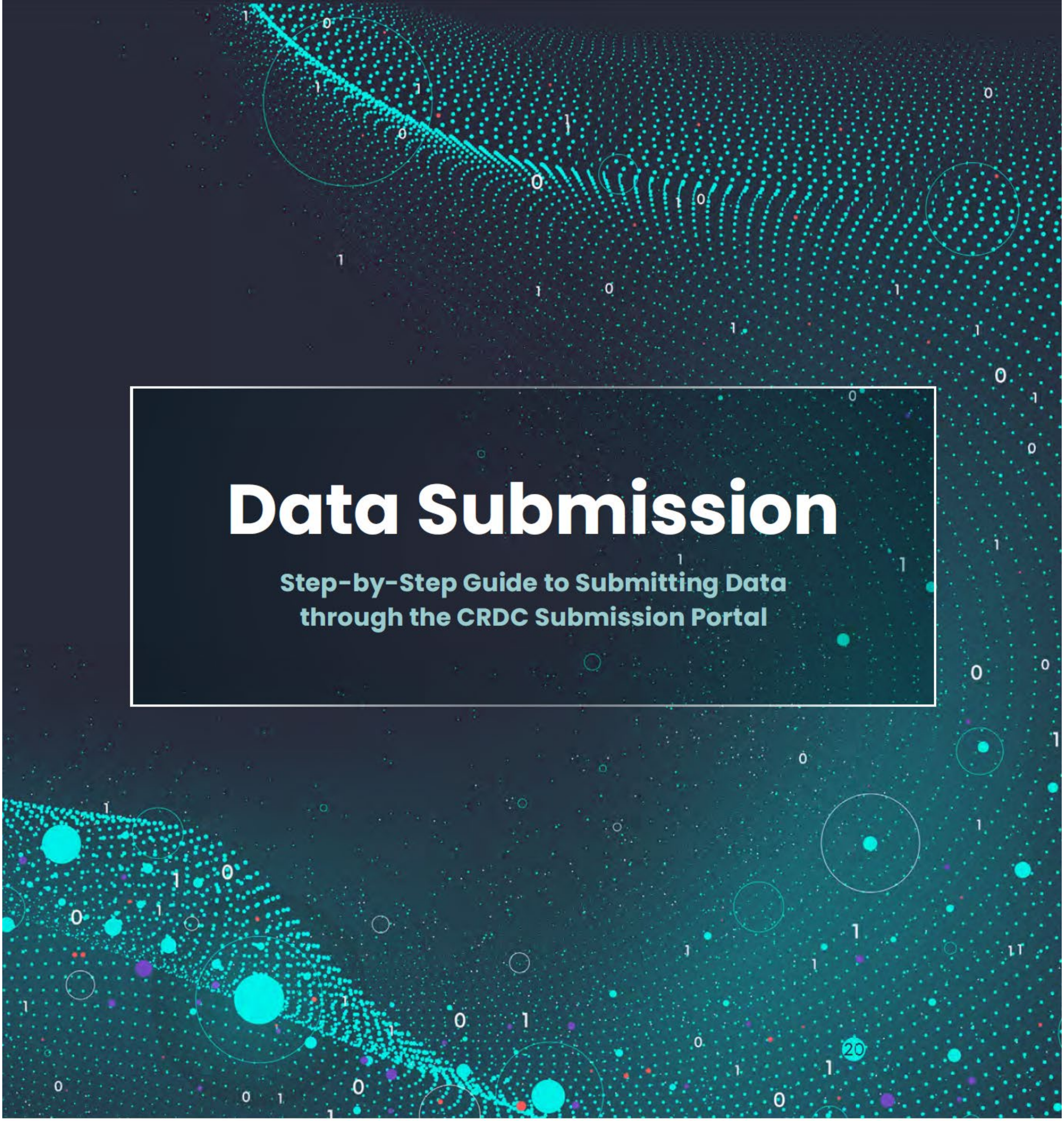


Data Submission

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 study or email the CRDC Help Desk (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 data. Approval notification will be found in 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 the submitter's/user's organization or institution, however, it is not a requirement. Using institutional email as User identity is a preference that can help us figure out the user's organization. Users can choose a personal email as their identifier. NIH staff can log in using their PIV card.
  Note: If you do not login on the CRDC Submission Portal account within 60 days, your access to the portal will be deactivated. To reactivate your access please contact the CRDC Help Desk (NCICRDC@mail.nih.gov).

- **Request Access** to the Submitter role: Users need to get the *Submitter role* assigned to submit data to the CRDC via the Submission Portal. The *Organization role* will allow users to oversee the relevant submissions on the CRDC Submission Portal. Refer to section V on page 3 for more details.

- If the study contains controlled access data, the study must be registered at the database of Genotypes and Phenotypes(dbGaP). dbGaP will provide a dbGaP ID/Accession number upon registration. The portal will not allow the user to begin the data submission process without a dbGaP ID/Accession number The user should email the dbGaP ID associated with their study to the CRDC Help Desk (NCICRDC@mail.nih.gov) or share through the Request Access form (see section V) to initiate the data submission process.

- Additionally, be aware that the CRDC Submission Portal uses 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. See Figure 1.

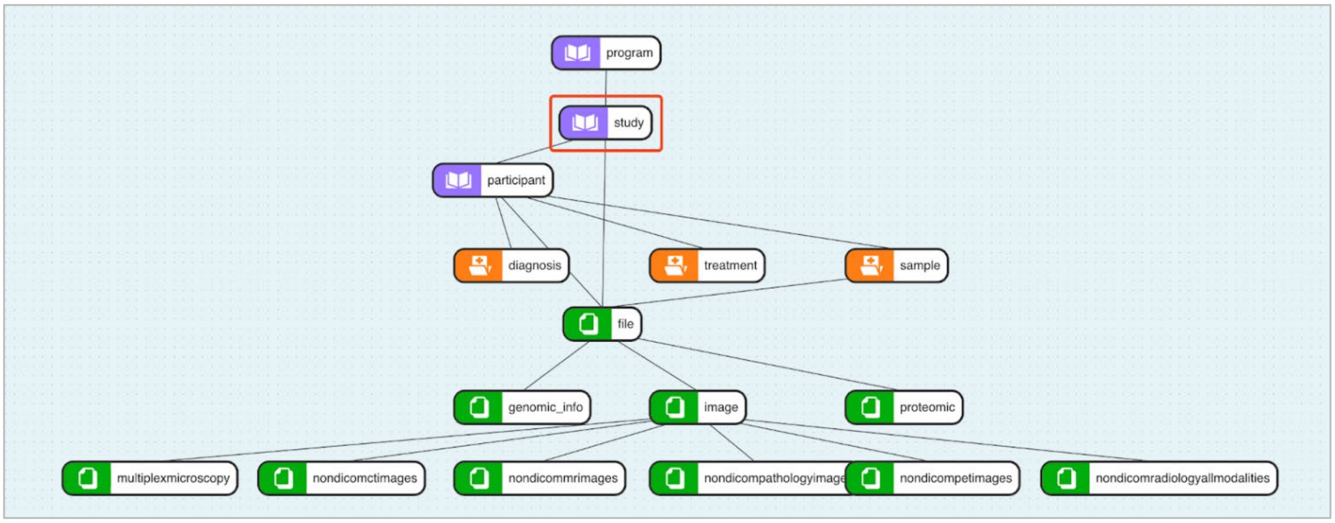


Figure 1. Study Node Relationships in the CDS Data Model

IV. DOCUMENTATION

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

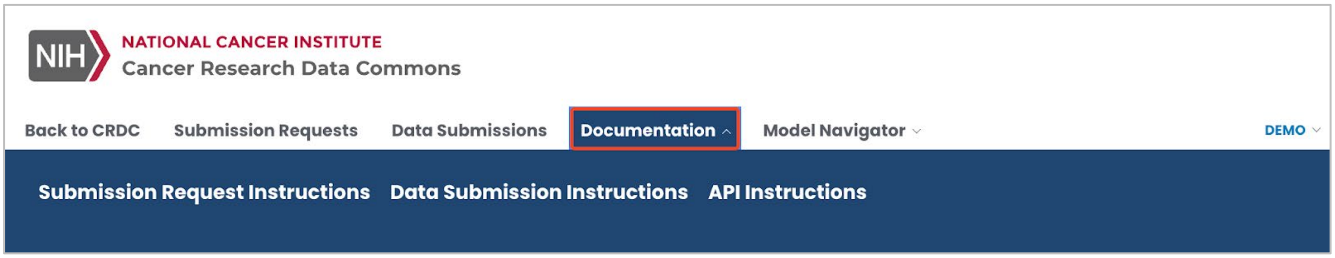
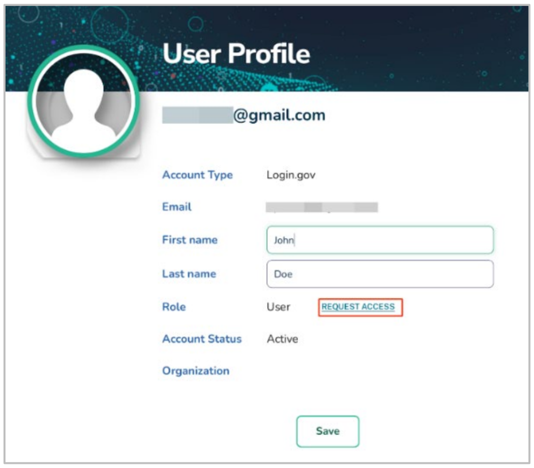

NIH
NATIONAL CANCER INSTITUTE
Cancer Research Data Commons
Back to CRDC
Submission Requests
Data Submissions
Documentation
Model Navigator
DEMO
Submission Request Instructions
Data Submission Instructions
API Instructions

Figure 2. Documentation Menu Showing Available Documents

V. REQUEST PERMISSIONS TO BEGIN A DATA SUBMISSION

Complete the following steps to request permissions to the Submitter role in order begin the data submission process:

1. Log in to the CRDC Submission Portal and go to the User Profile.
2. By default the account is assigned the User role. To request the Submitter or Organization Owner role, click on the **Request Access** button. See Figure 3.



User Profile

[redacted]@gmail.com

Account Type: Login.gov

Email: [redacted]

First name: John

Last name: Doe

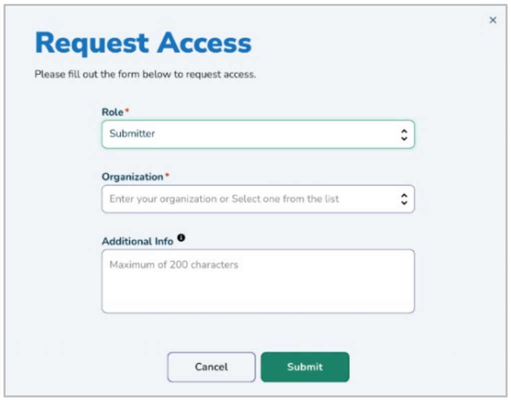
Role: User

Account Status: Active

Organization:

Figure 3. Request Access

3. A **Request Access Form** appears (see Figure 4) and the user needs to provide the required information. From the **Role** dropdown menu, select **Submitter** if you want to begin the data submission process, or **Organization Owner** if you only need to oversee the submissions associated with a specific organization. Then, choose your organization from the list provided. If your organization is not listed, users can manually enter the organization name.



Request Access

Please fill out the form below to request access.

Role*: Submitter

Organization*: Enter your organization or Select one from the list

Additional Info: Maximum of 200 characters

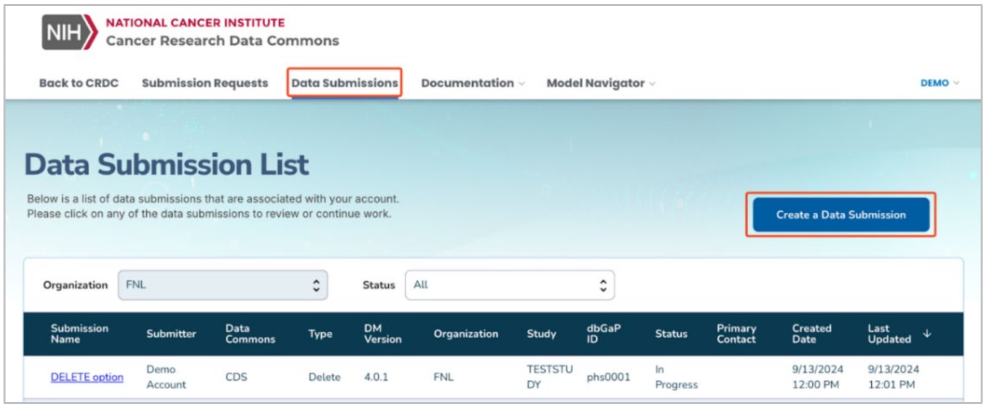
Figure 4. Request Access Form

4. Wait for the System Administrator to grant the requested access. The user will be notified via their login identity.
5. The user granted with **Submitter** role can start the data submission process and the user with **Organization Owner** role can oversee the relevant submissions.

Note: If you do not login on the CRDC Submission Portal account within 60 days, your access to the portal will be deactivated. If so, please contact the CRDC Help Desk to reactivate your account.

VI. STARTING A NEW SUBMISSION

Once logged in with a Submitter role, navigate to the Data Submissions tab. The Data Submissions List page loads. If this is the Submitter's first data submission, the table showing the list of Data Submissions will be empty. If the Submitter has multiple submissions, the filters at the top of the table can be used to narrow down the list. See Figure 5.



Data Submission List

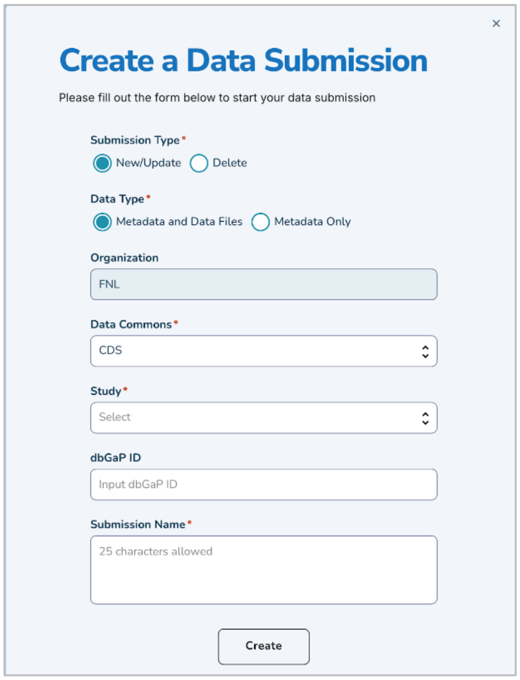
Below is a list of data submissions that are associated with your account.
Please click on any of the data submissions to review or continue work.

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 5. Create a Data Submission

To start a new data submission, click on the **Create a Data Submission** button and a dialog box as shown in Figure 6 will appear. Fill out all the requested and required information as described below.



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 6. 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. The deletion request will go through a validation process by the CRDC Team, and once approved, the deletion can be processed.

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 it is not preselected already. If you are trying to submit to another CRDC Data Commons not already listed, email the CRDC Help Desk at NCICRDC@mail.nih.gov.

4. The **Study** dropdown 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 Help Desk.

5. If your study contains controlled access data, enter the **dbGaP ID** /accession number will be pre-populated if it was entered in the Submission Request Form or emailed to CRDC Help Desk. The system will not allow the Submitter to initiate a data submission without providing the dbGaP ID.

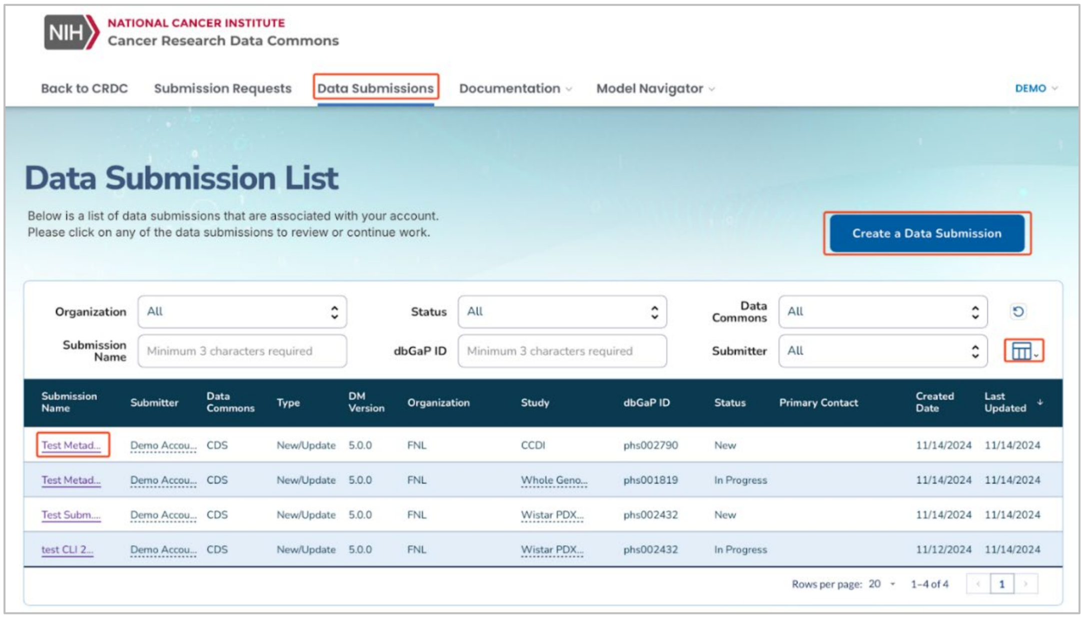
6. Optionally, enter a **Submission Name**, which is a free-text field available to users to label their submissions. This label will appear on the Submissions List table, which is accessible when you click the **Create** button.

7. The new submission is created by clicking the **Create** button.

   Once a submission is created, CRDC assigns a Primary Contact to your submission. You can find the email address of the CRDC Data Submission team member assigned to your submission at the top of the new submission page. From this step onwards, all questions related to the data submission should be directed to the assigned Primary Contact.

VII. 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. If the submission name exceeds 10 characters, hover over the name to view a pop-up with the full submission name (see Figure 7). Additionally, users can customize the columns displayed in the submission table by clicking the Table icon in the top-right corner, selecting the desired columns, and applying the changes.



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
Data Commons: All
Submission Name: Minimum 3 characters required
dbGaP ID: Minimum 3 characters required
Submitter: All

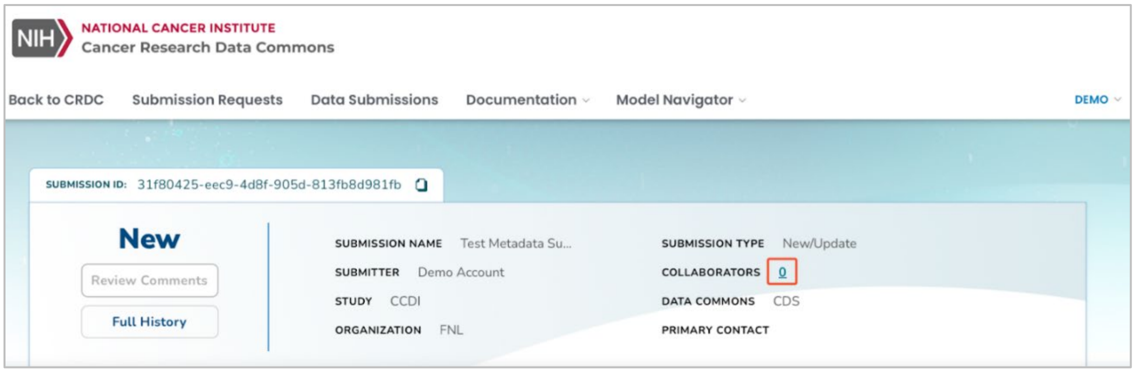
Submission Name	Submitter	Data Commons	Type	DM Version	Organization	Study	dbGaP ID	Status	Primary Contact	Created Date	Last Updated
Test Metad...	Demo Accou...	CDS	New/Update	5.0.0	FNL	CCDI	phs002790	New		11/14/2024	11/14/2024
Test Metad...	Demo Accou...	CDS	New/Update	5.0.0	FNL	Whole Geno...	phs001819	In Progress		11/14/2024	11/14/2024
Test Subm....	Demo Accou...	CDS	New/Update	5.0.0	FNL	Wistar PDX...	phs002432	New		11/14/2024	11/14/2024
test CLI 2...	Demo Accou...	CDS	New/Update	5.0.0	FNL	Wistar PDX...	phs002432	In Progress		11/12/2024	11/14/2024

Rows per page: 20 1–4 of 4 1

Figure 7. Active Data Submissions List

VIII. ADDING AND MANAGING COLLABORATORS

The submitter can add collaborators to a given submission and manage their access. The collaborators can upload and validate data on the Submission Portal. By default, the collaborator count is set to zero, and as collaborators are added, the count will be updated on the dashboard. See Figure 8.



SUBMISSION ID: 31f80425-eec9-4d8f-905d-813fb8d981fb

New

SUBMISSION NAME Test Metadata Su...
SUBMITTER Demo Account
STUDY CCDI
ORGANIZATION FNL
SUBMISSION TYPE New/Update
COLLABORATORS 0
DATA COMMONS CDS
PRIMARY CONTACT

Figure 8. Count of Collaborators on the Data Submission Dashboard

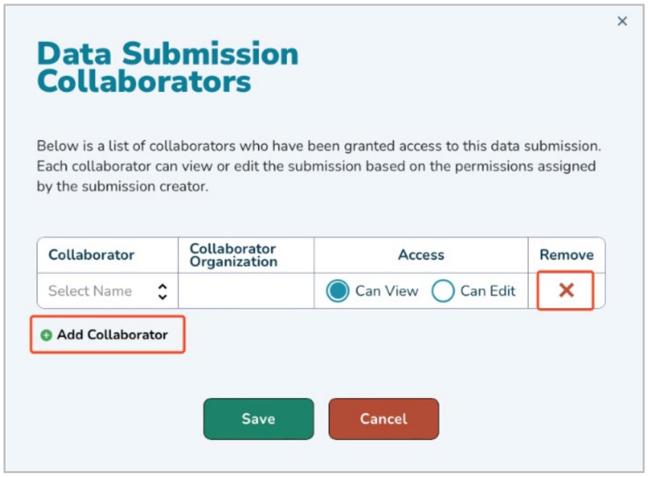
To add and manage collaborators, click the hyperlink next to **Collaborators** on the Submission dashboard, which will open a new Data submission Collaborators window. See Figure 9.

In the Data Submission Collaborators window, the submitter can select collaborator/s by selecting names from the drop-down menu under **Collaborator**. The collaborator's organization will get pre-populated under the **Collaborator Organization** field. Please ensure that the collaborator has an authorized account on the CRDC Submission Portal with the Submitter role and is affiliated to the same study.

In the **Access** column, the submitter can assign either **Read Only** or **Edit** permission to the collaborator:

- **Can Edit** permission allows the Collaborator to upload and validate the data on the portal and
- **Can View** permission allows the Collaborator to oversee the submission process on the portal.

The submitter can remove a collaborator by simply clicking the remove icon, add multiple collaborators by clicking the **Add Collaborator** button, and repeating the process for each additional collaborator. Be sure to click **Save** before closing the window to retain any changes.



Data Submission Collaborators

Below is a list of collaborators who have been granted access to this data submission. Each collaborator can view or edit the submission based on the permissions assigned by the submission creator.

Collaborator	Collaborator Organization	Access	Remove
Select Name		● Can View ○ Can Edit	✕

Figure 9. Add and Manage Collaborators for a Given Submission

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 10) 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. Models of other Data Commons will be added as and when they are integrated with CRDC Submission Portal.

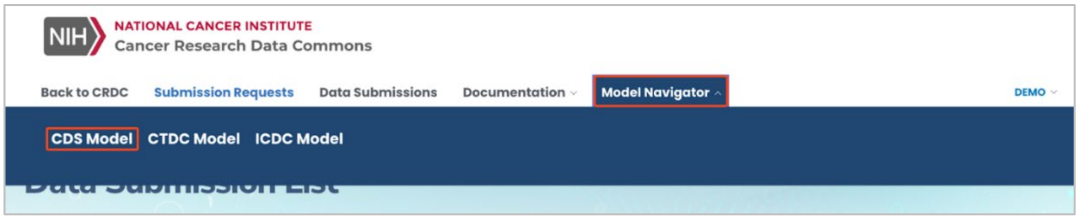


Figure 10. Use the Menu Bar to Navigate to the Data Model Viewer

Once you select the Data Model, the Data Model Viewer appears in the next screen, as seen in Figure 11. On this page, you can view the data model in detail and download the submission templates.

Note: Figure 11 shows the CDS Data Model as an example.

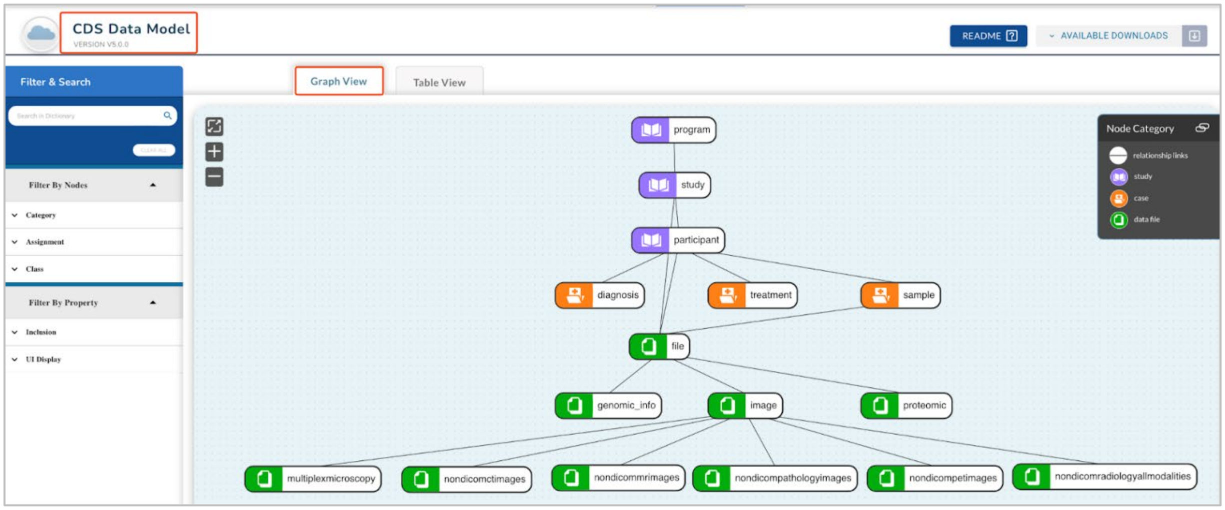
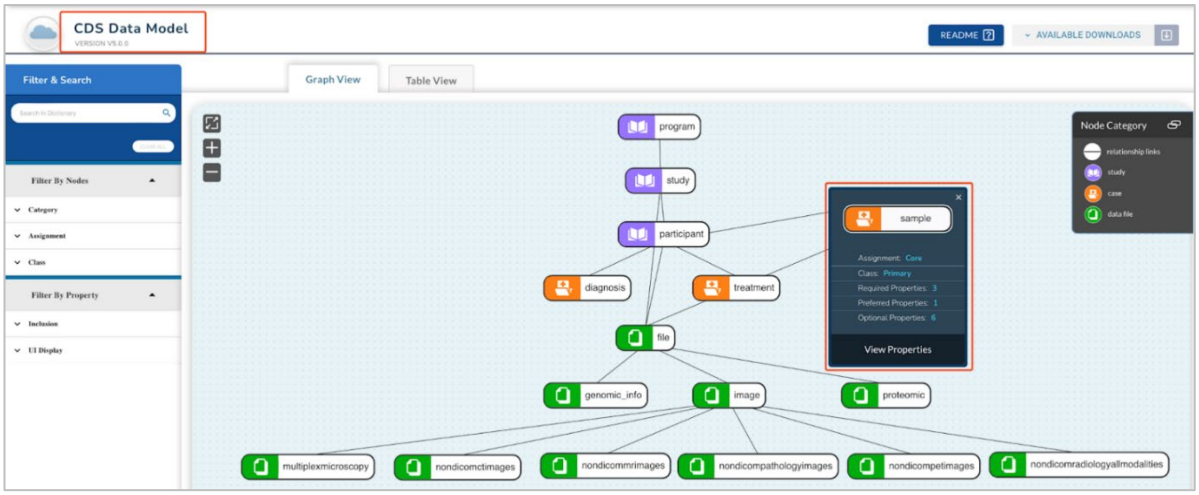


Figure 11. Data Model Viewer Graph View

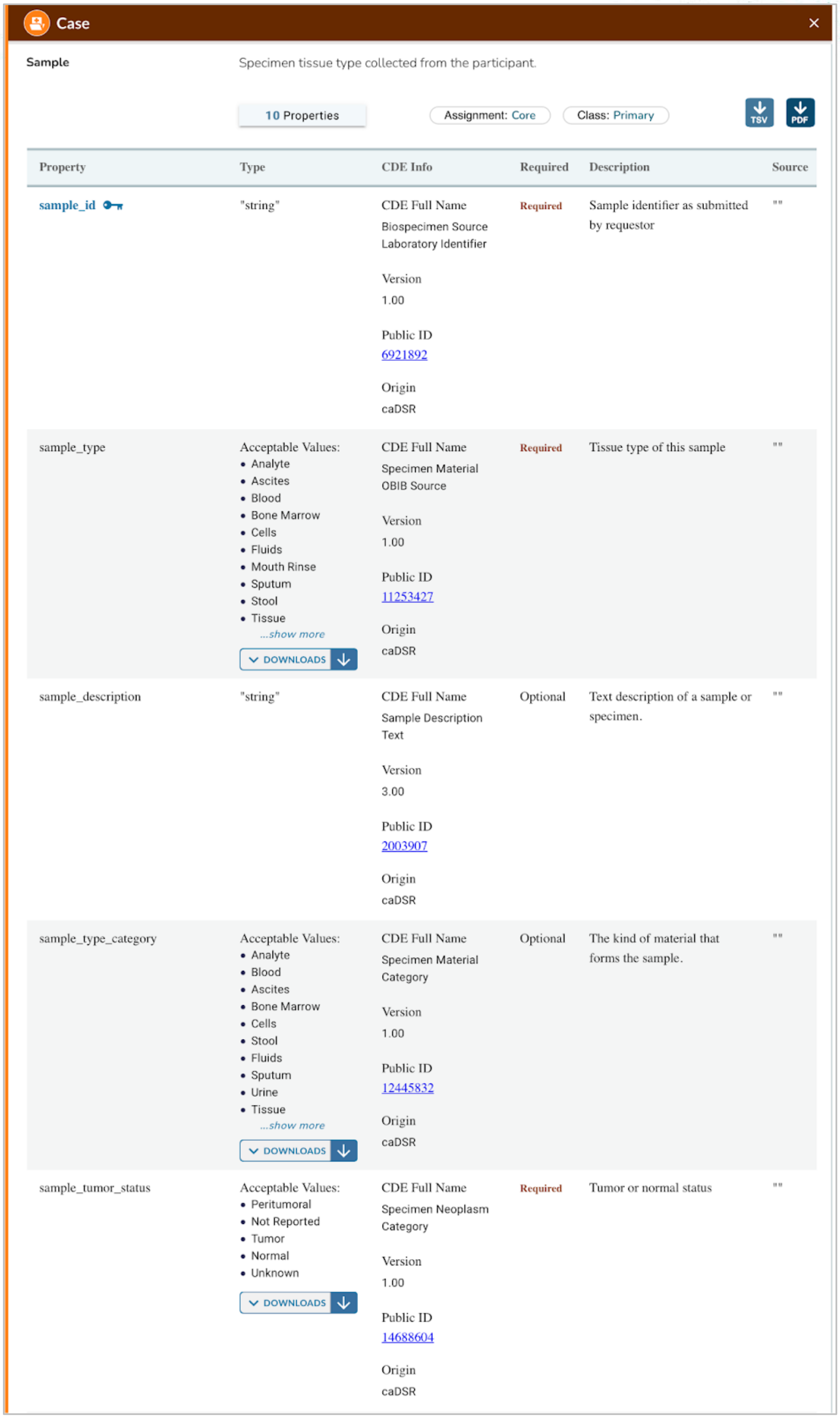
Use the Data Model Viewer to explore the data elements that submissions require or can accept. On the **Graph View** tab, click a node in the graph to open its summary. At the bottom of the node summary, click the **View Properties** menu to open a table view of the selected node.



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

Figure 12. Click a Node to View a Summary and Open the Table View

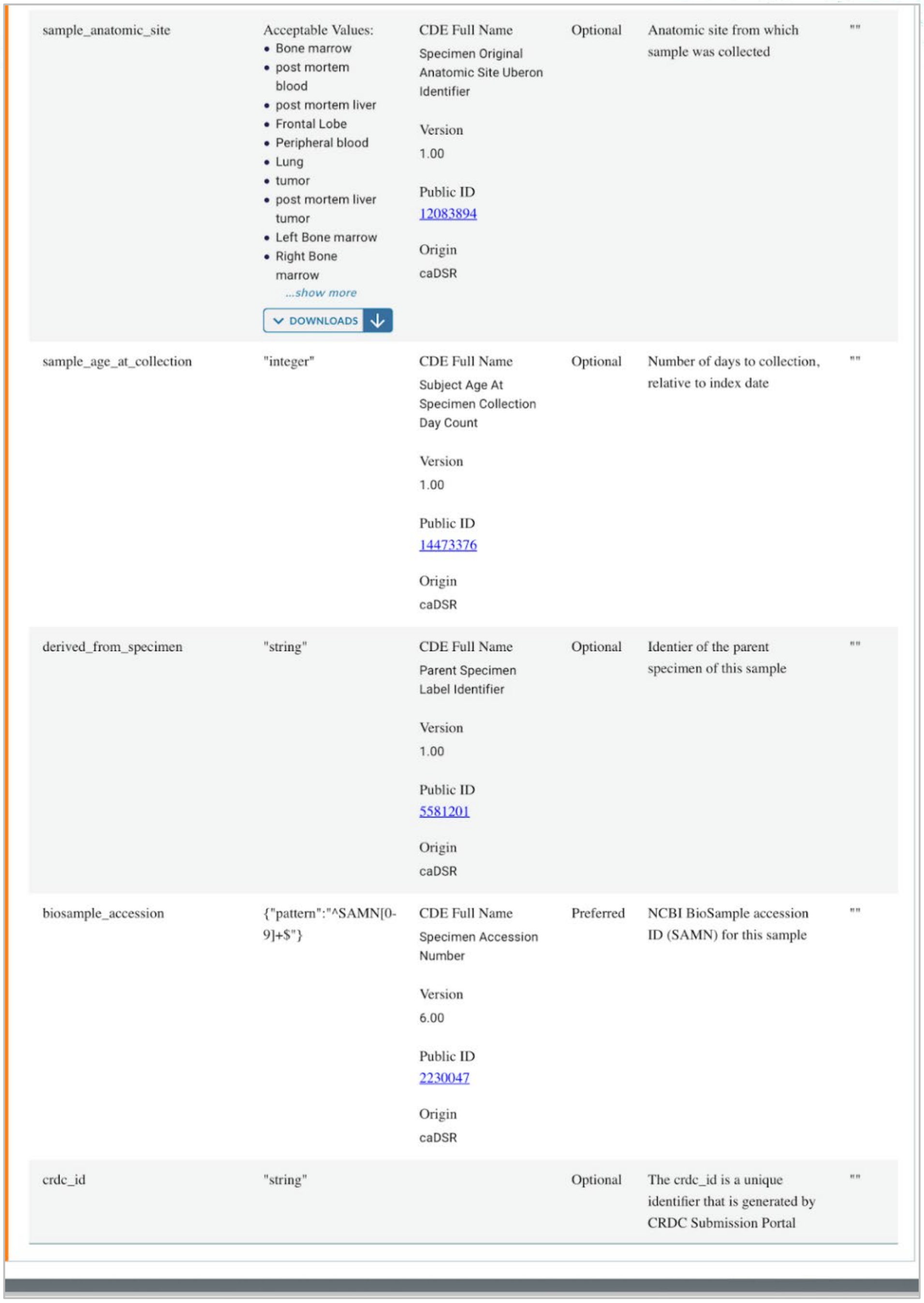
The table view includes the description of Properties that are expected in that field (such as strings, integers, etc.) and which fields are required. See Figure 13.



Sample Specimen tissue type collected from the participant.

Property	Type	CDE Info	Required	Description	Source
sample_id	"string"	CDE Full Name Biospecimen Source Laboratory Identifier Version 1.00 Public ID 6921892 Origin caDSR	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	CDE Full Name Specimen Material OBIB Source Version 1.00 Public ID 11253427 Origin caDSR	Required	Tissue type of this sample	""
sample_description	"string"	CDE Full Name Sample Description Text Version 3.00 Public ID 2003907 Origin caDSR	Optional	Text description of a sample or specimen.	""
sample_type_category	Acceptable Values: • Analyte • Blood • Ascites • Bone Marrow • Cells • Stool • Fluids • Sputum • Urine • Tissue ...show more DOWNLOADS	CDE Full Name Specimen Material Category Version 1.00 Public ID 12445832 Origin caDSR	Optional	The kind of material that forms the sample.	""
sample_tumor_status	Acceptable Values: • Peritumoral • Not Reported • Tumor • Normal • Unknown DOWNLOADS	CDE Full Name Specimen Neoplasm Category Version 1.00 Public ID 14688604 Origin caDSR	Required	Tumor or normal status	""

Figure 13. Table View of a Sample Node



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	CDE Full Name Specimen Original Anatomic Site Uberon Identifier Version 1.00 Public ID 12083894 Origin caDSR	Optional	Anatomic site from which sample was collected	""
sample_age_at_collection	"integer"	CDE Full Name Subject Age At Specimen Collection Day Count Version 1.00 Public ID 14473376 Origin caDSR	Optional	Number of days to collection, relative to index date	""
derived_from_specimen	"string"	CDE Full Name Parent Specimen Label Identifier Version 1.00 Public ID 5581201 Origin caDSR	Optional	Identier of the parent specimen of this sample	""
biosample_accession	{"pattern":"^SAMN[0-9]+$"}	CDE Full Name Specimen Accession Number Version 6.00 Public ID 2230047 Origin caDSR	Preferred	NCBI BioSample accession ID (SAMN) for this sample	""
crdc_id	"string"		Optional	The crdc_id is a unique identifier that is generated by CRDC Submission Portal	""

Figure 13. Table View of a Sample Node (continued)

2. Downloading Data Dictionary and Submission Templates

To download the Data Dictionary and the submission templates, select **Available Downloads**, as shown in Figure 14. Then select one of the following files from the dropdown menu and click the download arrow to start the download. CRDC Vocabularies for the selected Data Commons Model and Examples Templates can also be downloaded.

- **Data Dictionary (PDF)**
- **Submission Templates (TSV)**
- **All Vocabularies (TSV)**
- **All Vocabularies (JSON)**
- **Example Templates (TSV)**

Submitters can use the **Submission Templates** to format and upload metadata in the CRDC Submission Portal. These templates ***must*** be in TSV format. Templates in any other format such as Microsoft Excel format (.xls, .xlsx) etc. will fail. Software like ModernCSV can be used to work with these Submission Templates, as it handles CSV and TSV as tables without automatically modifying the data

The **Data Dictionary** document provides detailed information about the metadata structure, content, and 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. Each Data Commons has its own unique Data Model with corresponding data dictionary, set of nodes, properties, and submission templates The **Example Templates** are examples of completed submission templates with mock data, 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.

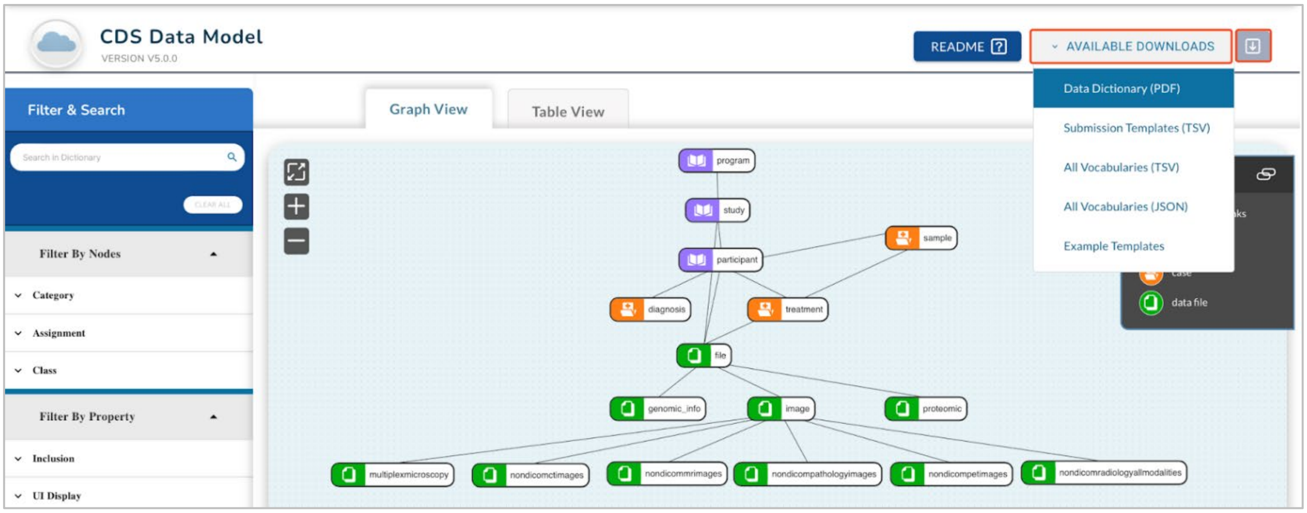
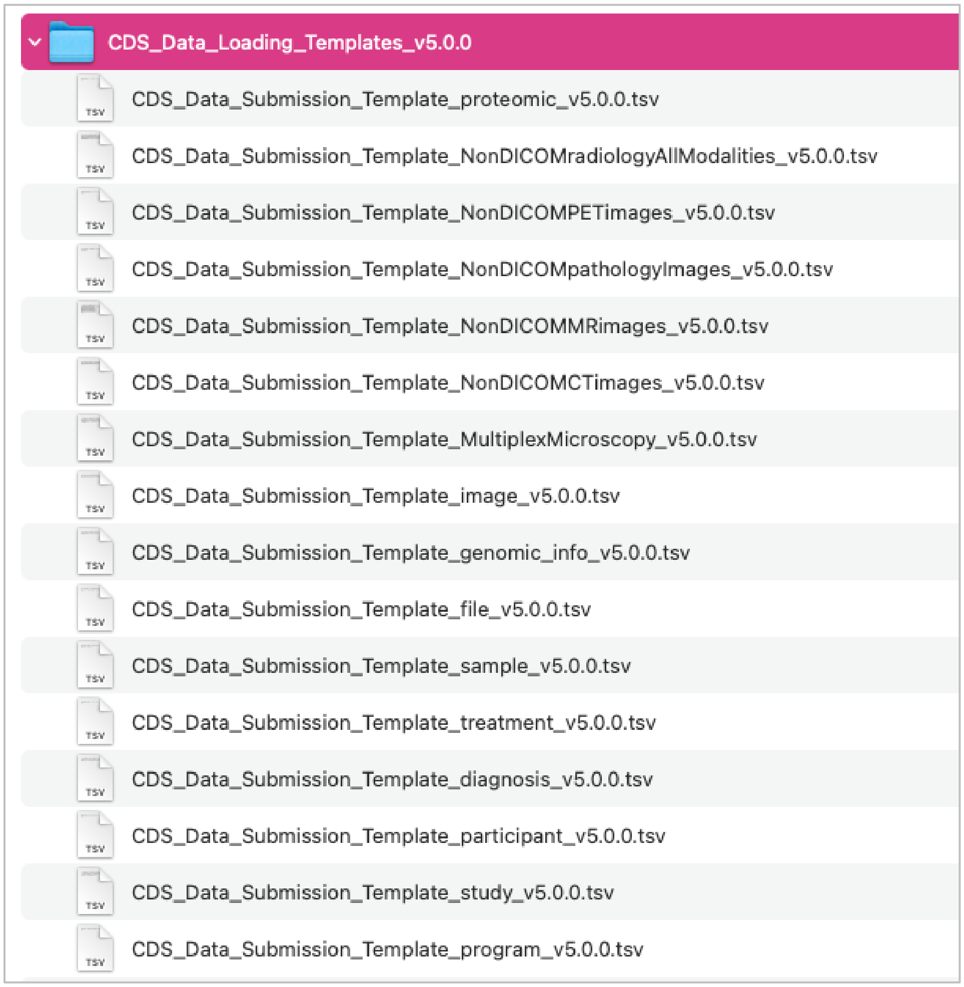


Figure 14. Using the Download Menu

The downloaded files are provided as a zip archive. 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 15.

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



CDS_Data_Loading_Templates_v5.0.0

- CDS_Data_Submission_Template_proteomic_v5.0.0.tsv
- CDS_Data_Submission_Template_NonDICOMradiologyAllModalities_v5.0.0.tsv
- CDS_Data_Submission_Template_NonDICOMPETImages_v5.0.0.tsv
- CDS_Data_Submission_Template_NonDICOMpathologyImages_v5.0.0.tsv
- CDS_Data_Submission_Template_NonDICOMMRimages_v5.0.0.tsv
- CDS_Data_Submission_Template_NonDICOMCTimages_v5.0.0.tsv
- CDS_Data_Submission_Template_MultiplexMicroscopy_v5.0.0.tsv
- CDS_Data_Submission_Template_image_v5.0.0.tsv
- CDS_Data_Submission_Template_genomic_info_v5.0.0.tsv
- CDS_Data_Submission_Template_file_v5.0.0.tsv
- CDS_Data_Submission_Template_sample_v5.0.0.tsv
- CDS_Data_Submission_Template_treatment_v5.0.0.tsv
- CDS_Data_Submission_Template_diagnosis_v5.0.0.tsv
- CDS_Data_Submission_Template_participant_v5.0.0.tsv
- CDS_Data_Submission_Template_study_v5.0.0.tsv
- CDS_Data_Submission_Template_program_v5.0.0.tsv

Figure 15. 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 node in the model; 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 submitter does not need to fill or submit the imaging submission template.

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 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 in the first column is prefilled with the correct node name for that specific template (e.g., in the CDS_Data_Submission_Template_study_v5.0.0.tsv template, the second row in the first column is filled in as ‘Study.’). 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 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.

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 (CLI) for uploading data to its temporary CRDC storage. You can install and use the CLI 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. However, submitters should ensure that they are using the latest version of the CLI tool and the configuration file.

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. See Figure 16.

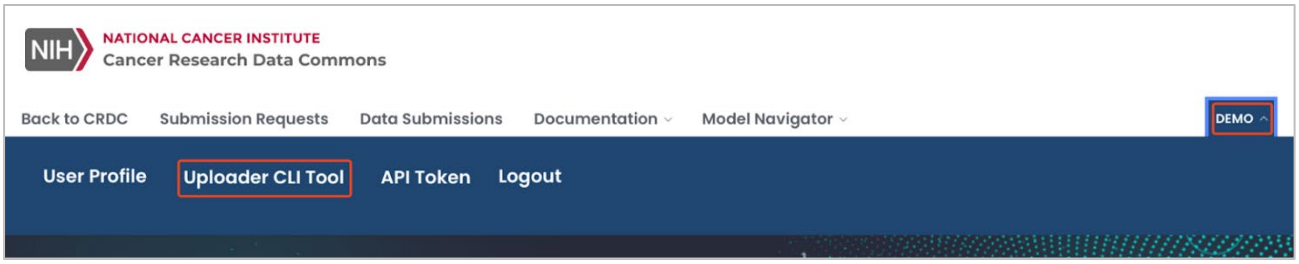
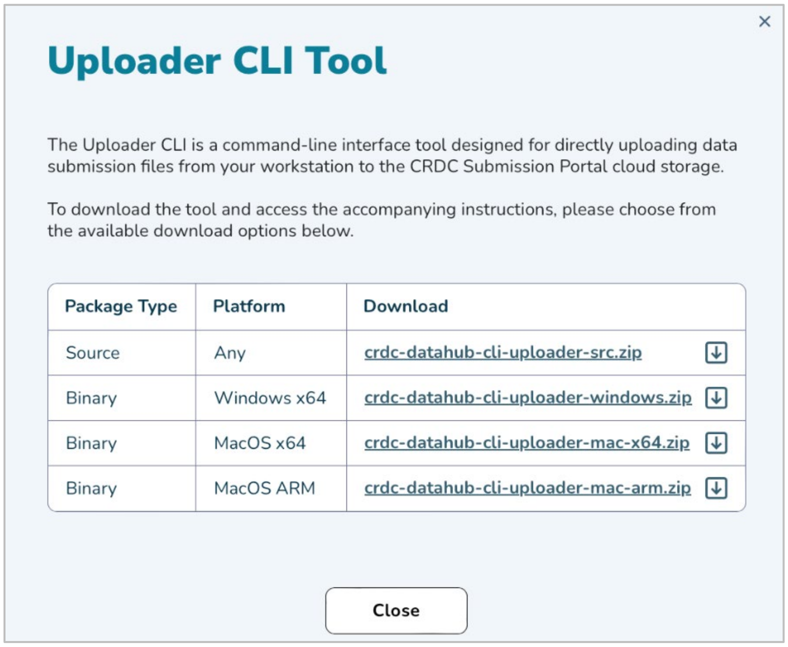


Figure 16. Menu with the Uploader CLI Tool Download Option

Click the **Download** icon next to the available options to download the CLI tool and its accompanying instructions (see Figure 17). A ZIP archive will be saved to your local machine.



Uploader CLI Tool

The Uploader CLI is a command-line interface tool designed for directly uploading data submission files from your workstation to the CRDC Submission Portal cloud storage.

To download the tool and access the accompanying instructions, please choose from the available download options below.

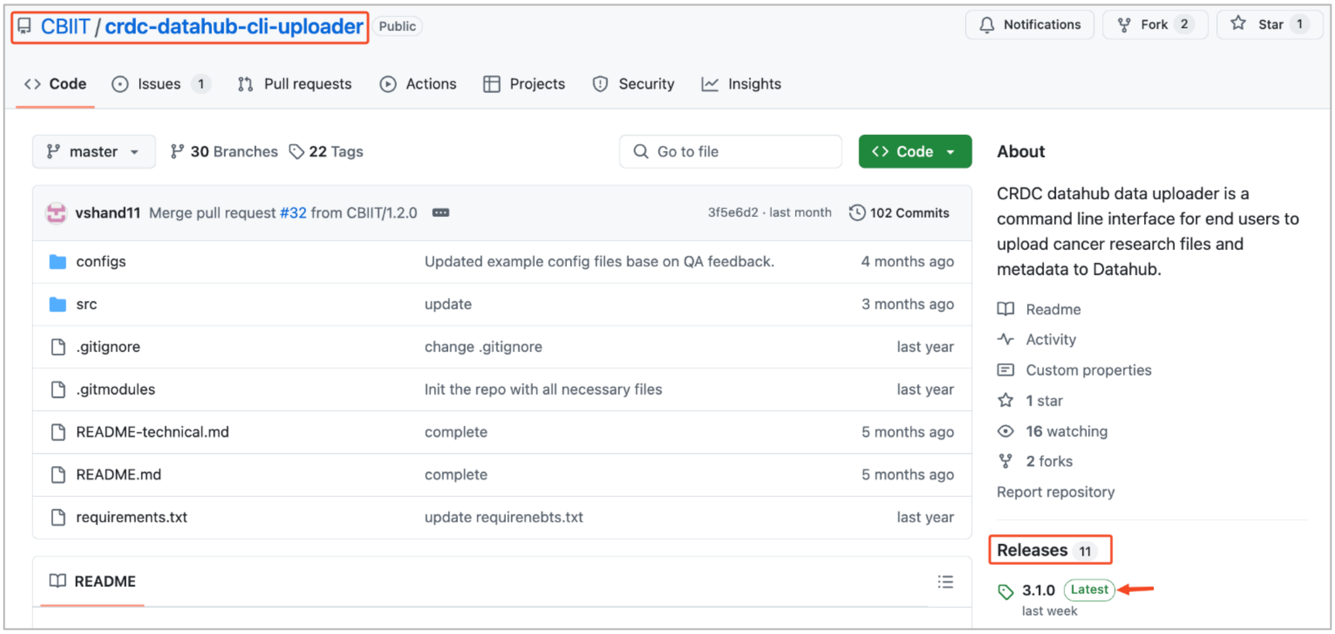
Package Type	Platform	Download
Source	Any	crdc-datahub-cli-uploader-src.zip
Binary	Windows x64	crdc-datahub-cli-uploader-windows.zip
Binary	MacOS x64	crdc-datahub-cli-uploader-mac-x64.zip
Binary	MacOS ARM	crdc-datahub-cli-uploader-mac-arm.zip

Figure 17. 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
```



vshand11 Merge pull request #32 from CBIIT/1.2.0 3f5e6d2 · last month 102 Commits

configs	Updated example config files base on QA feedback.	4 months ago
src	update	3 months ago
.gitignore	change .gitignore	last year
.gitmodules	Init the repo with all necessary files	last year
README-technical.md	complete	5 months ago
README.md	complete	5 months ago
requirements.txt	update requirenebts.txt	last year

About

CRDC datahub data uploader is a command line interface for end users to upload cancer research files and metadata to Datahub.

Figure 18. Uploader CLI Tool as it Appears in GitHub

3.3 Setting Up the Python Environment

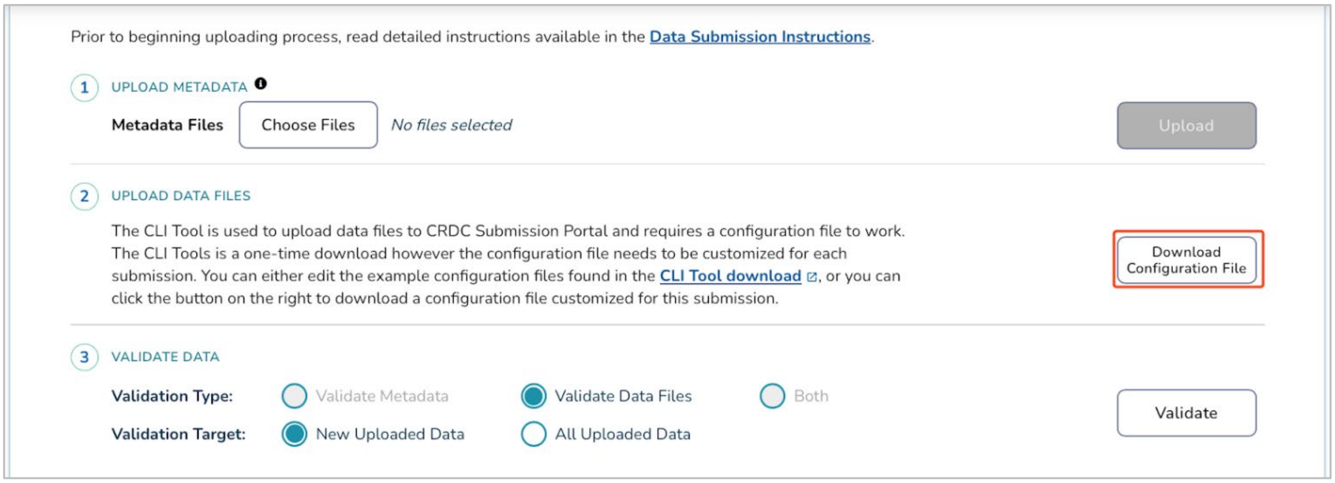
The Uploader CLI Tool has Python library dependencies that you must install before running the CLI. These dependencies can be installed by running the command `pip3 install -r requirements.txt`. The `requirements.txt` contains the list of dependencies, described below. If you want to install the dependencies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

- 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. See Figure 20. Enter the path to the local folder containing your Data Files and the local path to the Metadata 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]

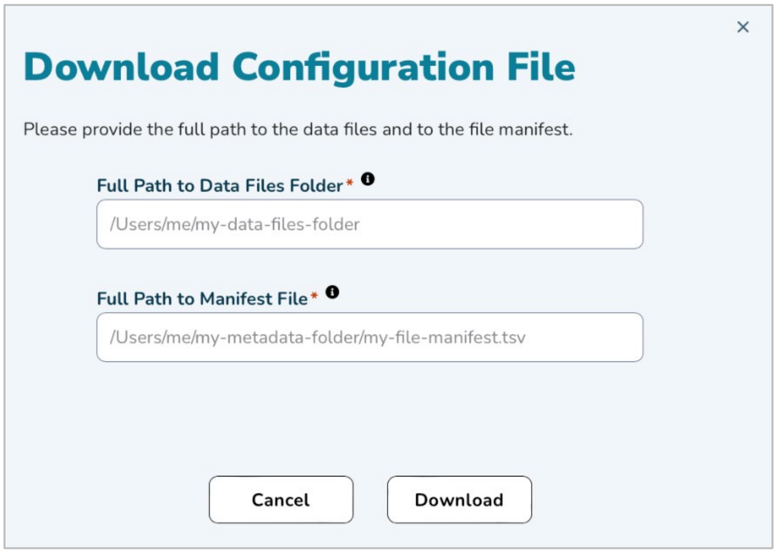
3 VALIDATE DATA

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

Validation Target: ● New Uploaded Data ○ All Uploaded Data

[Validate]

Figure 19. Download Configuration File



Download Configuration File

Please provide the full path to the data files and to the file manifest.

Full Path to Data Files Folder*

/Users/me/my-data-files-folder

Full Path to Manifest File*

/Users/me/my-metadata-folder/my-file-manifest.tsv

Figure 20. Dependencies to Download Configuration File

Additionally, if you need to populate the configuration file 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 file 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 study that the uploaded files will be associated with. To find the correct submission ID, log into the system and select the study from the Data Submission List 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 study consists of one or more submissions (often many more), with each Submission ID linked to the parent study. A single user working on multiple studies must carefully track which Submission IDs they are uploading to ensure the data is associated with the correct study.
- **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.
- **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 and size exists in the CRDC 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 document called a File Manifest to upload the data files to the temporary CRDC storage. This File Manifest is a simple table (a TSV file) with all the required properties as defined by the Data Model except the File IDs as the IDs are generated by the CLI tool. Submitters can use the file.tsv template from the Data Model viewer to create this File Manifest, saving the effort of creating a duplicate file.

The file.tsv template downloaded from Data Model viewer does not include a column for file IDs/Keys because the CLI Tool generates those IDs. Once all the data files are uploaded, the CLI Tool creates a “final” version of the File Manifest (e.g. file-final.tsv) and saves it in the same location of the original File Manifest. The Uploader Tool will automatically upload this final manifest for you, so you don’t need to upload it manually to the CRDC portal.

If you need to make changes to the File Manifest, be sure to edit the final version before uploading it to the portal. If you need to re-upload your data files for any reason, you can reuse the final manifest provided by the tool. In that case, it will use the existing file IDs/Keys instead of creating 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. Also, the following commands assume your current directory is in the unzipped CLI directory.

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

When running Windows version, the command should look like the following:

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

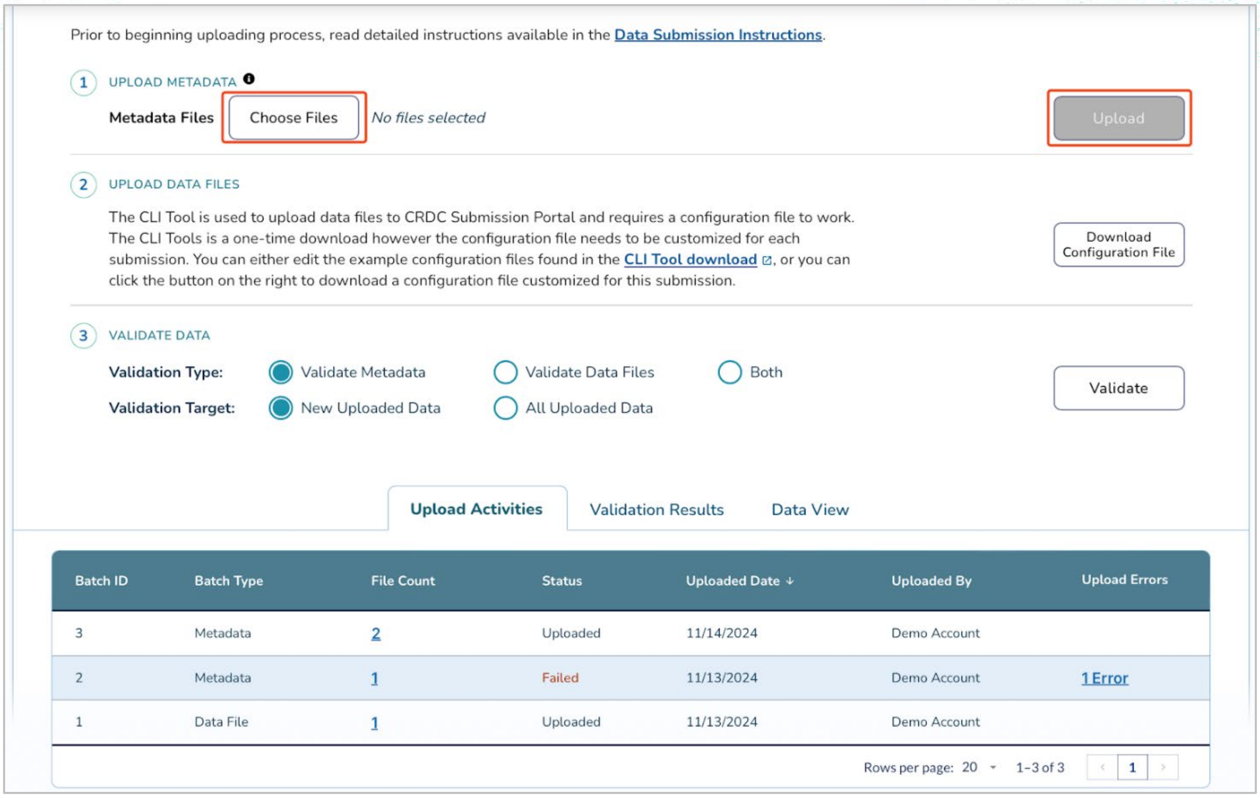
When running Mac version, the command should be:

```
   $ ./uploader --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 22.

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: *No files selected*

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.

3 VALIDATE DATA

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

Validation Target: ● New Uploaded Data ○ All Uploaded Data

Batch ID	Batch Type	File Count	Status	Uploaded Date ↓	Uploaded By	Upload Errors
3	Metadata	2	Uploaded	11/14/2024	Demo Account	
2	Metadata	1	Failed	11/13/2024	Demo Account	1 Error
1	Data File	1	Uploaded	11/13/2024	Demo Account	

Rows per page: 20 1–3 of 3 1

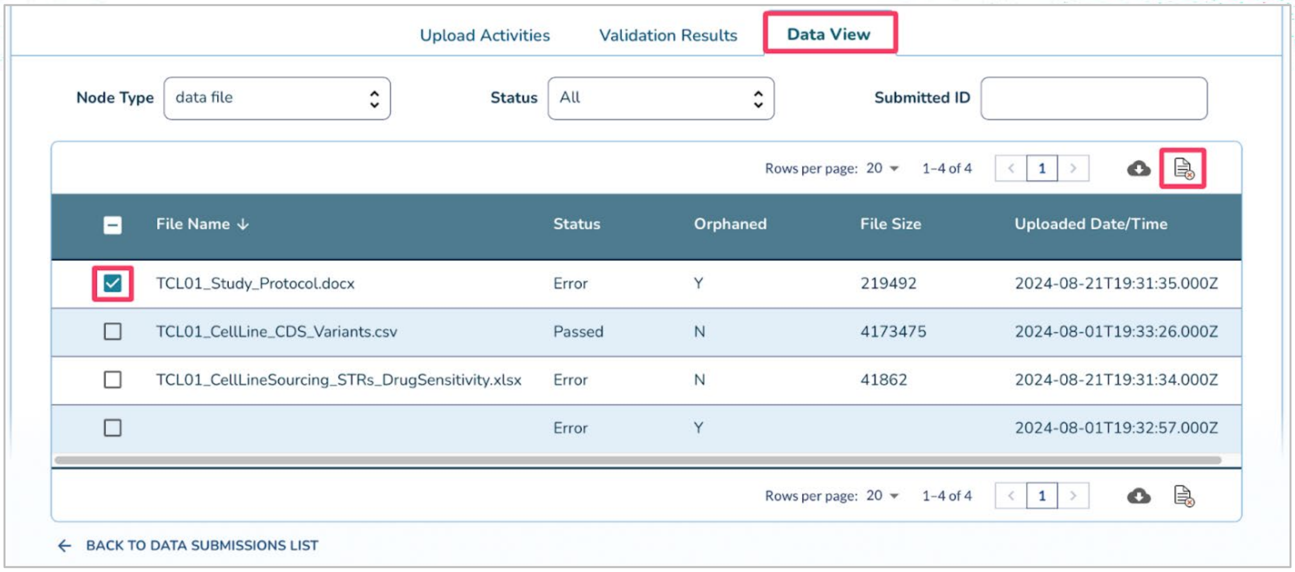
Figure 21. Upload and Validate Data and Metadata

As depicted in Figure 22, to start the upload process, click the ***Choose Files*** button and then select the metadata Submission Manifests 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 the primary validation checks, the data is not uploaded and 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 22. 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.



Node Type: data file
Status: All
Submitted ID:

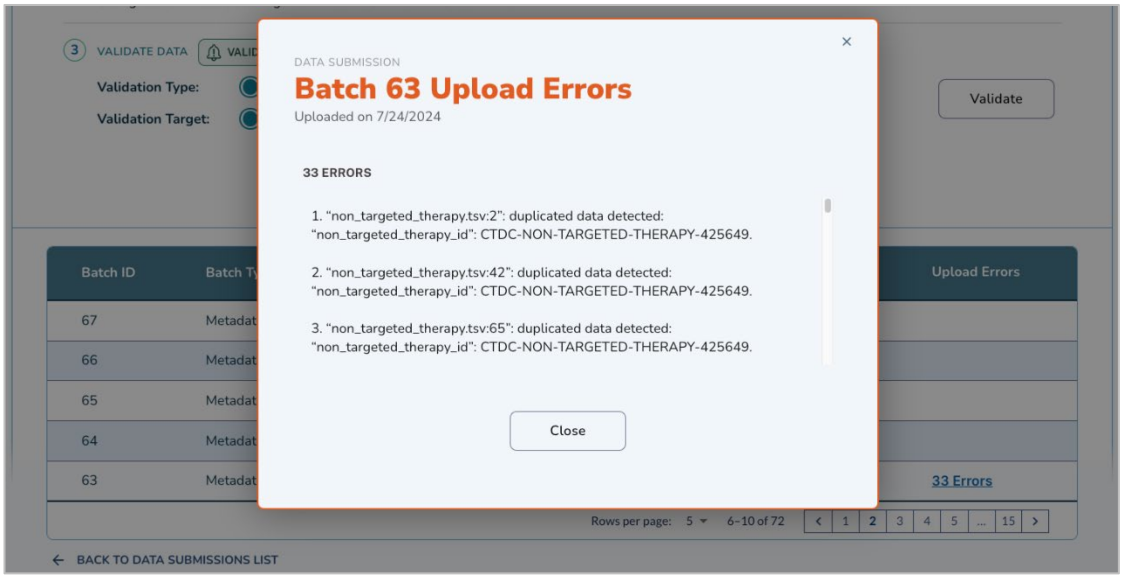
Rows per page: 20 1–4 of 4

☐	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

Figure 22. 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 23.



DATA SUBMISSION

Batch 63 Upload Errors

Uploaded on 7/24/2024

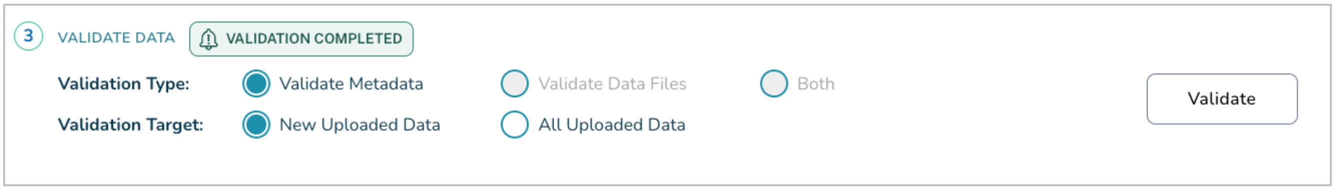
33 ERRORS

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.

Figure 23. 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 24.



3 VALIDATE DATA VALIDATION COMPLETED

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

Validation Target: ● New Uploaded Data ○ All Uploaded Data

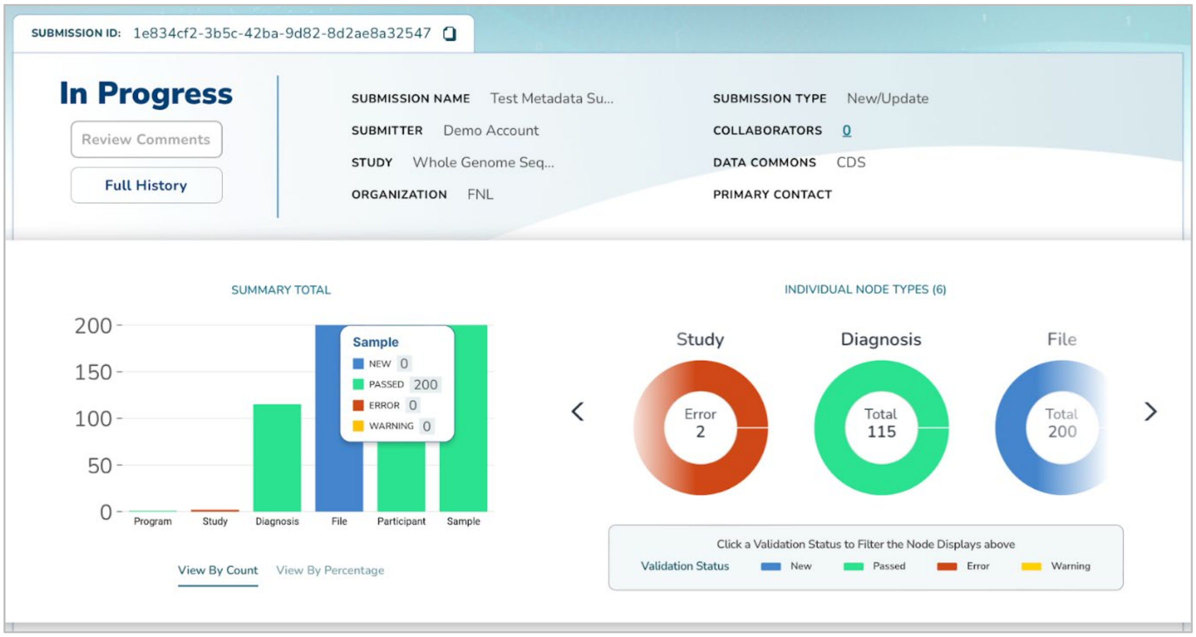
Figure 24. Validate Data Options

The first step is selecting which files to validate. The **Validate Metadata** option runs validations only on the Submission Metadata Manifest, 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 25.



SUBMISSION ID: 1e834cf2-3b5c-42ba-9d82-8d2ae8a32547

In Progress

SUBMISSION NAME Test Metadata Su...

SUBMITTER Demo Account

STUDY Whole Genome Seq...

ORGANIZATION FNL

SUBMISSION TYPE New/Update

COLLABORATORS 0

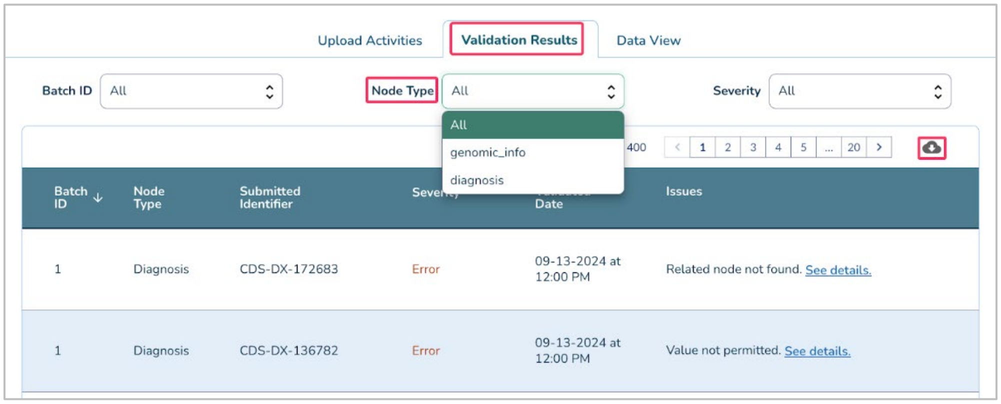
DATA COMMONS CDS

PRIMARY CONTACT

Figure 25. Validation Summary

The left graph in Figure 26 displays a list of the nodes and indicates the status of the uploaded and validated data, with green representing data that passed and red representing data that failed the validations. Hovering over each bar generates a more detailed summary for that node. For instance, in Figure 26, the file node displays 200 uploaded files, all of which have successfully passed validation. The blue color indicates data that has been uploaded but has not yet been validated. 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 **Al** 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.

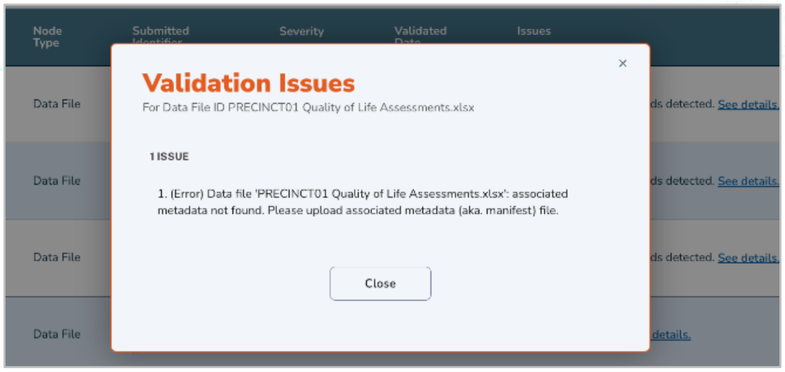


Batch ID	Node Type	Submitted Identifier	Severity	Date	Issues
1	Diagnosis	CDS-DX-172683	Error	09-13-2024 at 12:00 PM	Related node not found. See details.
1	Diagnosis	CDS-DX-136782	Error	09-13-2024 at 12:00 PM	Value not permitted. See details.

Figure 26. 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 27 presents an example of 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 27. Validation Error Details

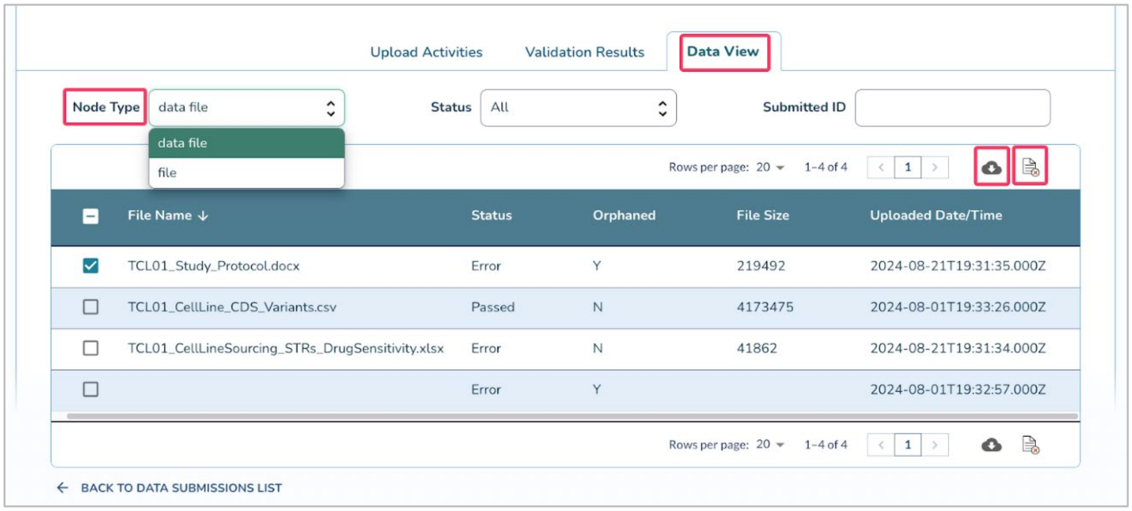
4.2 Correcting Errors

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:

☐	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 28. 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.