

The background of the entire page is a dark blue space filled with a dense field of small, glowing teal dots. Scattered throughout are binary digits '0' and '1' in a light teal color. Several thin, white circles of varying sizes are also present, some of which enclose clusters of the glowing dots. A prominent, bright teal diagonal line runs from the top left towards the center of the page.

Data Submission

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

Table of Contents

I. INTRODUCTION	1
II. PREREQUISITES	1
III. CRDC DATA MODELS	2
IV. DOCUMENTATION	2
V. REQUEST SUBMITTER ROLE	3
VI. STARTING A NEW SUBMISSION	4
VII. CONTINUING AN EXISTING SUBMISSION	6
1. Adding and Managing Collaborators.....	6
2. Obtaining Submission Templates	7
3. Downloading Data Dictionary and Submission Templates	11
3.1 Submission Templates and Properties	13
3.1.1 Special Columns.....	13
3.1.2 Type Column	14
3.1.3 Relationship Columns (Parent Mapping Columns).....	14
4. Uploading Data Files and Metadata Manifests	14
4.1 Uploader CLI Tool.....	14
4.1.1 Introduction.....	14
4.2 Downloading the Uploader CLI Tool	15
4.2.1 Download the Uploader CLI Tool from the CRDC Submission Portal.....	15
4.2.2 Cloning the Uploader CLI Tool from GitHub	16
4.3 Setting Up the Python Environment	16
4.4 Using the Uploader CLI Tool.....	17
4.4.1 Uploader CLI Tool Configuration File.....	17
4.4.2 File Manifest	19
4.5 Starting the Upload Process.....	19
4.6 Using the CRDC Submission Portal's Graphical Interface to Upload Metadata Submission Templates	19
5. Running Validations	22
5.1 Reviewing Validation Results	23
5.1.1 Viewing and Filtering Validation Results	23
5.2 Correcting Errors.....	25
5.3 Remove Specific Files.....	25
6. Submitting Your Final Dataset	26
7. What to Expect After Submission	26

I. INTRODUCTION

This tutorial walks you through the process of submitting data to CRDC through the [CRDC Submission Portal](#). If you have questions that are not answered here, please contact the Data Submissions team member assigned to your submission once you create a submission successfully 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/Rejection notification will be found on the portal under [Submission Request](#) and also in an email sent to the requestor when the request gets approved or rejected. If rejected, consider [other repositories for sharing data at NIH](#).
- 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 an 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 log in on the CRDC Submission Portal account within 60 days, your access to the portal will be deactivated. This applies even while you are working on an active, ongoing submission. 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 either submit data to the CRDC via the Submission Portal or oversee the relevant submissions. See “V. Request Submitter Role” on page 3 for more details.
- If the submission contains controlled access data, the study must be registered in the [database of Genotypes and Phenotypes \(dbGaP\)](#). dbGaP will provide a dbGaP ID/Accession number (phs000####) upon registration. The portal will not allow users to submit controlled access data without a dbGaP ID/Accession number. To initiate the data submission process, the user should email the dbGaP ID associated with their study to the CRDC Help Desk (NCICRDC@mail.nih.gov) if it was not shared through the CRDC Request Access form (see “V. Request Submitter Role” on page 3). Furthermore, for controlled access studies, data will be released on the Data Commons Portal after it is publicly available on the dbGaP website. Therefore, it is recommended to register the study and work on dbGaP submissions concurrently with the submission to CRDC.
- The CRDC Submission Portal uses CRDC standard Common Data Elements (CDEs), and all 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](#). It is recommended that the submitters get familiar with the CRDC standards before starting a submission. [On the caDSR website](#), 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 property is equivalent to column in a relational model. Relationships serve a similar purpose as foreign keys in a relational model. For example, in the GC Data Model, the Participant node is a child of the Study node and a parent of the Diagnosis, Treatment, Sample, and File nodes. See Figure 1.

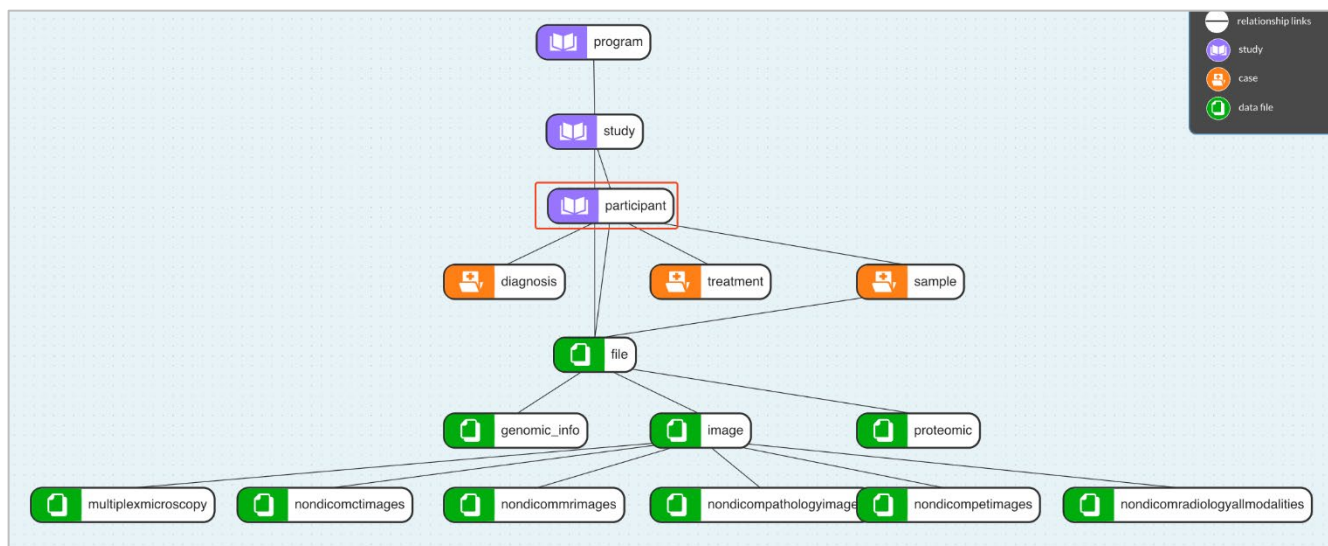


Figure 1. Participant Node Relationships in the GC Data Model

IV. DOCUMENTATION

Submitters can find this instructions document, which discusses submitting data on the CRDC Submission Portal, under the **Documentation** tab. Additionally, the instructions on using APIs for submitting data are also available.

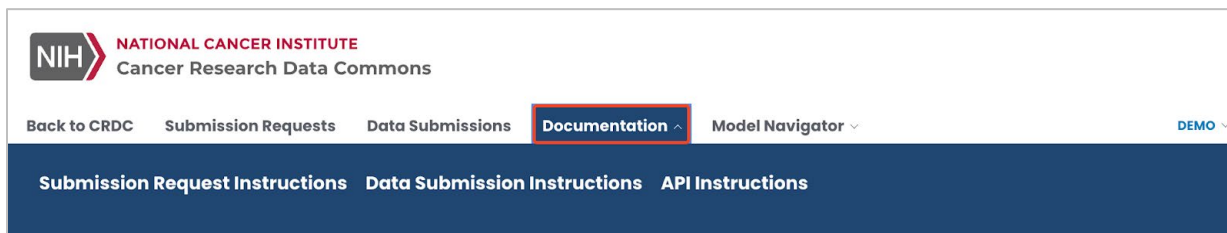
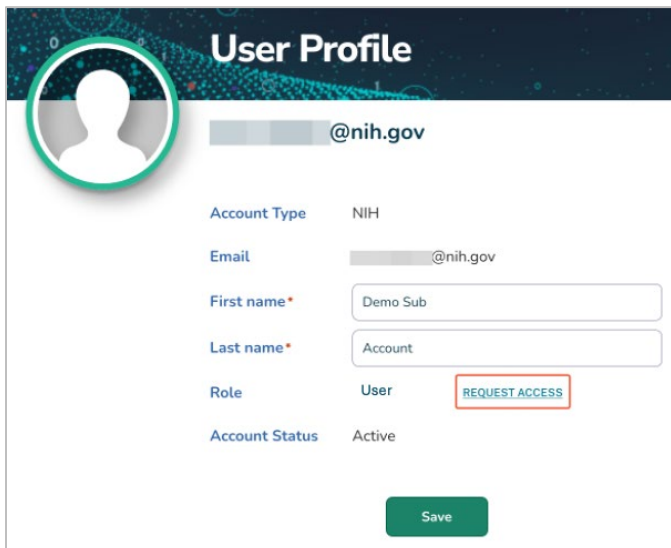


Figure 2. Documentation Menu Showing Available Documents

V. REQUEST SUBMITTER ROLE

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 role, click on **Request Access**. See Figure 3.



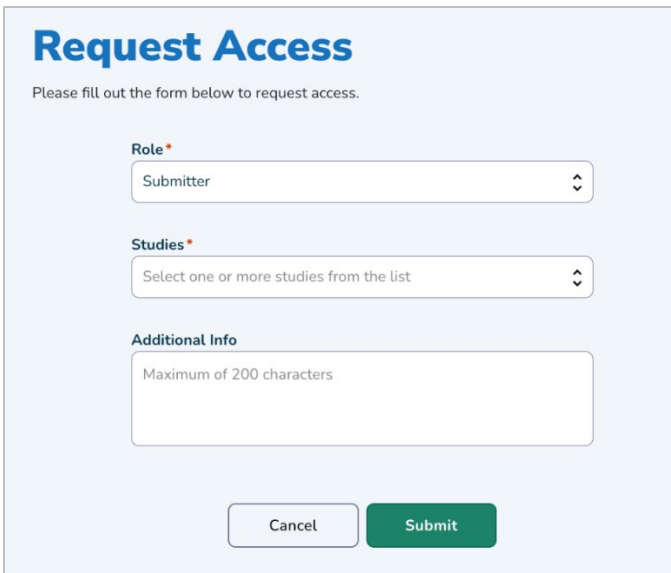
The 'User Profile' form displays the following information:

- Account Type:** NIH
- Email:** [redacted]@nih.gov
- First name:** Demo Sub
- Last name:** Account
- Role:** User (with a red-bordered button labeled 'REQUEST ACCESS' next to it)
- Account Status:** Active

A green 'Save' button is located at the bottom right of the form.

Figure 3. Request Access

3. A **Request Access Form** appears (see Figure 4) where you provide the required information. From the **Role** dropdown menu, select **Submitter** if you want to begin the data submission process or oversee the submissions for specific studies. From the **Studies** dropdown menu, select the relevant Studies from the list. Users can also provide additional details about their role in the **Additional Info** box before submitting the form.



The 'Request Access' form includes the following fields:

- Role:** A dropdown menu with 'Submitter' selected.
- Studies:** A dropdown menu with the text 'Select one or more studies from the list'.
- Additional Info:** A text area with a placeholder 'Maximum of 200 characters'.

At the bottom, there are 'Cancel' and 'Submit' buttons.

Figure 4. Request Access Form

4. The System Administrator grants the Submitter Role, and the user is notified via their login identity.
5. The user granted with the **Submitter** role can now start the data submission process.

Note: If you do not log in with the CRDC Submission Portal account within 60 days, your access to the portal will be deactivated. If that happens, please contact the CRDC Help Desk (NCICRDC@mail.nih.gov) to reactivate your account.

VI. STARTING A NEW SUBMISSION

Once logged in with a Submitter role, navigate to the Data Submissions tab on the CRDC Data Submission portal. The submitter is taken to the 'Data Submissions List' page. 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.

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Back to CRDC Submission Requests **Data Submissions** Documentation Model Navigator DEMO

Data Submission List

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

[Create a Data Submission](#)

Program: All Status: 6 statuses selected 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	Program	Study	dbGaP ID	Status	Primary Contact	Record Count	Created Date	Last Updated
Test3 GC	Demo Sb	GC	New/Update	v6.0.4	NA	GCTest	phs0000GC	In Progress		2,805	4/24/2025	4/25/2025
Test2 GC	Demo Sb	GC	New/Update	v6.0.4	NA	GCTest	phs0000GC	In Progress		2,805	4/24/2025	4/24/2025

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 appears. Fill out all the 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

Data Commons *
GC

Study *
CONTROLLED1

dbGaP ID *
<Not Provided>

Submission Name *
25 characters allowed

Create

Please contact NCICRDC@mail.nih.gov to submit your dbGaP ID once you have registered your study on dbGap.

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 the **Delete** option to remove files from a previous submission already released publicly by the Data Commons. Selecting the **Delete** option keeps only one **Data Type** option enabled, which would be **Metadata Only**. Submit the metadata associated with the data that needs to be deleted. The deletion request goes through a validation process by the CRDC Team and the CRDC Data Commons. Once approved, the deletion can be processed.
2. For Data Type, indicate whether you are submitting both metadata and data files or only metadata files by selecting one of the options, **Metadata and Data Files** or **Metadata Only**, respectively.
3. Select the **Data Commons** that you were approved to submit to by the Submission Review Committee (SRC) for your data, if it is not preselected already. If you do not see the CRDC Data Commons listed, email the CRDC Help Desk (NCICRDC@mail.nih.gov).
4. The **Study** dropdown menu displays the Study Title (or the Study Abbreviation) you previously shared through the Submission Request form. If you notice an error in this list, email the CRDC Help Desk (NCICRDC@mail.nih.gov).
5. If your study includes controlled access data, the **dbGaP ID** /accession number will be pre-populated as provided in the Submission Request Form or emailed to CRDC Help Desk (in case of the conditionally approved Submission) as shown in Figure 6. The system will not allow the submitter to initiate a data submission without providing the dbGaP ID by disabling the **Create** button.

- Submitters can give the submission a **Name** in the provided free-text field to label their submissions. This name appears in the Submissions List table on the Data Submissions List page, once the submission is created.
- Create the new submission by clicking the **Create** button, which will be enabled if the dbGaP ID is provided.

Once a submission is created, CRDC assigns a Primary Contact to your submission. You can find the email address of the CRDC Primary Contact assigned to your submission on the dashboard of your data submissions 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 Submissions** tab. A table, listing all the existing submissions, appears on that page. Under the **Submission Name** column, select the submission you want to continue with. To see the full version of the submission name, hover over the name (see Figure 7). The submitters 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. The filters at the top of the table provide useful ways to refine searches, especially when the list of submissions is extensive.

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Back to CRDC Submission Requests **Data Submissions** Documentation Model Navigator DEMO

Data Submission List

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

[Create a Data Submission](#)

Program: All Status: 6 statuses selected 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	Program	Study	dbGaP ID	Status	Primary Contact	Record Count	Created Date	Last Updated
Test3_GC	Demo Sb	GC	New/Update	v6.0.4	NA	GCTest	phs0000GC	In Progress		2,805	4/24/2025	4/29/2025
Test2_GC	Demo Sb	GC	New/Update	v6.0.4	NA	GCTest	phs0000GC	In Progress		2,805	4/24/2025	4/24/2025
Test_GC_2	Demo Sb	GC	New/Update	v6.0.4	NA	GCTest	phs0000GC	In Progress		3	4/24/2025	4/24/2025

Figure 7. Active Data Submissions List

1. Adding and Managing Collaborators

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

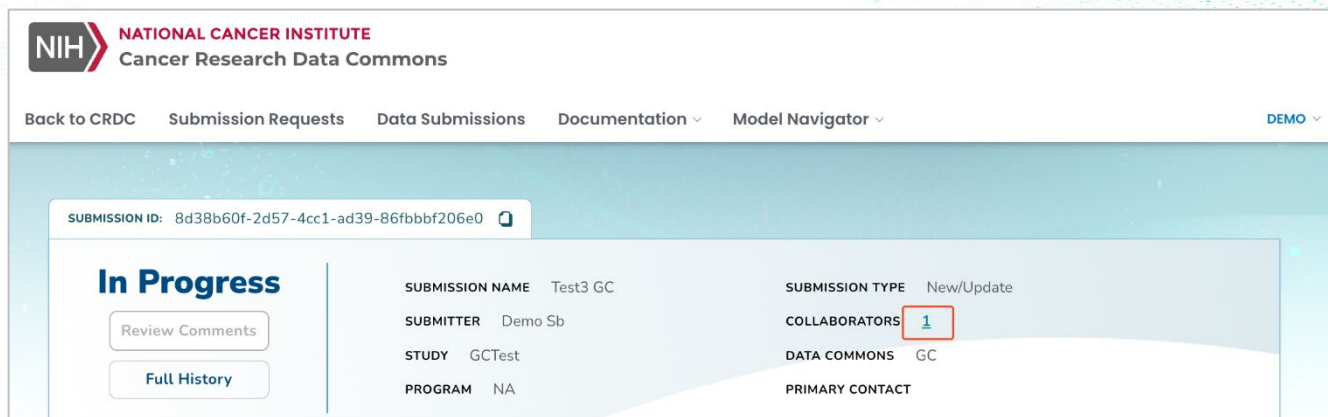


Figure 8. Count of Collaborators on the Data Submission Dashboard

To add and manage collaborators, click the hyperlink next to **Collaborators** on the Submission dashboard, which opens 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. Please ensure that the collaborator has an authorized account on the CRDC Submission Portal with the Submitter role, is affiliated with the same study, and has permissions from the data owner or Study PI/s, to submit data.

The submitter can **remove** a collaborator by clicking the remove icon [X], 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.

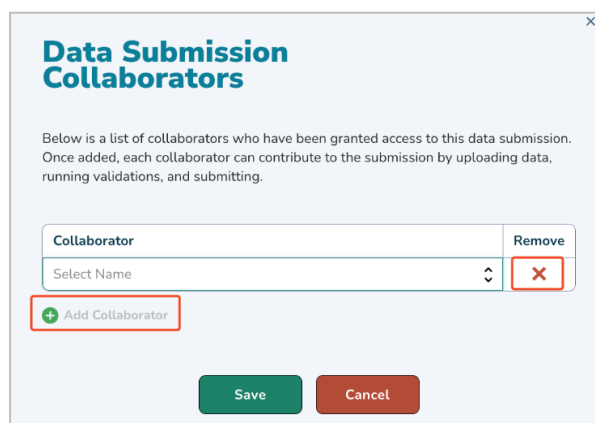


Figure 9. Add and Remove Collaborators for a Given Submission

2. Obtaining Submission Templates

Submitting data to CRDC requires you to submit your metadata into the submission templates. The metadata is then used to validate the information about the actual raw data. For instance, the file name and the file size of each uploaded raw data file will be compared with the file name and file size specified in the metadata manifest.

To get to the submission templates, click **Model Navigator** in the menu bar (see Figure 10), which lists the Data Models of the various Data Commons integrated under the CRDC Data Submission Portal. Select the Data Model respective to the Data Commons (DC) that the Submission Review Committee (SRC) has

approved you to submit data to. At this time, the submission templates for various Data Commons, including the General Commons (GC) Model, Clinical and Translational Data Commons (CTDC) Model, and Integrated Canine Data Commons (ICDC) Model are provided. Models of other Data Commons will be added when they are integrated with CRDC Submission Portal.

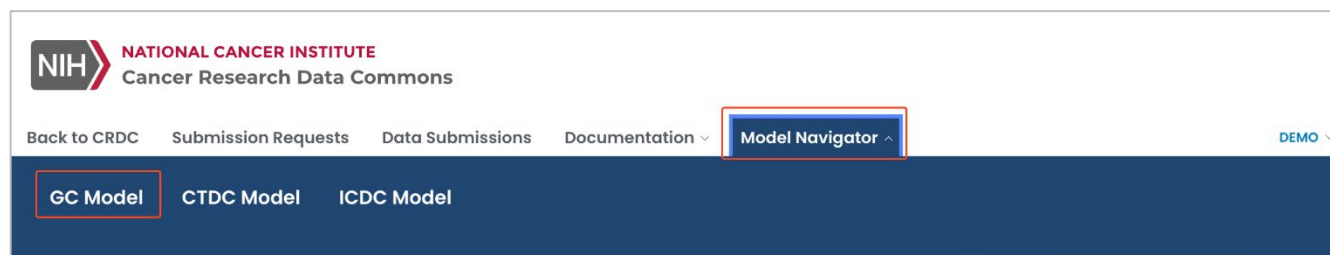


Figure 10. Use the Menu Bar to Navigate to the Data Model Viewer

Once you select the Data Model, you are taken to the Data Model Viewer page, as seen in Figure 11. On this page, you can view the data model in detail and also download the submission templates from the dropdown menu of Available Downloads. **Note:** Figure 11 shows the GC Data Model as an example.

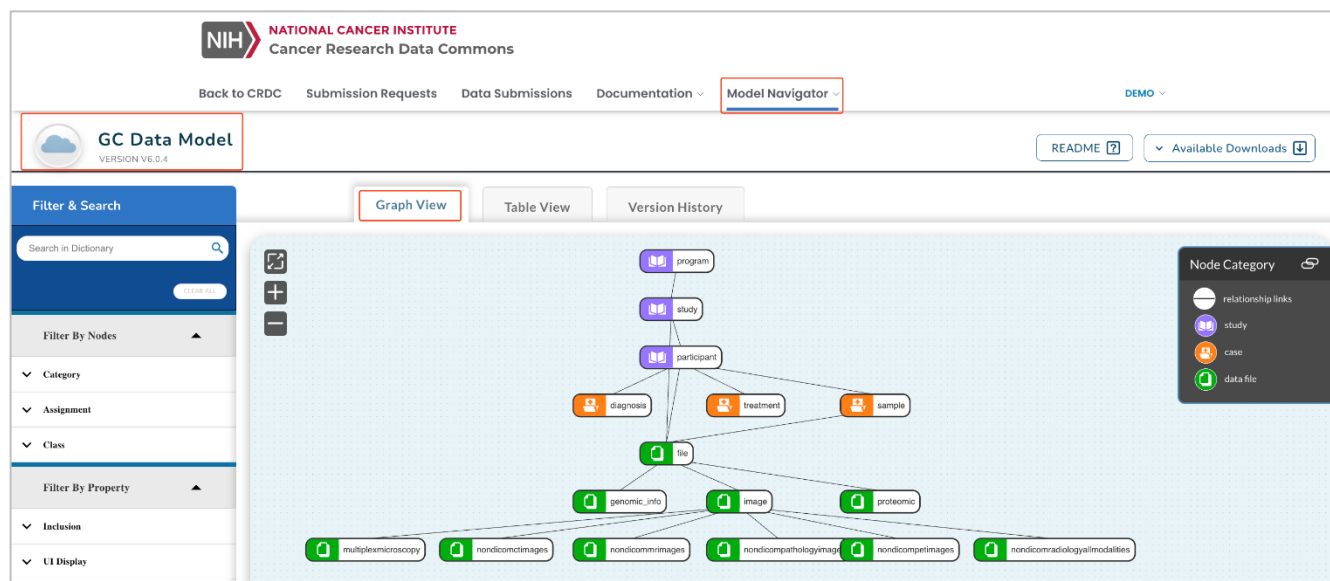


Figure 11. Data Model Viewer Graph View

Use the Data Model Viewer to explore the data elements that the Data Commons require or can accept. On the **Graph View** tab, the data model is represented in Nodes and Relationships. Clicking a node in the graph shows its summary. At the bottom of the node summary, click on the **View Properties** to open a **Table View** of the selected node. For instance, as shown in Figure 12 clicking on the **Sample** node opens its summary, with View Properties option at the bottom.

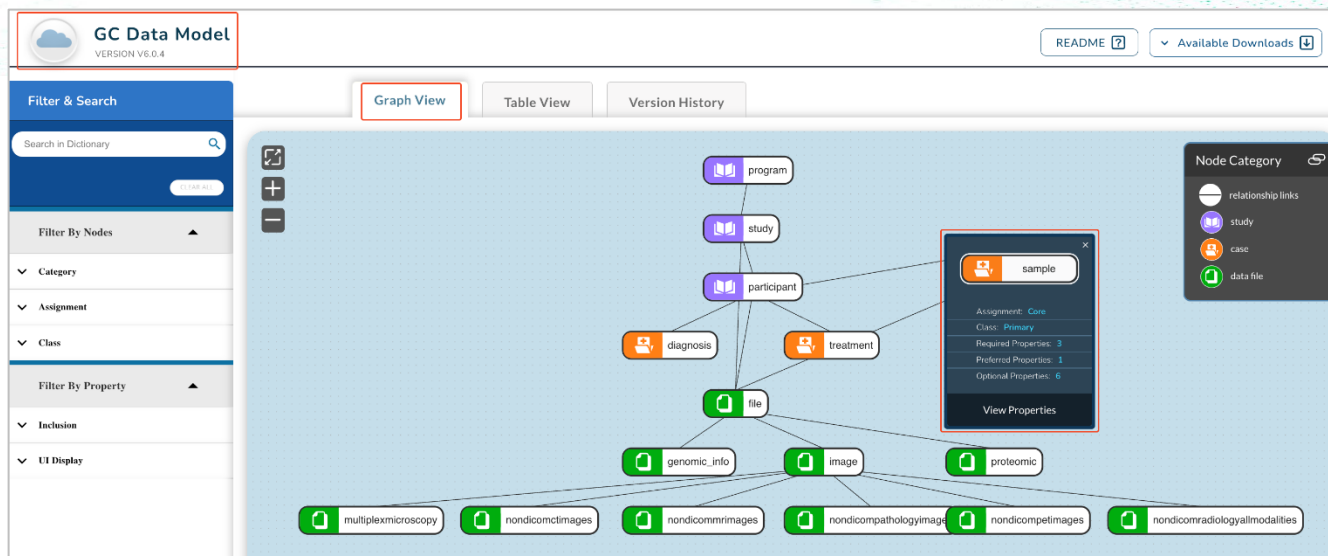


Figure 12. Click a Node to View a Summary and Open the Table View

The **Table View** lists all the data elements/properties of the data model and includes the description of each of these Properties (such as strings, integers, etc.) The table View also shows which of these properties are required. See Figure 13. Each of these properties are mapped to the Common Data Elements (CDEs) of the caDSR standards where applicable. Please note that CRDC data validations will only accept **Permissible Values** for elements mapped to the **Common Data Elements (CDEs)**. Details about the CDEs can be accessed by clicking on the Public ID for that specific Property. New CDEs and/or new Permissible Values can be requested by emailing the CRDC Help Desk (NCICRDC@mail.nih.gov), if you do not find the right match in the provided values. These requests will need to go through an approval process before you can use the new CDEs or the PVs for submission.

Additionally, in the **Table View** the **Submission Template** and the associated **Data Dictionary** for the specific node can be downloaded as shown in Figure 13 and described in the next section.

Case

Sample

Specimen tissue type collected from the participant.

10 Properties

Assignment: Core Class: Primary

Template Data Dictionary

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: • Urine • Tissue • Stool • Sputum • Mouth Rinse • Fluids • Cells • Bone Marrow • Blood • Ascites ...show more	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 • Body Fluid or Substance • Sputum • Urine • Tissue ...show more	CDE Full Name Specimen Material Category Version 1.00 Public ID 12445832 Origin caDSR	Optional	The kind of material that forms the sample.	**

Figure 13. Table View of an Excerpt of Sample Node

Additionally, any updates to the data model, such as property modifications, additions, or removals of permissible values, are reflected in the **Version History** tab, ensuring users have access to the changes. See Figure 14. Currently Version History is applicable and available for GC data model only.

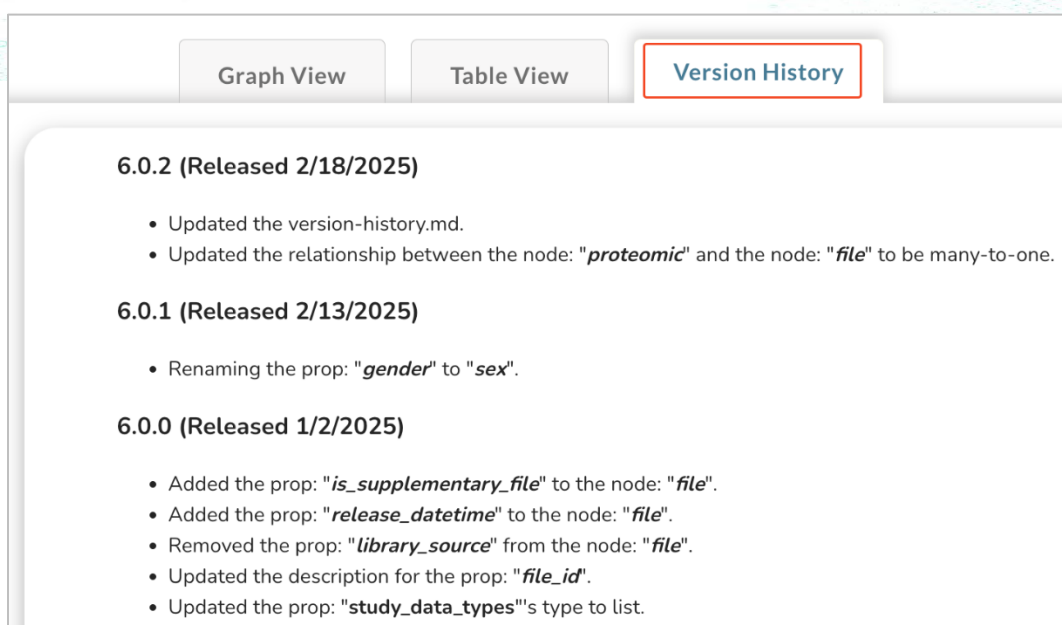


Figure 14. Excerpt of Version History of the GC Data Model

3. Downloading Data Dictionary and Submission Templates

Each Data Commons has its own unique Data Model with corresponding data dictionary, set of nodes, properties, and submission templates. To download the Data Dictionary and the submission templates, select **Available Downloads**, as shown in Figure 15. Then click on one of the options from the dropdown menu and the file in the selected format will be downloaded. CRDC Vocabularies for the selected Data Commons Model and Examples Templates can also be downloaded.

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

Submitters can use the **Submission Templates** to format and upload metadata to 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** provides detailed information about the metadata structure, content, and the required, preferred, and optional data elements for all nodes within the selected Data Model. Submitters can choose to download the Data Dictionary for either all Properties or only the Required Properties. The **All Vocabularies** document contains the permissible values for the data elements. 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.

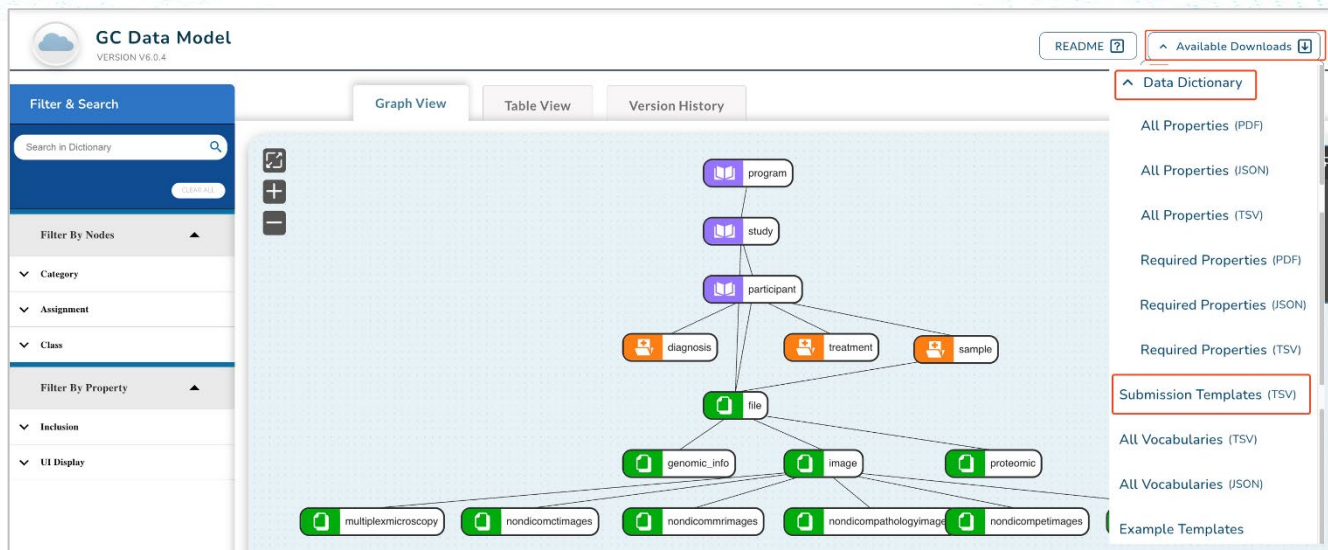


Figure 15. Using the Available Downloads Menu

The downloaded files are provided as a ZIP archive. The tab-separated text files can be viewed in any text editor or spreadsheet application like Microsoft Excel or OpenOffice Calc. Multiple metadata template files in TSV format are included within the ZIP archive of Submission Templates. The Submission Templates for the GC model is shown in Figure 16.

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

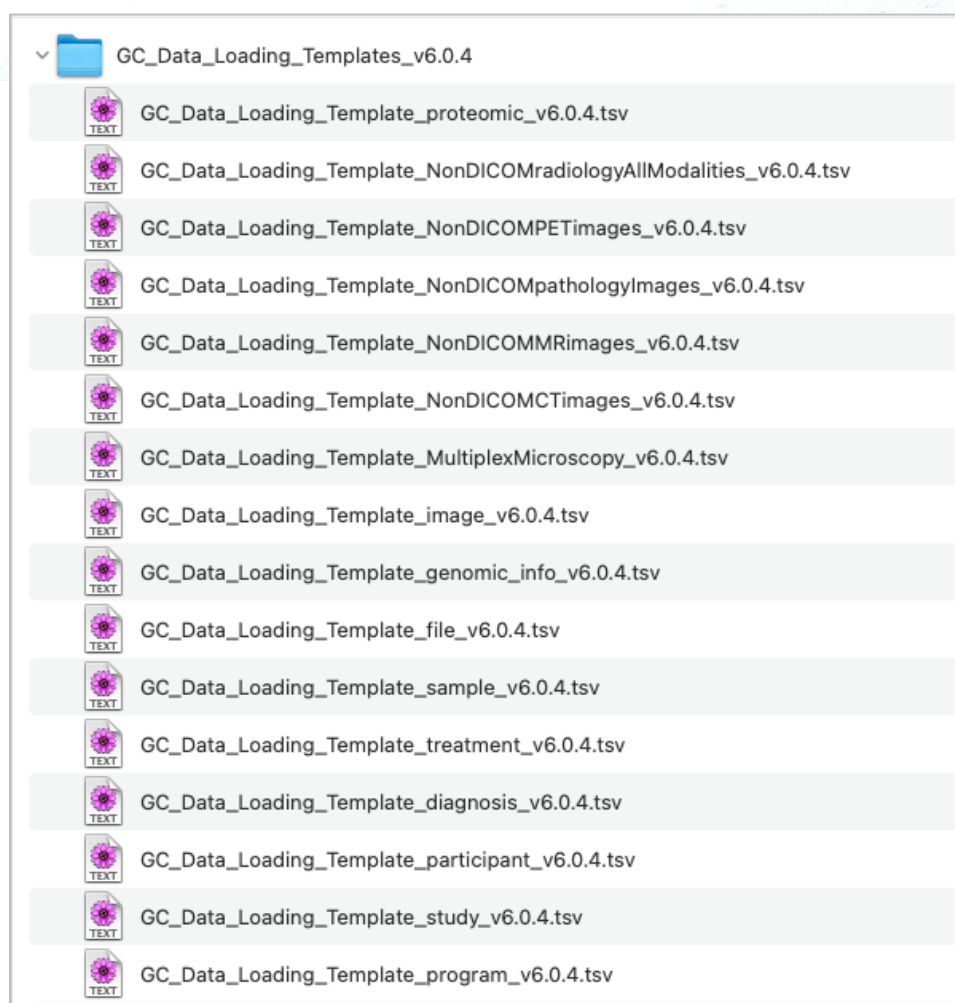


Figure 16. Submission Templates Downloaded from the CRDC Submission Portal Model Viewer

3.1 Submission Templates and Properties

Each of the submission templates covers information relevant to a specific node in the model; for example, the template 'GC_Data_Loading_Template_image_v6.0.4.tsv,' 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 'GC_Data_Loading_Template_image_v6.0.4.tsv'.

For every template that will be submitted, review the Data Dictionary (accessible through the **Available Downloads** menu) to understand what properties are required, as each individual template has few required properties, as well as preferred and optional properties 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.

3.1.1 Special Columns

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

3.1.2 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 ‘GC_Data_Loading_Template_study_v6.0.2.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.

3.1.3 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.

4. 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 to CRDC .
- **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.

4.1 Uploader CLI Tool

4.1.1 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. Binary versions of the CLI Tool are also available, which don’t require any installation or Python.

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.

4.2 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. However, downloading from the CRDC Submission Portal is recommended as it provides the latest version.

4.2.1 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, and it opens the menu. Select **Uploader CLI Tool** from the menu and a pop-up window will appear with the latest version of the CLI tool. See Figure 17.

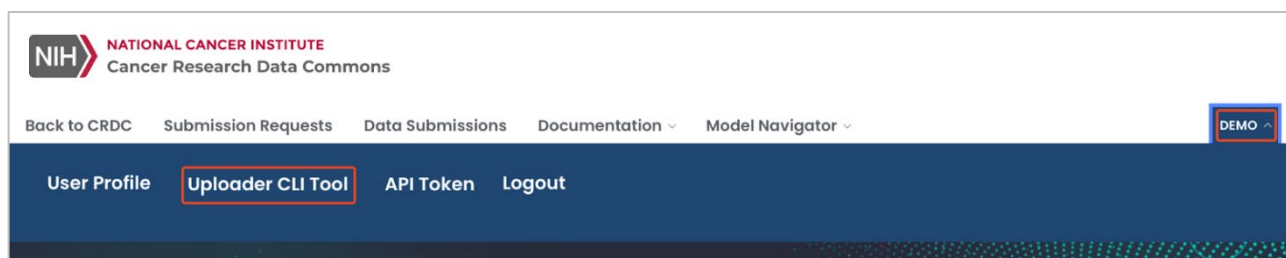


Figure 17. Menu with the Uploader CLI Tool Download Option

Click the **Download** icon next to the available Package Type options to download the CLI tool. The download comes with accompanying instructions (see Figure 18). A ZIP archive will be saved to your local machine.

Uploader CLI Tool

UPLOADER CLI VERSION: **v3.2**

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

Close

Figure 18. Download the Uploader CLI Tool

4.2.2 Cloning the Uploader CLI Tool from GitHub

The latest version of the Uploader CLI Tool can also be cloned from the Data Hub [GitHub repository](https://github.com/CBIIT/crdc-datahub-cli-uploader). 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>

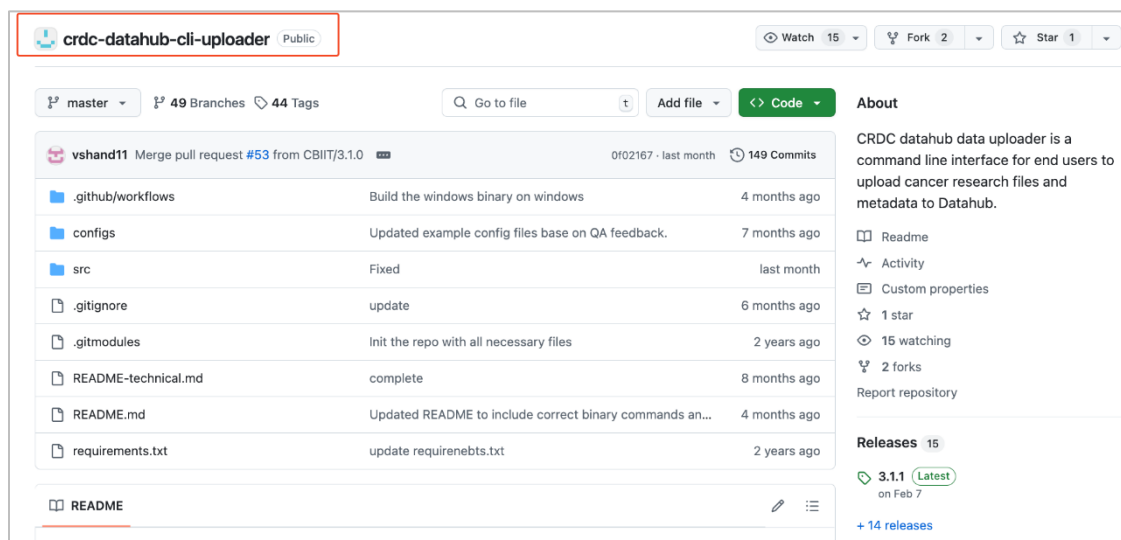


Figure 19. Uploader CLI Tool as it Appears in GitHub

4.3 Setting Up the Python Environment

Binary versions of CLI Tool are self-contained and don't require Python or installation of any dependencies. The Source version of 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
- rich

4.4 Using the Uploader CLI Tool

4.4.1 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 by clicking the **Download Configuration File** button, see Figure 20. This will open the **Download Configuration File** pop-up window, see Figure 21.

Enter the path to the local or S3 folder containing your Data Files and Metadata Manifest File (explained in Section 3.4.2) in the designated text boxes.

Note: If you enter an S3 URL in the **full path to Data Files folder**, the CLI tool will initiate an S3-to-S3 transfer.

Clicking the **Download** button will download the configuration file in YML format to your computer with prepopulated fields. Please note that uploading data files using the CLI does not support a nested folder structure. To proceed, either place all data files in a single folder or adjust the path in the configuration file to upload data from subfolders.

Prior to beginning uploading process, read detailed instructions available in the [Data Submission Instructions](#).

1

UPLOAD METADATA

CDS DATA MODEL: v6.0.2

Metadata Files

Choose Files

No files selected

Upload

2

UPLOAD DATA FILES

UPLOADER CLI VERSION: v3.2

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 20. 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

Cancel

Download

Figure 21. Dependencies to Download Configuration File

Additionally, if you choose 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 the configuration file needed by 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 the Configuration File needed by the Uploader CLI Tool 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 Configuration 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 follow.

- **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 Submissions 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.
- **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.

While users are expected to provide paths to their data folder and manifest file, they may choose to customize the values of the three parameters—**retries**, **overwrite**, and **dryrun**—to suit their needs.

Additionally, the **Data Submission Instructions** document, the version of the **Data Model** your submission utilizes, and the **CLI tool itself** can all be accessed on the same page. See Figure 20.

4.4.2 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 which are generated by the CLI tool. Submitters can use the `file.tsv` template downloaded from the Data Model viewer page, 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 after the CLI tool has uploaded it on the Portal, ensure you edit the final version (e.g., `file-final.tsv` saved locally on your computer) before re-uploading it through the User Interface. If you need to re-upload your data files for any reason, you can reuse the final manifest generated by the CLI tool. This will retain the existing file IDs/Keys instead of generating new ones.

4.5 Starting the Upload Process

Once the configuration file has been downloaded or 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 that 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
```

4.6 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 for a submission from the **Data View** table. Refer to Figure 22.

In case you intend to upload the metadata in phases, you should keep your new metadata, 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.

Batch ID	Batch Type	File Count	Status	Uploaded Date ↓	Uploaded By	Upload Errors
5	Metadata	16	Uploaded	2/25/2025	Demo Account	
4	Metadata	2	Failed	2/25/2025	Demo Account	2 Errors
3	Metadata	2	Failed	2/25/2025	Demo Account	7 Errors

Figure 22. 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 number under the **File Count** column displays a list of all files uploaded in that batch. If there are errors in the files being uploaded, 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 reupload the 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 reuploaded. An example of batch upload error message is presented in Figure 23.

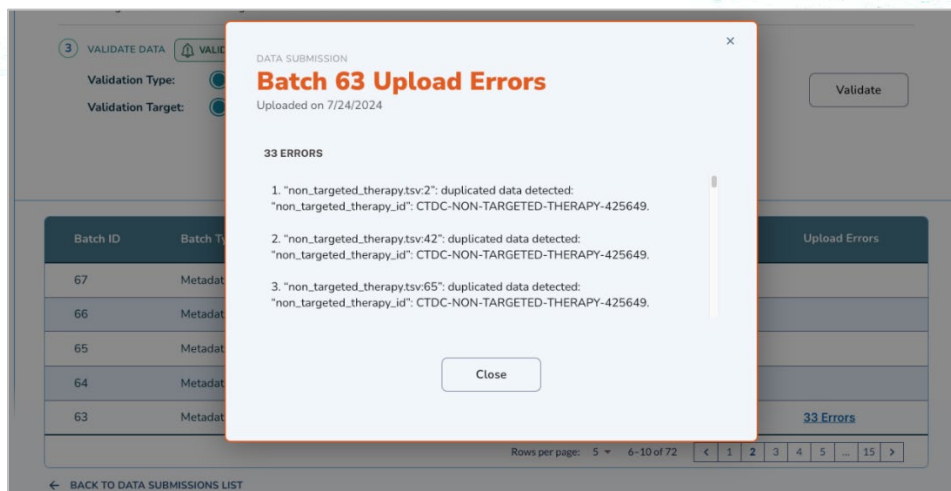


Figure 23. Batch Upload Errors

You can remove specific metadata previously uploaded in the system in the Data View table, as depicted in Figure 24. For instance, to remove metadata for a participant, select the file and click the **Delete** icon to remove it from your submission. Note, this action will also delete the associated metadata from the child nodes. To visualize the associated metadata from child the nodes, click on the file-id and a pop-up window show up. See Figure 25. 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.

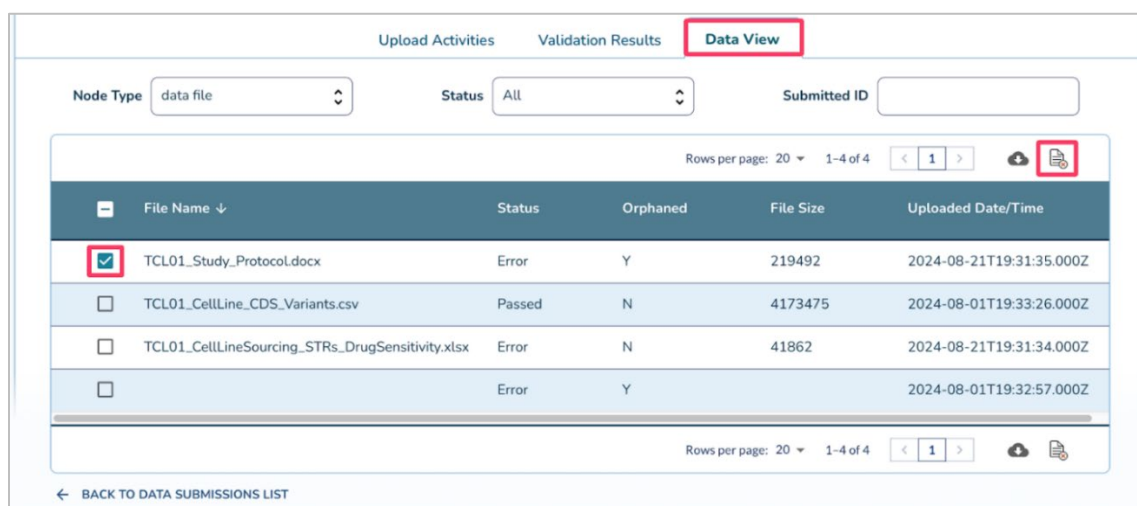


Figure 24. Delete and Download Metadata File(s)

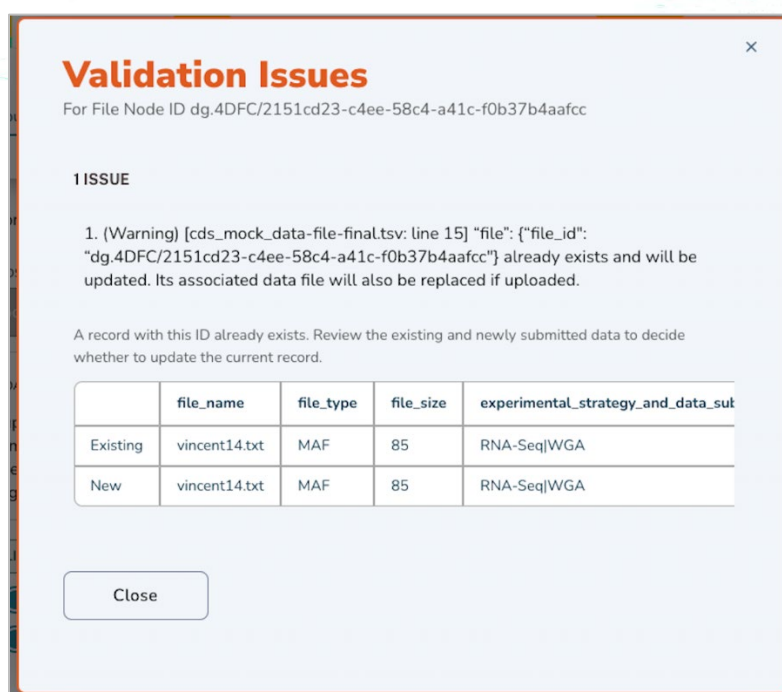


Figure 25. Visualize Associated Metadata from Child Nodes

5. 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 26.

3
VALIDATE DATA
VALIDATION COMPLETED

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

Validation Target:
☒ New Uploaded Data
☐ All Uploaded Data

Validate

Figure 26. 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.

5.1 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 27.

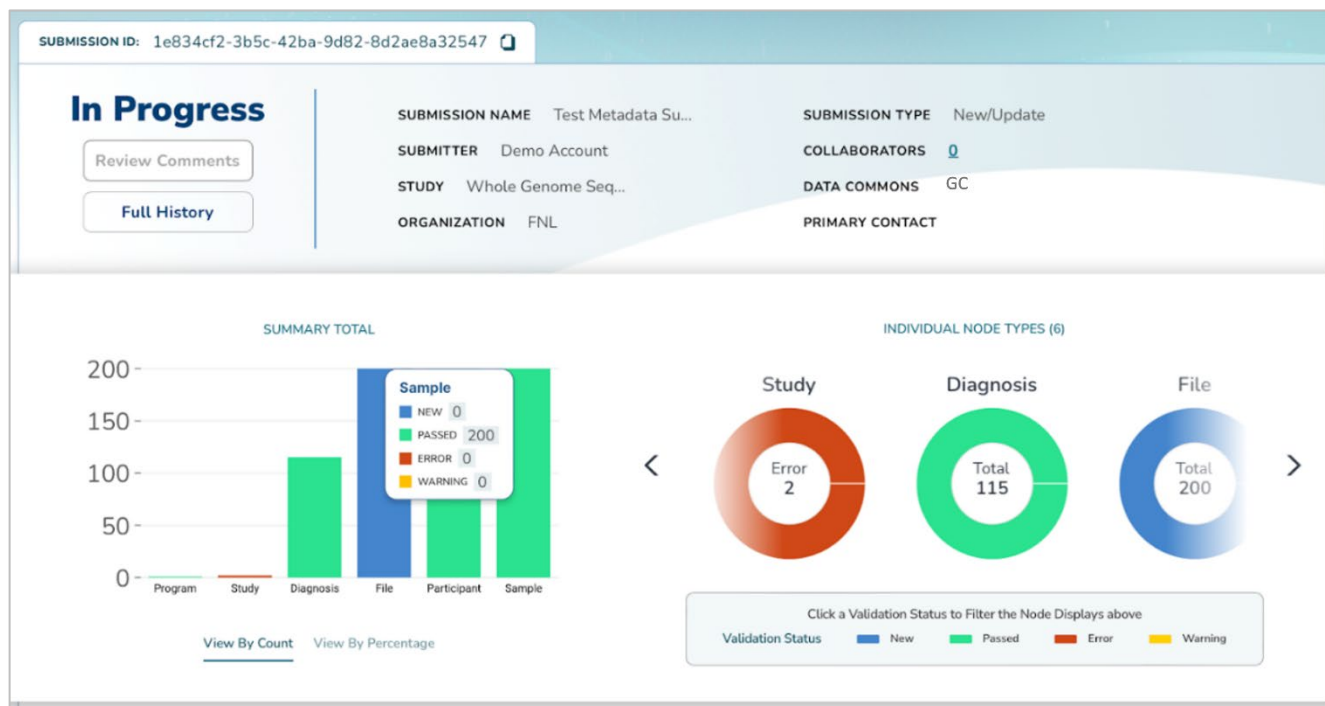


Figure 27. Validation Summary

The left graph in Figure 27 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 27, 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.

5.1.1 Viewing and Filtering Validation Results

By default, Submitters can view validation results in an **Aggregated** format, grouped by **Issue Type** as shown in Figure 28. To explore errors in more detail, follow these steps:

1. Expand detailed errors. See Figure 29.
 - a. Click the **Expand link** under the Expand column to view the detailed validation errors for the specific Issue Type.
 - b. Additional filters (Issue Type, Batch ID, Node Type, and Severity) can be helpful to narrow down the results. Selecting a **specific Node Type** refines the search to relevant validation errors. 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.

- c. View details for a specific Issue under the **Issue column** by clicking on the **See details** to access more specific information regarding each validation error.
- d. Users can download all the validation errors in a table by clicking on the cloud icon.

Issue Type	Severity	Count	Expand
Value not permitted	Error	1,659	Expand
Invalid Property	Warning	300	Expand
Invalid integer value	Error	300	Expand
Updating existing data	Warning	16	Expand

Figure 28. Aggregated Validation Results Table

Batch ID	Node Type	Submitted Identifier	Severity	Validated Date	Issues
4	Demographic	CTDC-DEMOGRAPHI...	Error	2/26/2025	Value not permi... See details.
4	Demographic	CTDC-DEMOGRAPHI...	Error	2/26/2025	Value not permi... See details.
4	Demographic	CTDC-DEMOGRAPHI...	Error	2/26/2025	Value not permi... See details.
4	Demographic	CTDC-DEMOGRAPHI...	Error	2/26/2025	Value not permi... See details.

Figure 29. Expanded 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 30 presents an example of Validation Issue details.

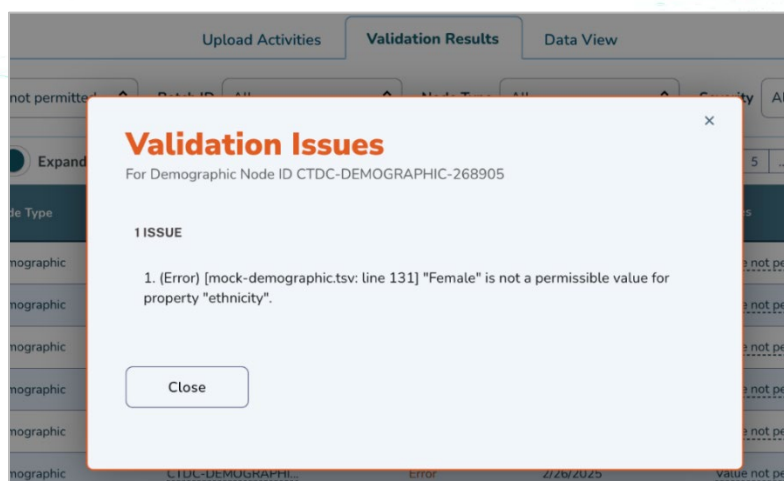


Figure 30. Validation Error Details

5.2 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.

5.3 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.

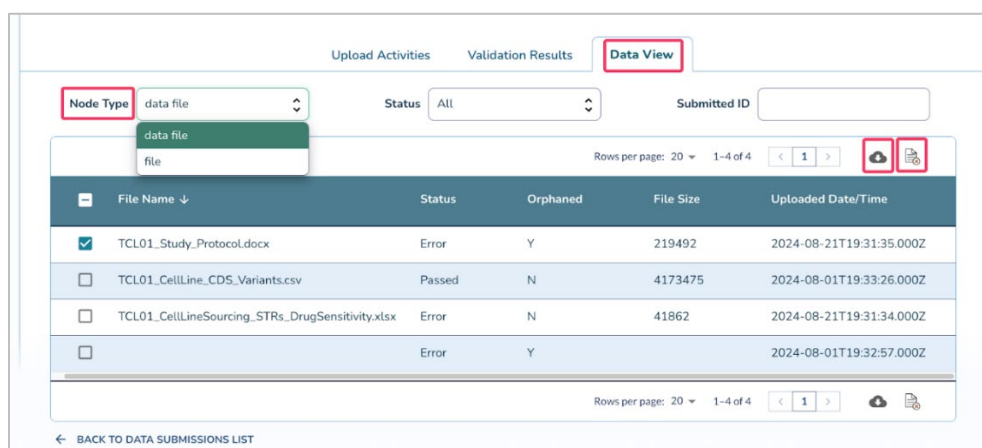


Figure 31. Removing Specific Files

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

- **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.

6. 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 CRDC 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 the control to you.

7. 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, GC, 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 for secondary sharing and will be made available and accessible on their portal.

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