NATIONAL CANCER INSTITUTE
GDC Documentation

GDC Data Submission Portal User's Guide

NCI Genomic Data Commons (GDC)

NCI GDC

Table of contents

1. Data Submission Portal

1.1 Before Submitting Data to the GDC Portal

1.1.1 Overview

The National Cancer Institute (NCI) Genomic Data Commons (GDC) Data Submission Portal User's Guide is the companion documentation for the GDC Data Submission Portal and provides detailed information and instructions for its use.

1.1.2 Steps to Submit Data to the GDC

The following tasks are required to submit data to the GDC Data Submission Portal.

1. Complete the GDC Data Submission Request Form. After submission, the reqest will be reviewed by the GDC Data Submission Review Committee. During this time, create an eRA Commons account if you do not already have one.
2. If the study is approved, contact a Genomic Program Administrator (GPA) to register the approved study in dbGaP. This includes registering the project as a GDC Trusted Partner study, registering cases, and adding authorized data submitters. For more information, see the Data Submission Process.
3. Contact GDC User Services to create a submission project. The User Services team will require a project ID, which is a two-part identifier, where the first portion is the **Program** followed by a hyphen (**-**) and the second portion is the **Project**. This must be alphanumeric and all caps only. An example would be `TCGA-BRCA`. You must also create a project name, which can be longer and has fewer requirements on length or character usage. An example would be `Breast Invasive Carcinoma`.

1.1.3 Key Features

The GDC Data Submission Portal is a platform that allows researchers to submit and release data to the GDC. The key features of the GDC Data Submission Portal are:

- **Upload and Validate Data**: Project data can be uploaded to the GDC project workspace. The GDC will validate the data against the GDC Data Dictionary.
- **Browse Data**: Data that has been uploaded to the project workspace can be browsed to ensure that the project is ready for processing.
- **Download Data**: Data that has been uploaded into the project workspace can be downloaded for review or update by using the API or the Data Transfer Tool.
- **Review and Submit Data**: Prior to submission, data can be reviewed to check for accuracy and completeness. Once the review is complete, the data can be submitted to the GDC for processing through Data Harmonization.
- **Release Data**: After harmonization, data can be released to the research community for access through the GDC Data Portal and other GDC Data Access Tools.
- **Status and Alerts**: Visual cues are implemented in the GDC Data Submission Portal Dashboard to easily identify incomplete submissions via panel displays summarizing submitted data and associated data elements.
- **Transactions**: A list of all actions performed in a project is provided with detailed information for each action.

1.1.4 Sections to the Data Submission Portal Guide

- **Data Submission Overview**: Graphical explanations used to display the life cycle of projects and their data.
- **Data Submission Process**: An overview of the data submission process using the GDC Data Submission Portal.
- **Data Submission Walkthrough**: Step-by-step instructions on GDC data submission and their relationship to the GDC Data Model.

1.1.5 HIPAA Compliance

The GDC will not accept any data for patients age 90 and over including any follow-up events in which the event occurs after a patient turns 90 to ensure that HIPAA compliance is maintained. To comply with these requirements data submitters may omit any data (entire cases or specific nodes) that would violate this rule or obfuscate associated dates. Please see the Date Obfuscation section for more information.

1.1.6 Release Notes

The Release Notes section of this User's Guide contains details about new features, bug fixes, and known issues.

1.2 Data Submission Overview

1.2.1 Overview

This section will walk users through two parts of the submission process. The first portion will be the steps taken by the users to go through the submission process from start to finish. The second portion will describe the lifecycle of a project and a file throughout the data submission process.

1.2.2 GDC Data Submission Workflow

The diagram below illustrates the process from uploading through releasing data in the GDC Data Submission Portal. To review the steps needed before beginning submission see Before Submitting Data to the GDC Portal

GDC Data Submission Portal Workflow Upload

Review GDC Dictionary and GDC Data Model - Submitter Activity

It is suggested that all submitters review the GDC Dictionary and GDC Data Model. It is beneficial for submitters to know which nodes will need metadata submission, how these nodes relate to each other, and what information is required for each node in the model.

Download Templates - Submitter Activity

After determining the required nodes for the submission, go to each node page in the GDC Dictionary. There will be a "Download Template" drop down list. Select the file format, either TSV or JSON, and download the template for the node. If numerous entries are being submitted all at one time, it is suggested that the user uses a TSV template. At this point, it is suggested to go through the template and remove fields that will not be populated by the metadata submission, but make sure to complete all fields that are required for the node. For more information about the Data Dictionary, please visit here.

`See GDC Data Dictionary here.`

Upload Case Information Including dbGaP Submitted Subject IDs - Submitter Activity

After registering the study in dbGaP, the first node to be created in the data model is the `case` node. The `case` node is important as it will contain a unique `submitter_id` that is registered in dbGaP under a particular project. This will connect the two databases, dbGaP and GDC, and allows for access to be granted to a controlled data set based on the study and its cases.

To submit the `case` nodes, a user must be able to login and access the GDC Submission Portal for their respective project. Metadata for all nodes are uploaded via the API or through the Submission Portal.

`See case example here.`

`See metadata upload example here.`

Upload Clinical and Biospecimen Data - Submitter Activity

With the creation of `case` nodes, other nodes in the data model can be uploaded. This includes the Clinical and Biospecimen nodes, with examples for each that can be found in the Data Upload Walkthrough.

`See clinical example here.`

`See biospecimen example here.`

`See metadata upload example here.`

Register Data Files - Submitter Activity

Registering data files is necessary before they can be uploaded. This allows the GDC to later validate the uploads against the user-supplied md5sum and file size. The submission of these files can range from clinical and biospecimen supplements to `submitted_aligned_reads` and `submitted_unaligned_reads` .

`See experiment data example here.`

Upload Data Using Data Transfer Tool - Submitter Activity

Before uploading the submittable data files to the GDC, a user will need to determine if the correct nodes have been created and the information within them are correct. This is accomplished using the Browse page in the Data Submission Portal. Here you can find the metadata and file_state, which must have progressed to `registered` for an associated file to be uploaded. You can find more about the file life cycle here.

Once the submitter has verified that the submittable data files have been registered, the user can obtain the submission manifest file that is found on the Project Overview page. From this point the submission process is described in the "Uploading the Submittable Data File to the GDC" section.

For strategies on data upload, further documentation for the GDC Data Submission process is detailed on the Data Submission Processes and Tools section of the GDC Website.

`See submittable data file upload example here.`

Verify Accuracy and Completeness of Project Data (Project QC) - Submitter Activity

The submitter is responsible for reviewing the data uploaded to the project workspace and ensuring there are no critical QC errors, see Data Submission Walkthrough, and ensuring that it is ready for processing by the GDC Harmonization Process. A user should be able to go through the Pre-Harmonization Checklist, and verify that their submission meets these criteria.

`See pre-harmonization checklist here.`

Request Data Harmonization - Submitter Activity

When the project is complete and ready for processing, the submitter will request harmonization. If the project is not ready for processing, the project can be re-opened and the submitter will be able to upload more data to the project workspace.

`See harmonization request example here.`

NOTE: The GDC requests that users submit their data to the GDC within six months from the first upload of data to the project workspace.

GDC Review/QC Submitted Data - GDC Activity

The Bioinformatics Team at the GDC runs the Quality Control pipeline on the submitted data. This pipeline mirrors the Pre-Harmonization Checklist and will determine if the submission is complete and is ready for the Harmonization pipeline. If the submission does contain problems, the GDC will contact the user to "Re-Open" the project and fix the errors in their submission.

Once the review is complete, all validated nodes will be changed to state "submitted". At this point users can submit more files to a project, but they will be considered as a different batch for harmonization.

GDC Harmonize Data - GDC Activity

After the submission passes the GDC Quality Control pipeline, it will be queued for the GDC Harmonization pipeline.

Submitter Review/QC of Harmonized Data - Submitter Activity

After the data is processed in the Harmonization pipeline, the GDC asks submitters to verify the quality of their harmonized data. It is the user's responsibility to notify the GDC of any errors in their harmonized data sets. The GDC will then work with the user to correct the issue and rerun the Harmonization pipeline if needed.

Release Data Within Six Months - Submitter Activity

Project release occurs after the data has been harmonized, and allows users to access this data with the GDC Data Portal and other GDC Data Access Tools. The GDC will release data according to GDC Data Sharing Policies. Data must be released within six months after GDC data processing has been completed, or the submitter may request earlier release.

`See release example here.`

Note: Released cases and/or files can be redacted from the GDC. For more information, visit the GDC Policies page (under GDC Data Sharing Policies).

GDC Releases Data - GDC Activity

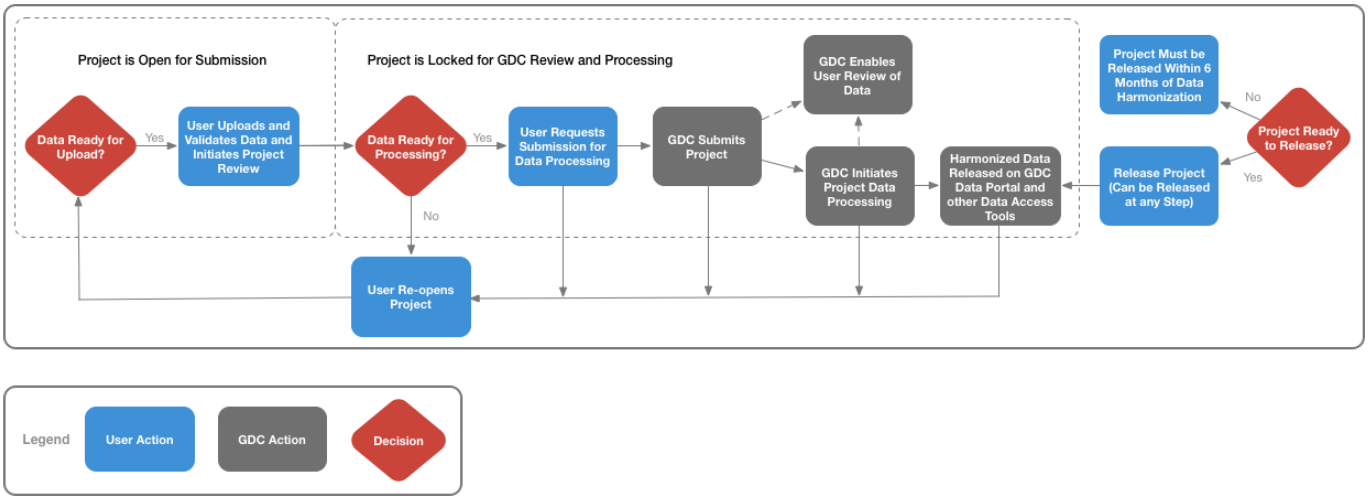
GDC data releases are not continuous, but instead are released in discrete data updates. Once harmonized data is approved and release request is approved, data will be available in an upcoming GDC Data Release.

1.2.3 Project and File Lifecycles

Project Lifecycle

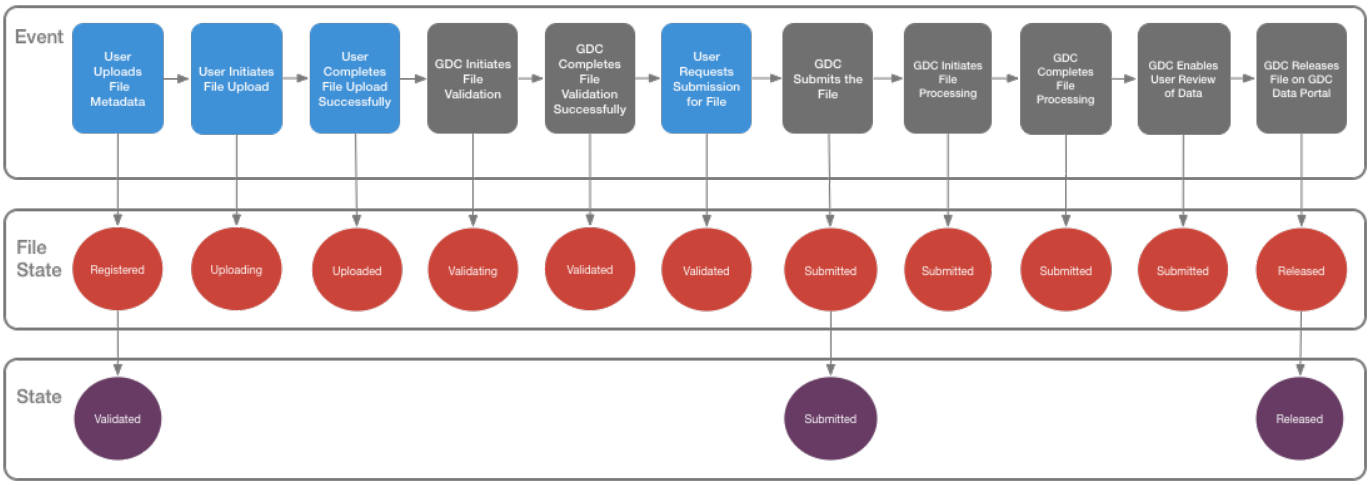
The diagram of the project lifecycle below demonstrates the transition of a project through the various states. Initially the project is open for data upload and validation. Any changes to the data must be made while the project status is open. When the data is uploaded and ready for review, the submitter changes the project state to review. During the review state, the project is locked and additional data cannot be uploaded. If data changes are needed during the review period, the project has to be re-opened.

The process of Harmonization does not occur immediately after submitted files are uploaded. After the submission is complete and all the necessary data and files have been uploaded, the user submits the data to the GDC for processing through the GDC Data Harmonization Pipelines and the project state changes to submitted. When the data has been processed, the project state changes back to open for new data to be submitted to the project and the submitter can review the processed data. After review of the processed data, the submitter can then release the harmonized data to the GDC Data Portal and other GDC Data Access Tools according to GDC Data Sharing Policies.



File Lifecycle

This section describes states pertaining to submittable data files throughout the data submission process. A submittable data file could contain data such as genomic sequences (such as a BAM or FASTQ) or pathology slide images. The file lifecycle starts when a submitter uploads metadata for a file to the GDC Data Submission Portal. This metadata file registers a description of the file as an entity on the GDC, the status for this is known as "state" and is represented by **purple** cirlces. The submitter can then use the GDC Data Transfer Tool to upload the actual file, which is represeneted by **red** circles. Throughout the lifecycle, the file and its associated entity transition through various states from when they are initially registered through file submission and processing. The diagram below details these state transitions.


Event
User Uploads File Metadata
User Initiates File Upload
User Completes File Upload Successfully
GDC Initiates File Validation
GDC Completes File Validation Successfully
User Requests Submission for File
GDC Submits the File
GDC Initiates File Processing
GDC Completes File Processing
GDC Enables User Review of Data
GDC Releases File on GDC Data Portal
File State
Registered
Uploading
Uploaded
Validating
Validated
Validated
Submitted
Submitted
Submitted
Submitted
Released
State
Validated
Submitted
Released

1.3 Data Submission Portal

1.3.1 Overview

This section will walk users through the submission process using the GDC Data Submission Portal to upload files to the GDC.

1.3.2 Authentication

Requirements

Accessing the GDC Data Submission Portal requires eRA Commons credentials with appropriate dbGaP authorization. To learn more about obtaining the required credentials and authorization, see Obtaining Access to Submit Data.

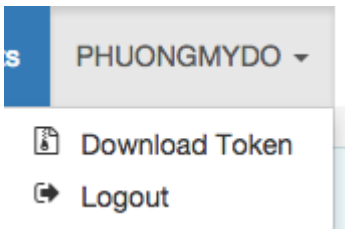
Authentication via eRA Commons

Users can log into the GDC Data Submission Portal with eRA Commons credentials by clicking the "Login" button. If authentication is successful, the user will be redirected to the GDC Data Submission Portal front page and the user's eRA Commons username will be displayed in the upper right corner of the screen.

GDC AUTHENTICATION TOKENS

The GDC Data Portal provides authentication tokens for use with the GDC Data Transfer Tool or the GDC API. To download a token:

1. Log into the GDC using your eRA Commons credentials.
2. Click the username in the top right corner of the screen.
3. Select the "Download Token" option.



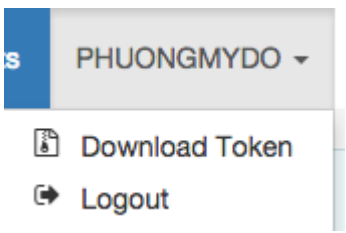
A new token is generated each time the `Download Token` button is clicked.

For more information about authentication tokens, see Data Security.

NOTE: The authentication token should be kept in a secure location, as it allows access to all data accessible by the associated user account.

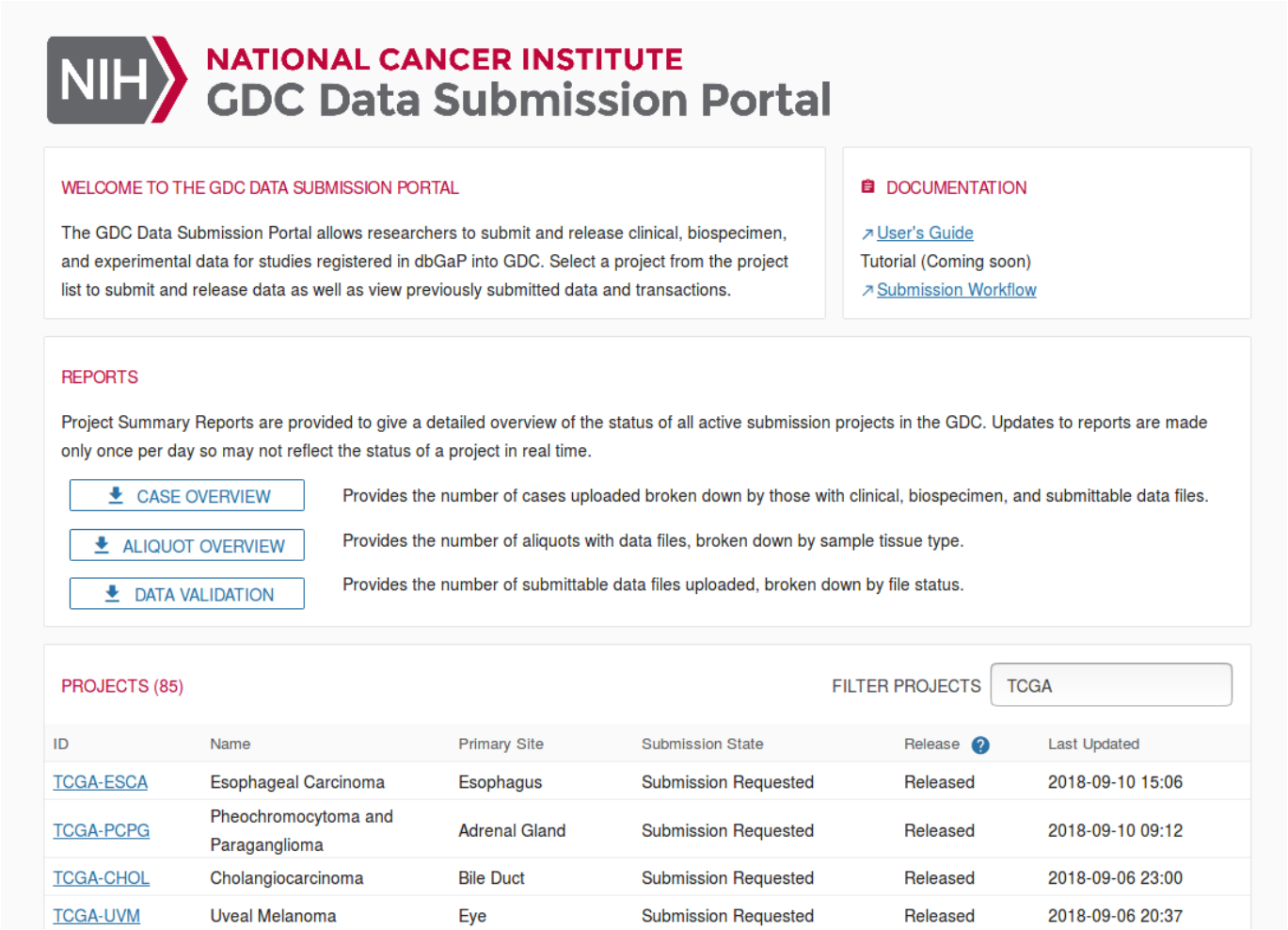
LOGGING OUT

To log out of the GDC, click the username in the top right corner of the screen, and select the Logout option. Users will automatically be logged out after 15 minutes of inactivity.



1.3.3 Homepage

After authentication, users are redirected to a homepage. The homepage acts as the entry point for GDC data submission and provides submitters with access to a list of authorized projects, reports, and transactions. Content on the homepage varies based on the user profile (e.g. submitter, program office).



Reports

Project summary reports can be downloaded at the Submission Portal homepage at three different levels: `CASE OVERVIEW`, `ALIQUOT OVERVIEW`, and `DATA VALIDATION`. Each report is generated in tab-delimited format in which each row represents an active project.

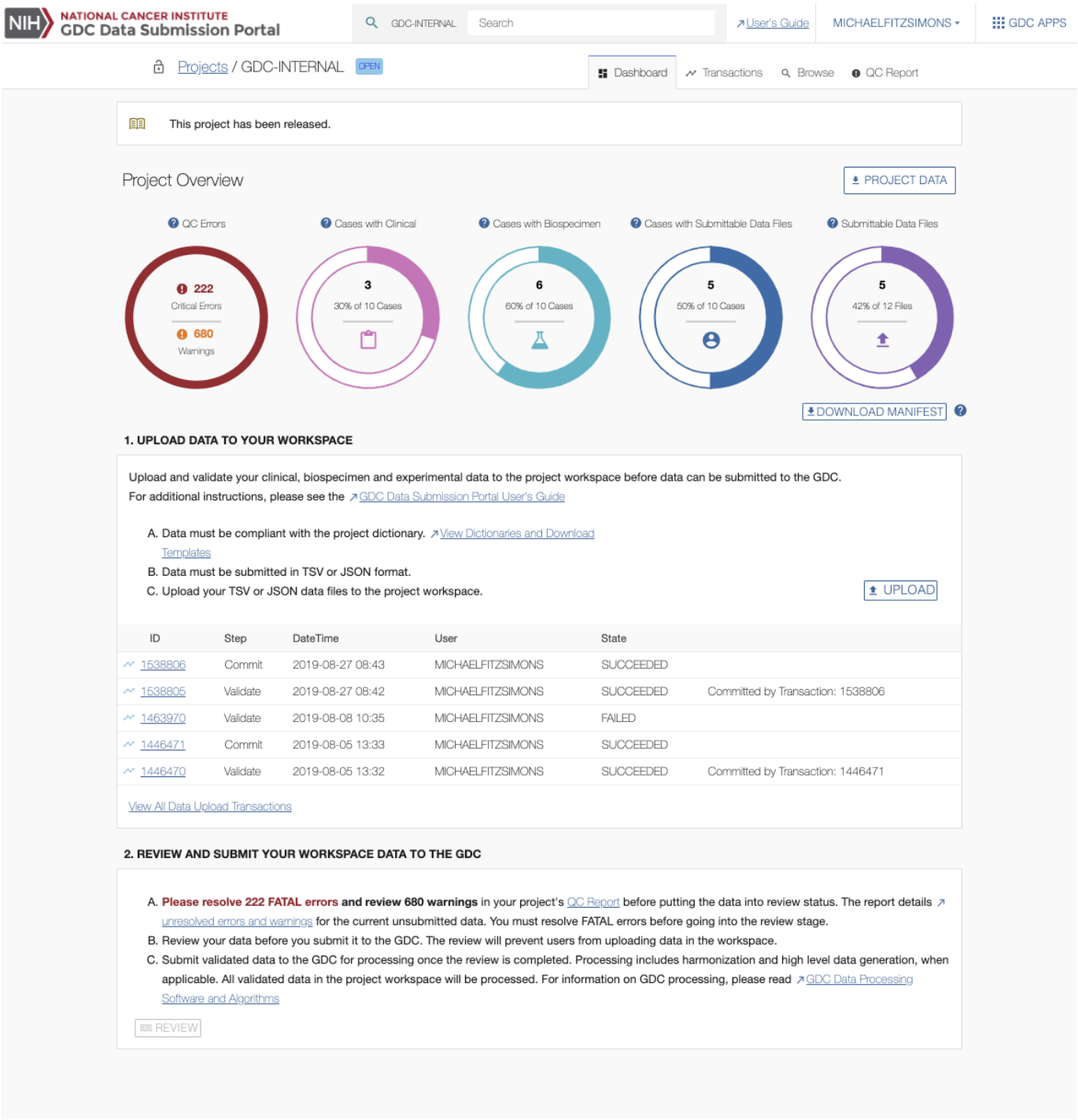
- **`CASE OVERVIEW`:** This report describes the number of cases with associated biospecimen data, clinical data, or submittable data files (broken down by data type) for each project.
- **`ALIQUOT OVERVIEW`:** This report describes the number of aliquots in a project with associated data files. Aliquot numbers are broken down by sample tissue type.
- **`DATA VALIDATION`:** This report categorizes all submittable data files associated with a project by their file status.

Projects

The projects section in the homepage lists the projects that the user has access to along with basic information about each project. For users with access to a large number of projects, this table can be filtered using the 'FILTER PROJECTS' field. Selecting a project ID will direct the user to the project's Dashboard. The button used to release data for each project is also located on this screen, see Release for details.

1.3.4 Dashboard

The GDC Data Submission Portal dashboard provides details about a specific project.



The dashboard contains various visual elements to guide the user through all stages of submission, from viewing the Data Dictionary, support of data upload, to submitting a project for harmonization.

To better understand the information displayed on the dashboard and the available actions, please refer to the Data Submission Walkthrough.

Project Overview

The Project Overview sections of the dashboard displays the most current project state (open / review / submitted / processing) and the GDC Release, which is the date in which the project was released to the GDC.

The search field at the top of the dashboard allows for submitted entities to be searched by partial or whole `submitter_id`. When a search term is entered into the field, a list of entities matching the term is updated in real time. Selecting one of these entities links to its details in the Browse Tab.

The remaining part of the top section of the dashboard is broken down into four status charts:

- **QC Errors**: The number of errors found in the uploaded data. For more details please refer to the QC Report Section.
- **Cases with Clinical**: The number of `cases` for which Clinical data has been uploaded.
- **Cases with Biospecimen**: The number of `cases` for which Biospecimen data has been uploaded.
- **Cases with Submittable Data Files**: The number of `cases` for which experimental data has been uploaded.
- **Submittable Data Files**: The number of registered submittable data files that have been successfully uploaded through the GDC Data Transfer Tool. Totals do not include files that have been submitted for harmonization. For more information on this status chart, please refer to File Lifecycle.
- **`DOWNLOAD MANIFEST:`** This button below the status chart allows the user to download a manifest for registered files in this project that have not yet been uploaded.

Action Panels

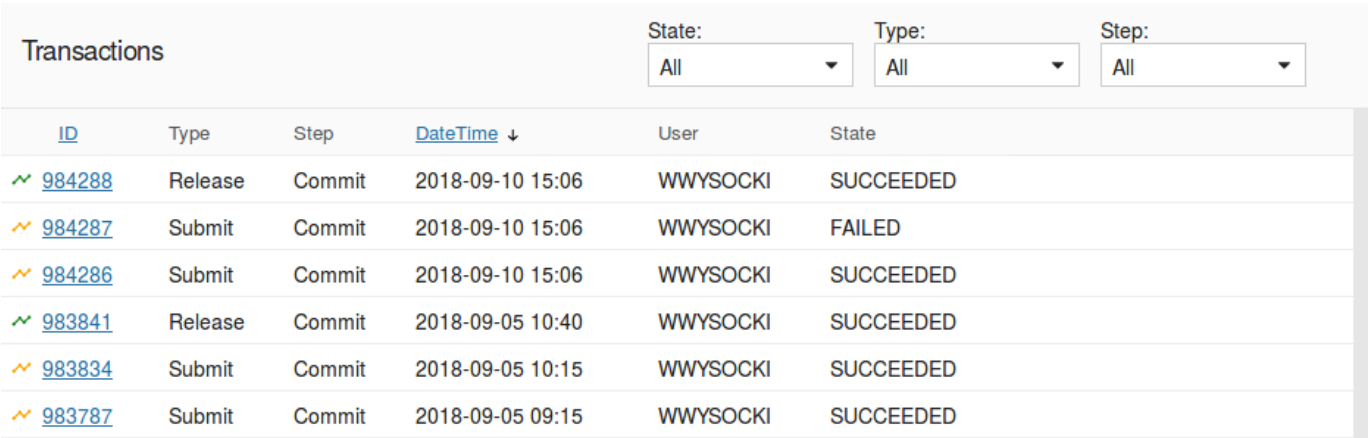
There are two action panels available below the Project Overview.

- UPLOAD DATA TO YOUR WORKSPACE: Allows a submitter to upload project data to the GDC project workspace. The GDC will validate the uploaded data against the GDC Data Dictionary. This panel also contains a table that displays details about the five latest transactions. Clicking the IDs in the first column will bring up a window with details about the transaction, which are documented in the transactions page. This panel will also allow the user to commit file uploads to the project.
- REVIEW AND SUBMIT YOUR WORKSPACE DATA TO THE GDC: Allows a submitter to review project data which will lock the project to ensure that additional data cannot be uploaded while in review. Once the review is complete, the data can be submitted to the GDC for processing through the GDC Harmonization Process.

These actions and associated features are further detailed in their respective sections of the documentation.

1.3.5 Transactions

The transactions page lists all of the project's transactions. The transactions page can be accessed by choosing the Transactions tab at the top of the dashboard or by choosing "View All Data Upload Transactions" in the first panel of the dashboard.



ID	Type	Step	DateTime ↓	User	State
984288	Release	Commit	2018-09-10 15:06	WWYSOCKI	SUCCEEDED
984287	Submit	Commit	2018-09-10 15:06	WWYSOCKI	FAILED
984286	Submit	Commit	2018-09-10 15:06	WWYSOCKI	SUCCEEDED
983841	Release	Commit	2018-09-05 10:40	WWYSOCKI	SUCCEEDED
983834	Submit	Commit	2018-09-05 10:15	WWYSOCKI	SUCCEEDED
983787	Submit	Commit	2018-09-05 09:15	WWYSOCKI	SUCCEEDED

The types of transactions are the following:

- **Upload:** The user uploads data to the project workspace. Note that submittable data files uploaded using the GDC Data Transfer tool do not appear as transactions. Uploaded submittable data can be viewed in the Browse tab.
- **Delete:** The user deletes data from the project workspace.
- **Review:** The user reviews the project before submitting data to the GDC.
- **Open:** The user re-opens the project if it was under review. This allows the upload of new data to the project workspace.
- **Submit:** The user submits uploaded data to the GDC. This triggers the data harmonization process.
- **Release:** The user releases harmonized data to be available through the GDC Data Portal and other GDC data access tools.

Transactions List View

The transactions list view displays the following information:

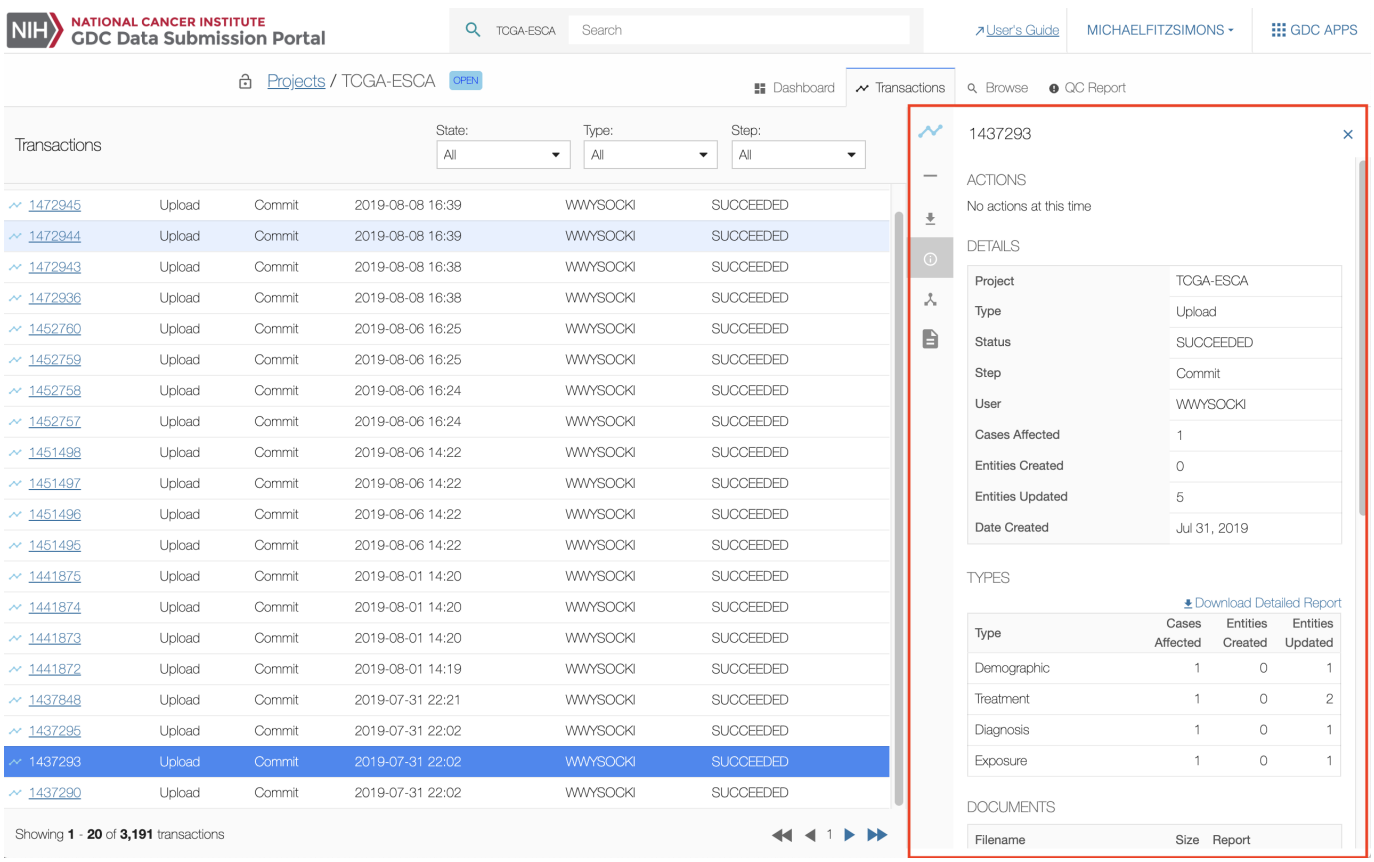
Column	Description
ID	Identifier of the transaction
Type	Type of the transaction (see the list of transaction types in the previous section)
Step	The step of the submission process that each file is currently in. This can be Validate or Commit. "Validate" represents files that have not yet been committed but have been uploaded using the submission portal or the API.
DateTime	Date and Time that the transaction was initiated
User	The username of the submitter that performed the transaction
State	Indicates the status of the transaction: `SUCCEEDED` , `PENDING` , or `FAILED`
Commit/Discard	Two buttons appear when data has been uploaded using the API or the submission portal. This allows for validated data to be incorporated into the project or discarded. This column will then display the transaction number for commited uploads and "Discarded" for the uploads that are discarded.

Transaction Filters

Choosing from the drop-down menu at the top of the table allows the transactions to be filtered by those that are in progress, to be committed, succeeded, failed, or discarded. The drop-down menu also allows for the transactions to be filtered by type and step.

Transactions Details

Clicking on a transaction will open the details panel. Data in this panel is organized into multiple sections including actions, details, types, and documents as described below.



Navigation between the sections can be performed by either scrolling down or by clicking on the section icon displayed on the left side of the details panel.

ACTIONS

The Actions section allows a user to perform an action for transactions that provide actions. For example, if a user uploads read groups and file metadata, a corresponding manifest file will be available for download from the transaction. This manifest is used to upload the actual files through the GDC Data Transfer Tool.

DETAILS

The Details section provides details about the transaction itself, such as its project, type, and number of affected cases.

DETAILS

Project	GDC-INTERNAL
Type	Upload
Status	SUCCEEDED
Step	Commit
User	--
Cases Affected	1
Entities Created	5
Entities Updated	0
Date Created	Aug 25, 2016

TYPES

The Types section lists the type of files submitted and the number of affected cases and entities.

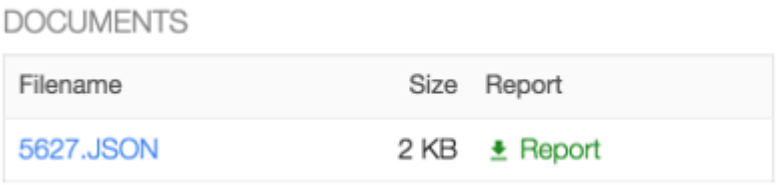
TYPES

Download Detailed Report

Type	Cases Affected	Entities Created	Entities Updated
Case	1	1	0
Sample	1	1	0
Aliquot	1	1	0
Read_group	1	1	0
Submitted_unaligned_reads	1	1	0

DOCUMENTS

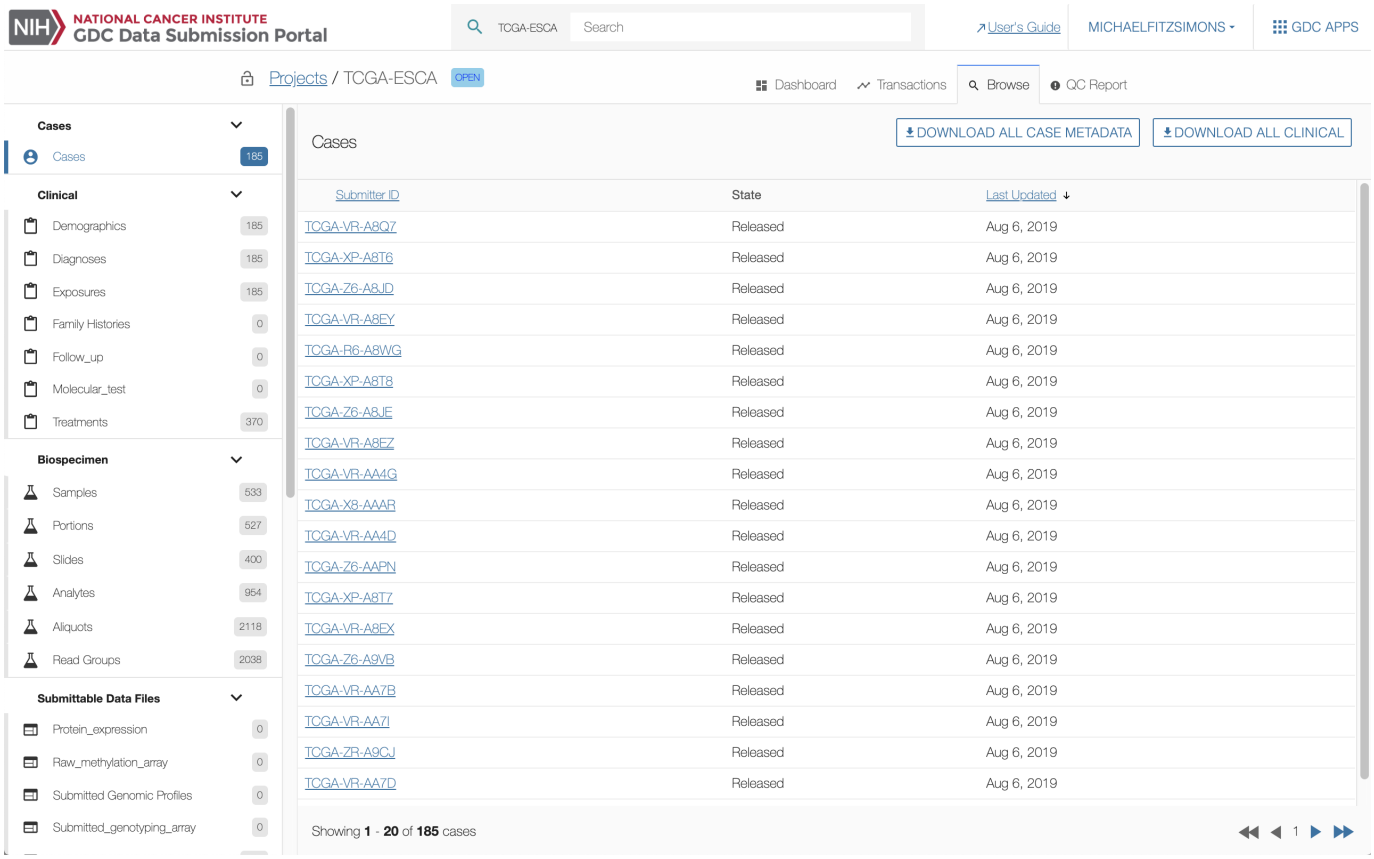
The Documents section lists the files submitted during the transaction. The user can download the original files from the transaction, a report detailing the transaction, or the errors that originated from the transaction that has failed.



1.3.6 Browse

The `Browse` menu provides access to all of a project's content. Most content is driven by the GDC Data Dictionary and the interface is dynamically generated to accommodate the content.

Please refer to the GDC Data Dictionary Viewer for specific details about dictionary-generated fields, columns, and filters.



Main Interface Elements

FILTERS

A wide set of filters are available for the user to select the type of entity to be displayed. These filters are dynamically created based on the GDC Data Dictionary.

Current filters are:

Filter	Description
Cases	Display all `Cases` associated with the project.
Clinical	Display all Clinical data uploaded to the project workspace. This is divided into subgroups including `Demographics`, `Diagnoses`, `Exposures`, `Family Histories`, `Follow_up`, `Molecular_tests`, and `Treatments`.
Biospecimen	Display all Biospecimen data uploaded to the project workspace. This is divided into subgroups including `Samples`, `Portions`, `Slides`, `Analytes`, `Aliquots`, and `Read Groups`.
Submittable Data Files	Displays all data files that have been registered with the project. This includes files that have been uploaded and those that have been registered but not uploaded yet. This category is divided into groups by file type.
Annotations	Lists all annotations associated with the project. An annotation provides an explanatory comment associated with data in the project.
Harmonized Data Files	Lists all data files that have been harmonized by the GDC. This category is divided into groups by generated data.

LIST VIEW

The list view is a paginated list of all entities corresponding to the selected filter.

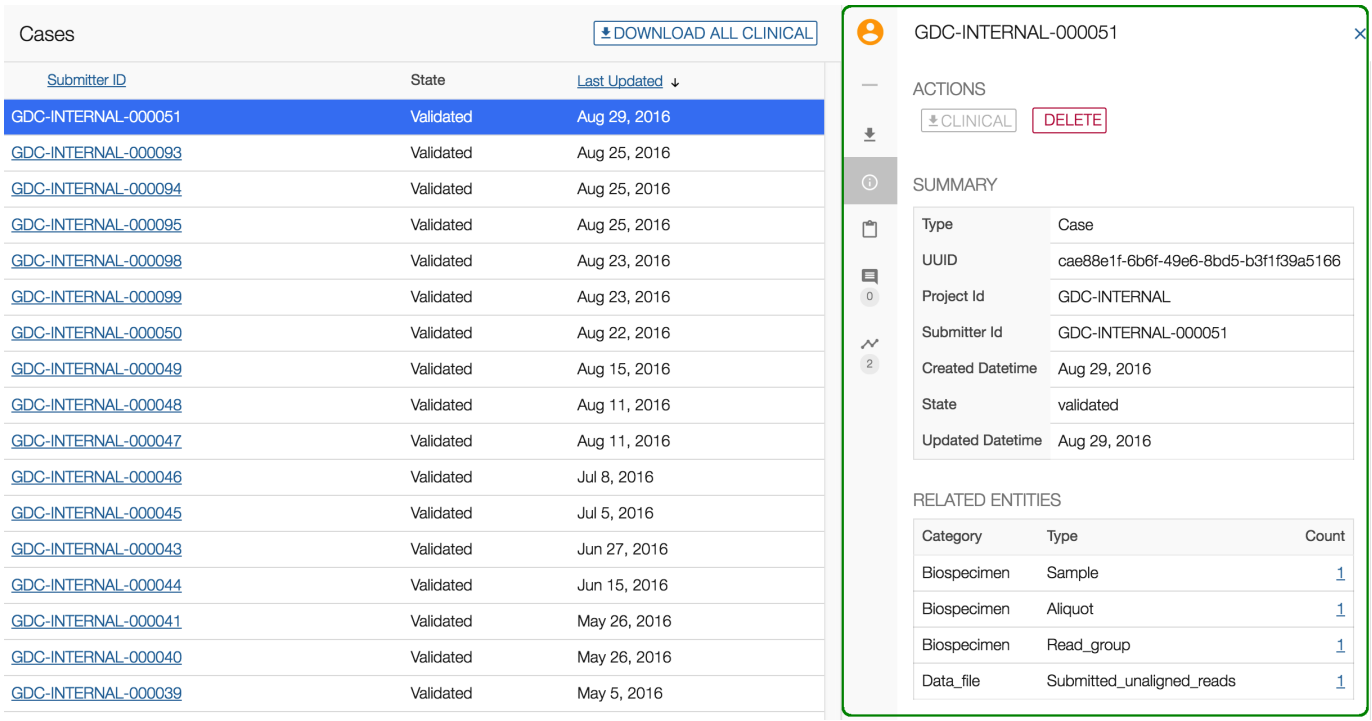
On the top-right section of the screen, the user can download data about all entities associated with the selected filter.

- For the case filter, it will download all Clinical data or all Metadata.
- For all other filters, it will download the corresponding metadata (e.g., for the `demographic` filter, it will download all `demographic` data).

DETAILS PANEL

Clicking on an entity will open the details panel. Data in this panel is broken down into multiple sections depending on the entity type. The main sections are:

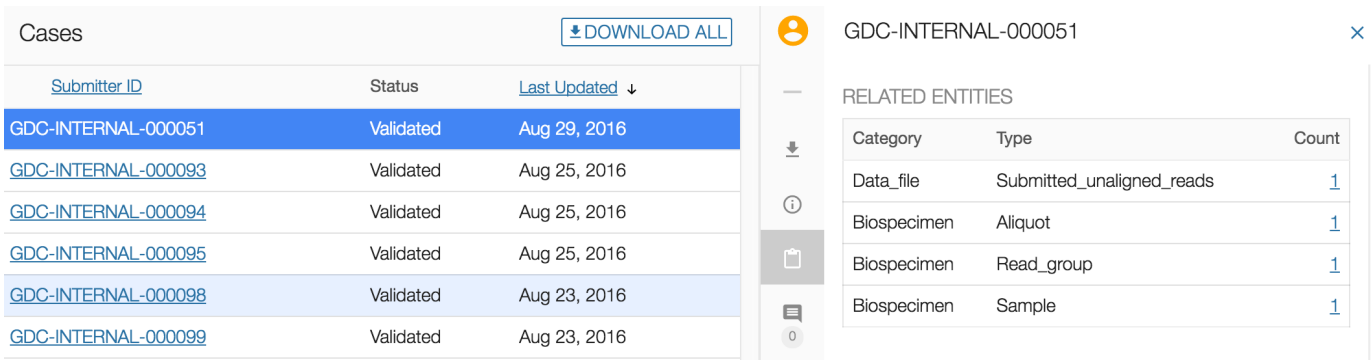
- **Actions**: Actions that can be performed relating the entity. This includes downloading the metadata (JSON or TSV) or submittable data file pertaining to the entity and deleting the entity. See the Deleting Entities guide for more information.
- **Summary**: A list of IDs and system properties associated with the entity.
- **Details**: Properties of the entity (not associated with cases).
- **Hierarchy** or **Related Entities**: A list of associated entities.
- **Annotations**: A list of annotations associated with the entity.
- **Transactions**: A list of previous transactions that affect the entity.



The sections listed above can be navigated either by scrolling down or by clicking on the section icon on the left side of the details panel.

RELATED ENTITIES

The Related Entities table lists all entities, grouped by type, related to the selected `case`. This section is only available at the `case` level.



This table contains the following columns:

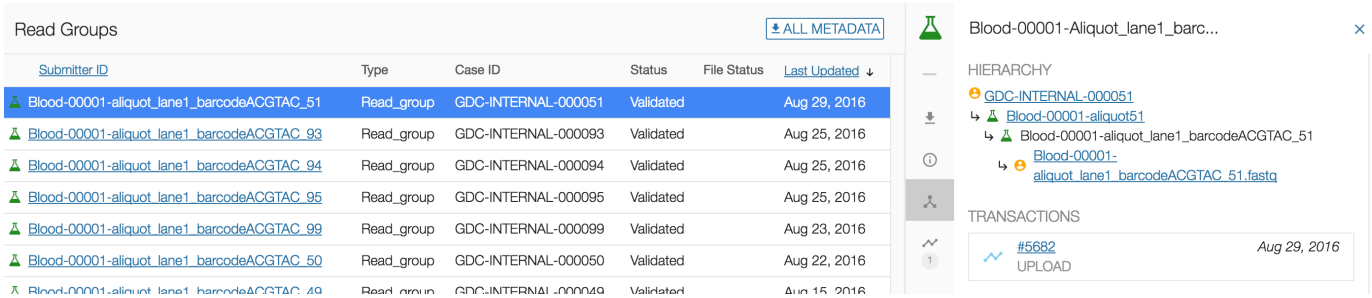
- **Category**: category of the entity (Clinical, Biospecimen, submittable data file).
- **Type**: type of entity (based on Data Dictionary).
- **Count:** number of occurrences of an entity associated with the `case`. Clicking on the count will open a window listing those entities within the Browse page.

HIERARCHY

The hierarchy section is available for entities at any level (e.g., Clinical, Biospecimen, etc.), except for `case`. The user can use the hierarchy section to navigate through entities.

The hierarchy shows:

- The `case` associated with the entity.
- The **direct** parents of the entity.
- The **direct** children of the entity.

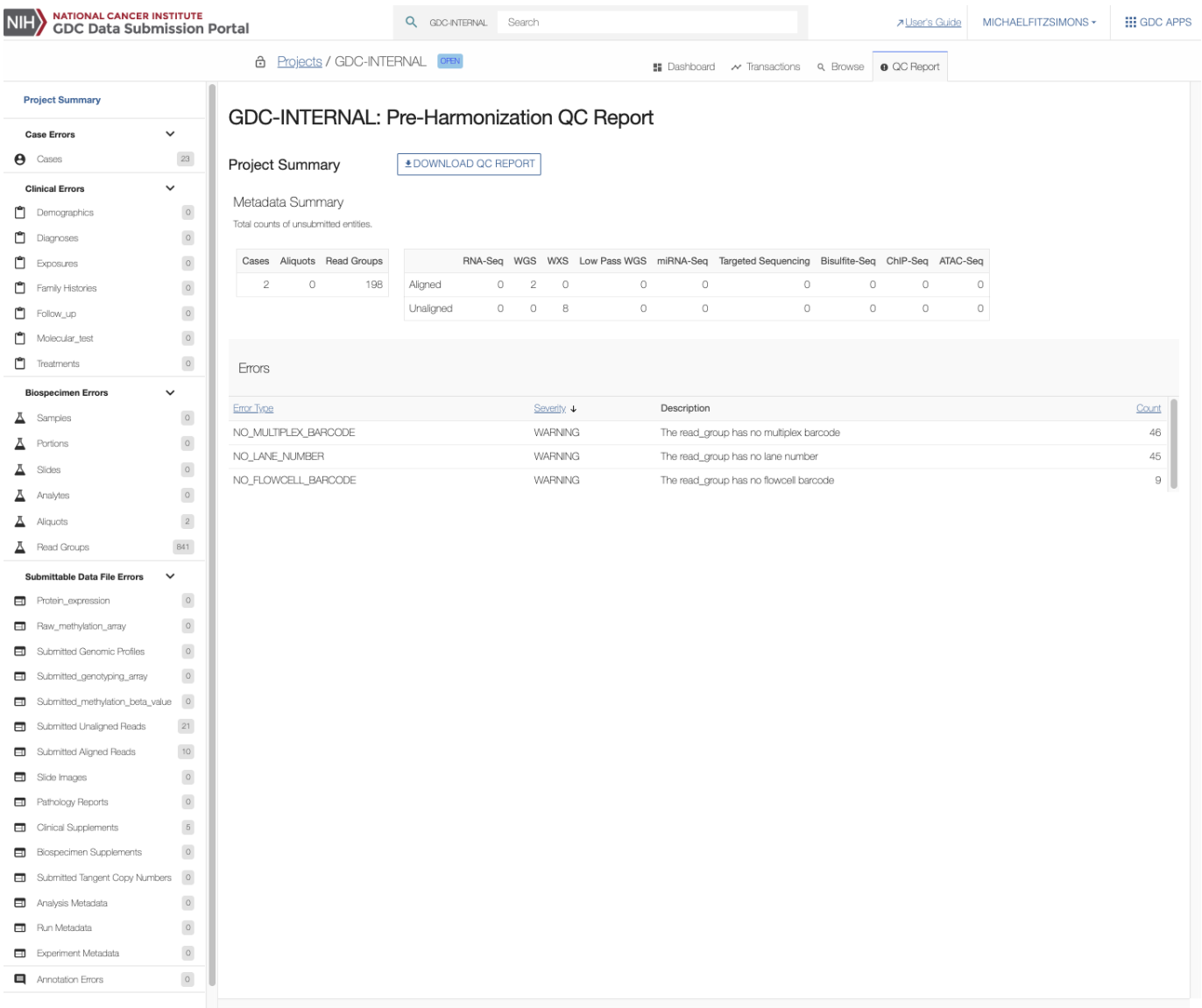


After uploading data to the workspace on the GDC Data Submission Portal, data will need to be reviewed by the submitter and then submitted to the GDC for processing.

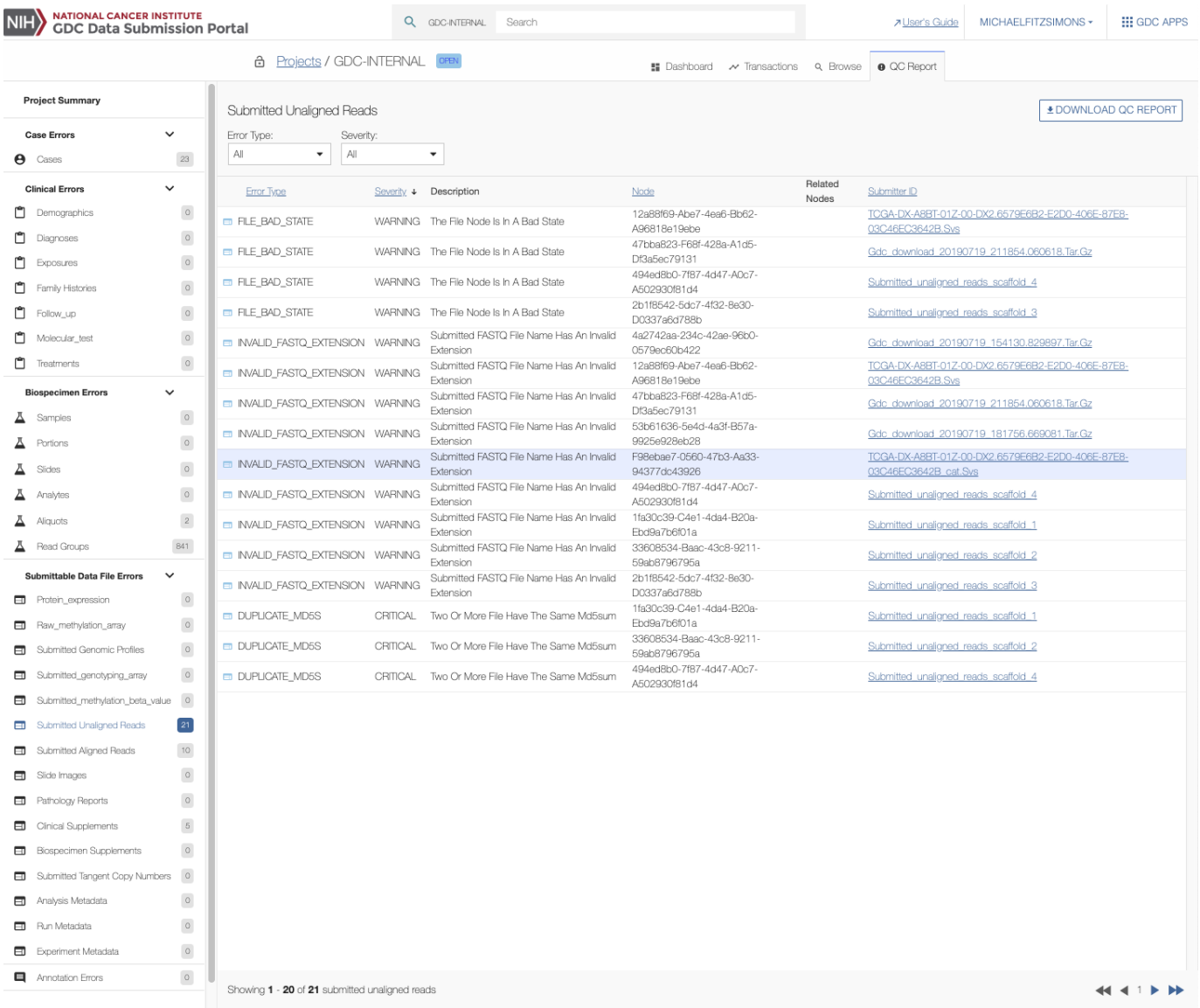
1.3.7 QC Reports

The QC Reports section allows users to see errors identified by the GDC for the current data that has not yet been submitted for harmonization. This includes all nodes in state `validated`. Data with error type `Critical` indicates errors that must be fixed before a submitter can Request Harmonization. Errors with error type `Warning` should be reviewed by the submitter as they may indicate discrepancies or problematic data.

You can see in the QC Reports Tab highlights of what data are present and the types of errors found in the project.



To find specific details for any node that contains errors you can click on the facet panel on the left to see those errors and to download a list of errors for that respective node. All potential errors are listed in the Pre-harmonization Checklist.



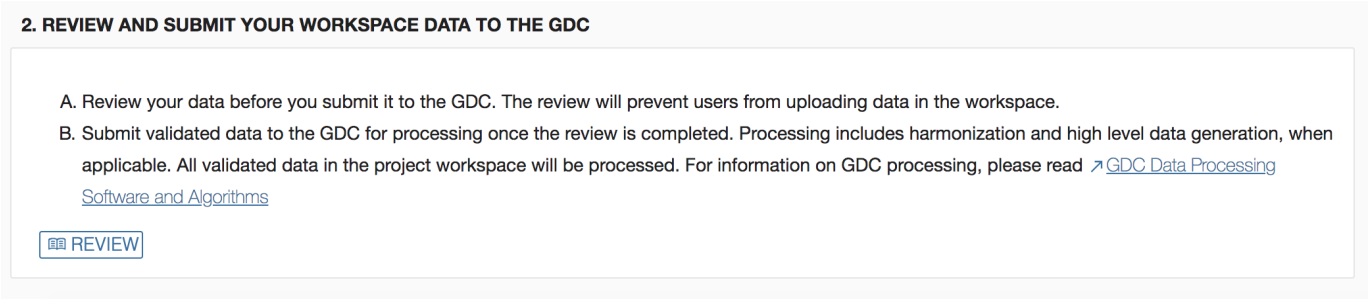
1.3.8 Submit Your Workspace Data to the GDC

The GDC Data Submission process is detailed on the Data Submission Processes and Tools section of the GDC Website.

Review

The submitter is responsible for reviewing the data uploaded to the project workspace (see Data Submission Walkthrough), and ensuring that it is ready for processing by the GDC Harmonization Process.

The user will be able to view the section below on the dashboard. The `REVIEW` button is available only if the project is in "OPEN" state.



Setting the project to the "REVIEW" state will lock the project and prevent users from uploading additional data. During this period, the submitter can browse the data in the Data Submission Portal or download it. Once the review is complete, the user can request to submit data to the GDC.

Once the user clicks on `REVIEW`, the project state will change to "REVIEW":

Pre-Harmonization Checklist

The Harmonization step is **NOT** an automatic process that occurs when data is uploaded to the GDC. The GDC performs batch processing of submitted data for Harmonization only after verifying that the submission is complete.

QC checks are automatically run on all supplied metadata and data files. The results are displayed within the QC Reports. These errors fall into two categories: Critical or Warning. If an error is deemed Critical it must be resolved before a submitter can request harmonization. If an error is categorized as Warning then the submitter should review this to verify the data have been submitted correctly. A list of the errors and their meanings are found in the table below:

CRITICAL ERRORS

Error Message	Description	How to Fix / Error Meaning
INVALID_CHARACTER	This entity submitter_id includes invalid characters	Upload new entity without invalid characters. The acceptable characters are alphanumeric characters [a-z, A-Z, 0-9] and `_`, `.`, `-`. Any other characters will interfere with the Harmonization workflow.
MORE_THAN_ONE_SAMPLE_TYPE	The aliquot is associated with more than one sample type	Ensure there is no `aliquot` attached to multiple `sample` nodes of more than one sample_type.
TWO_NODE_TYPES	The aliquot is associated with two or more node types	Ensure aliquot is only connected to a single type of node.
PE_FASTQ_FILE_COUNT	The number of FASTQ files for PE readgroup is not 2	Ensure that if a read group is paired end, that it has two FASTQ files. For the `read_group` node, make sure that the `is_paired_end` is set to `true` for paired end sequencing and `false` for single end sequencing.
SE_FASTQ_FILE_COUNT	The number of FASTQ files for SE readgroup is not 1	Ensure that if a read group is single end, that it has one FASTQ file. For the `read_group` node, make sure that the `is_paired_end` is set to `true` for paired end sequencing and `false` for single end sequencing.
CAPTURE_KIT_INADEQUATE	WXS/Targeted Sequencing ReadGroup lacks valid target capture kit	Modify read group entity to have a valid target capture kit from data dictionary. The `target_capture_kit` property is completed when the selected `library_strategy` is `WXS`. Errors will occur if `Not Applicable` or `Unknown` is selected.
TARGET_SEQ_LIBRARY_SELECTION	ReadGroup has library strategy Targeted Sequencing but does not have PCR or Hybrid Selection as its library selection	If library strategy is Target Sequencing, modify library selection to be either PCR or Hybrid Selection
WXS_LIBRARY_SELECTION	ReadGroup has library strategy WXS but does not have Hybrid Selection as its library selection	Modify library selection to be Hybrid Selection for WXS read groups
WGS_LIBRARY_SELECTION	ReadGroup has library strategy WGS but does not have Random as its library selection	For WGS read groups, ensure library strategy is set to Random
NO_READ_PAIR_NUMBER	The FASTQ is paired but has no read_pair_number	Include a read_pair_number for paired end FASTQ files
DUPLICATE_MD5S	Two or more files have the same md5sum	This means there are duplicate files in the submission. You must delete one of these files
SAMPLE_FIELD_MISMATCH_NORMAL	The tumor_descriptor field should be populated with 'Not Applicable' for normal samples.	Set tumor_descriptor to 'Not Applicable' for any normal sample.

WARNING ERRORS

Error Message	Description	How to Fix / Error Meaning
FILE_BAD_STATE	The file node is in a bad state	There are some files in a bad file_state. All files that are registered must been uploaded and validated. If file_state is `Error` You will have to delete the file using the data transfer tool, and re-upload it, or upload a file if the state is `Registered`
INCONSISTENT_READGROUPS	ReadGroups sharing a library_strategy under a given aliquot have properties that do not match	Verify the properties of shared read groups under the same aliquot are consistent.
NO_CLINICAL_SUPPLEMENT	The case has no associated clinical supplement	Upload an optional clinical supplement file. This is a file that contains clinical data about one or more cases in a user specified format
NO_BIOSPECIMEN_SUPPLEMENT	The case has no associated biospecimen supplement	Upload an optional biospecimen supplement file. This is a file that contains biospecimen data about one or more cases in a user specified format
NO_DEMOGRAPHIC	The case has no associated demographic information	Provide demographic information on the case. This will be required before data can be released.
NO_DIAGNOSIS	The case has no associated diagnosis information	Provide diagnosis information on the case. This will be required before data can be released.
MORE_THAN_ONE_SAMPLE	The aliquot is associated with more than one sample	Review whether multiple samples were actually combined to make a single aliquot. This is uncommon, but potentially correct.
MULTIPLE_ALIGNED_BAMS	The read_group has multiple submitted aligned BAMs	Review whether one read group actually appears in multiple BAM files. This is uncommon.
NO_MULTIPLEX_BARCODE	The read_group has no multiplex barcode	Provide multiplex barcode for the read_group.
NO_FLOWCELL_BARCODE	The read_group has no flowcell barcode	Provide flowcell barcode for the read_group
NO_LANE_NUMBER	The read_group has no lane number	Provide lane number for the read_group
MULTIPLE_SARS_ON_ALIQUOT	Multiple submitted aligned reads of the same experimental strategy are associated with one aliquot.	Each `aliquot` node is only associated with one `submitted_aligned_reads` file of the same `experimental_strategy`.
FASTQ_UNKNOWN_EXTENSION	The FASTQ filename has an unknown extension	FASTQ file extension should be `.fq` or `.fq.gz`. Impermissible extensions are `tar.gz` and `tar`.
MULTIPLE_FASTQ_READGROUPS	Submitted FASTQ file has links to multiple read groups	Ensure `submitted_unaligned_reads` of data_format `FASTQ` is not linked to multiple `read_group` nodes.
INVALID_FASTQ_EXTENSION	Submitted FASTQ file name has an invalid extension	FASTQ file extension should be `.fq` or `.fq.gz`. Impermissible extensions are `tar.gz` and `tar`.
FASTQ_TOO_LARGE	FASTQ exceeds 10GB in size	The `submitted_unaligned_reads` file is larger than 10 GB.
NO_ASSOCIATED_FILES	ReadGroup has no associated genomic files	Ensure that all read groups have genomic files attached - or delete them if they are no longer needed
SAMPLE_FIELD_MISMATCH_TUMOR	The tumor_descriptor field should be specified for tumor samples.	Set the tumor_descriptor to a value for tumor samples if possible.

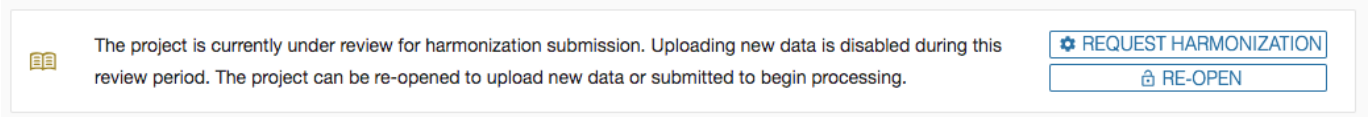
Error Message	Description	How to Fix / Error Meaning
CASE_LTFU_DAYS	If the patient is not lost to follow-up, the days to lost to follow-up question should not be answered.	Set days_to_lost_to_follow_up to 'null' if the patient was not lost to follow up.

Once user review is complete and all Critical errors are resolved, clicking the `REQUEST HARMONIZATION` button will indicate to the GDC Team and pipeline automation system that data processing can begin.

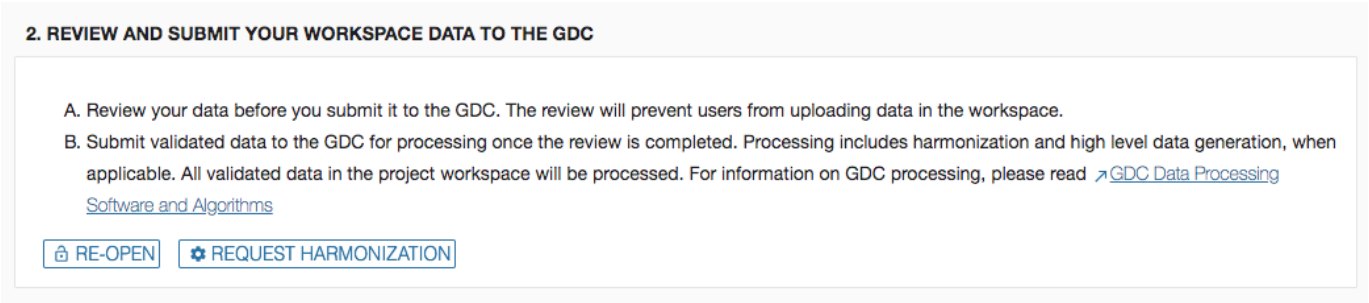
Submit to the GDC for Harmonization

When the project is ready for processing, the submitter will request to submit data to the GDC for Harmonization. If the project is not ready for processing, the project can be re-opened. Then the submitter will be able to upload more data to the project workspace.

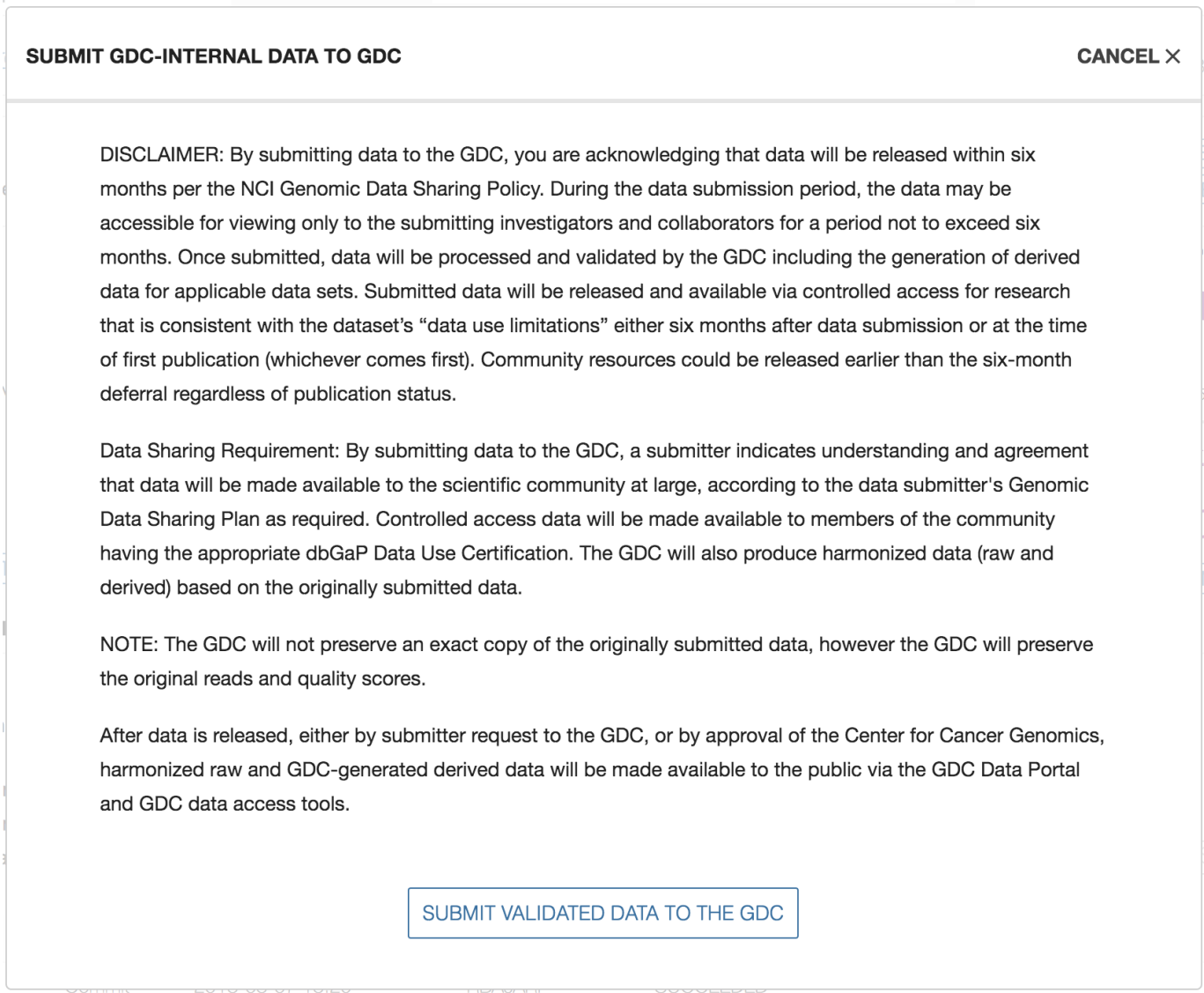
The `REQUEST HARMONIZATION` button is available only if the project is in "REVIEW" state. At this point, the user can decide whether to re-open the project to upload more data or to request harmonization of the data to the GDC. When the project is in "REVIEW" the following panel appears on the dashboard:



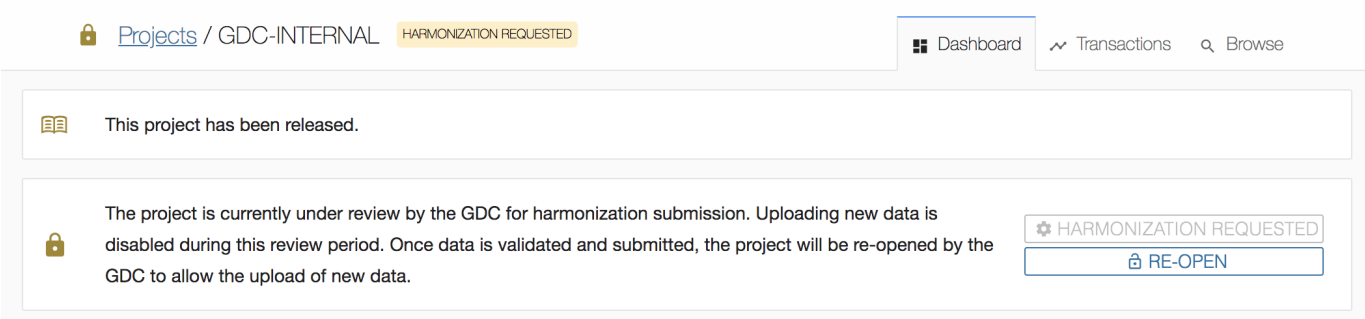
Once the user submits data to the GDC, they cannot modify the submitted nodes and files while harmonization is underway. Additional project data can be added during this period and will be considered a separate batch. To process an additional batch the user must again review the data and select `REQUEST HARMONIZATION`.



When the user clicks on the action `REQUEST HARMONIZATION` on the dashboard, the following popup is displayed:



After the user clicks on `SUBMIT VALIDATED DATA TO THE GDC`, the project state becomes "Harmonization Requested":



The GDC requests that users submit their data to the GDC for harmonization within six months from the first upload of data to the project workspace.

Reviewing Harmonized Data

After harmonization and prior to release, the GDC provides data submitters with access to their harmonized data. This allows the submitter to perform a check of the data, and let the GDC know if anything is incorrect before the data are released to the GDC Data Portal. How and in what detail the submitter wants to perform such a review is up to them, but here are a few suggestions for what a submitter may want to check.

Are all expected data present? More specifically, you could review the following questions: * Are the number of cases correct? * Are the number of cases associated with a given experimental strategy correct? * Are there any cases or experimental strategies I want to hold back that are still within the 6 month embargo period? * Does the clinical data appear as I expect? * Do the alignment statistics look acceptable? The GDC produces alignment metrics which are available via the API. This will allow users to see whether coverage, alignment, and other statistics are in line with expectations. The complete list can be found here.

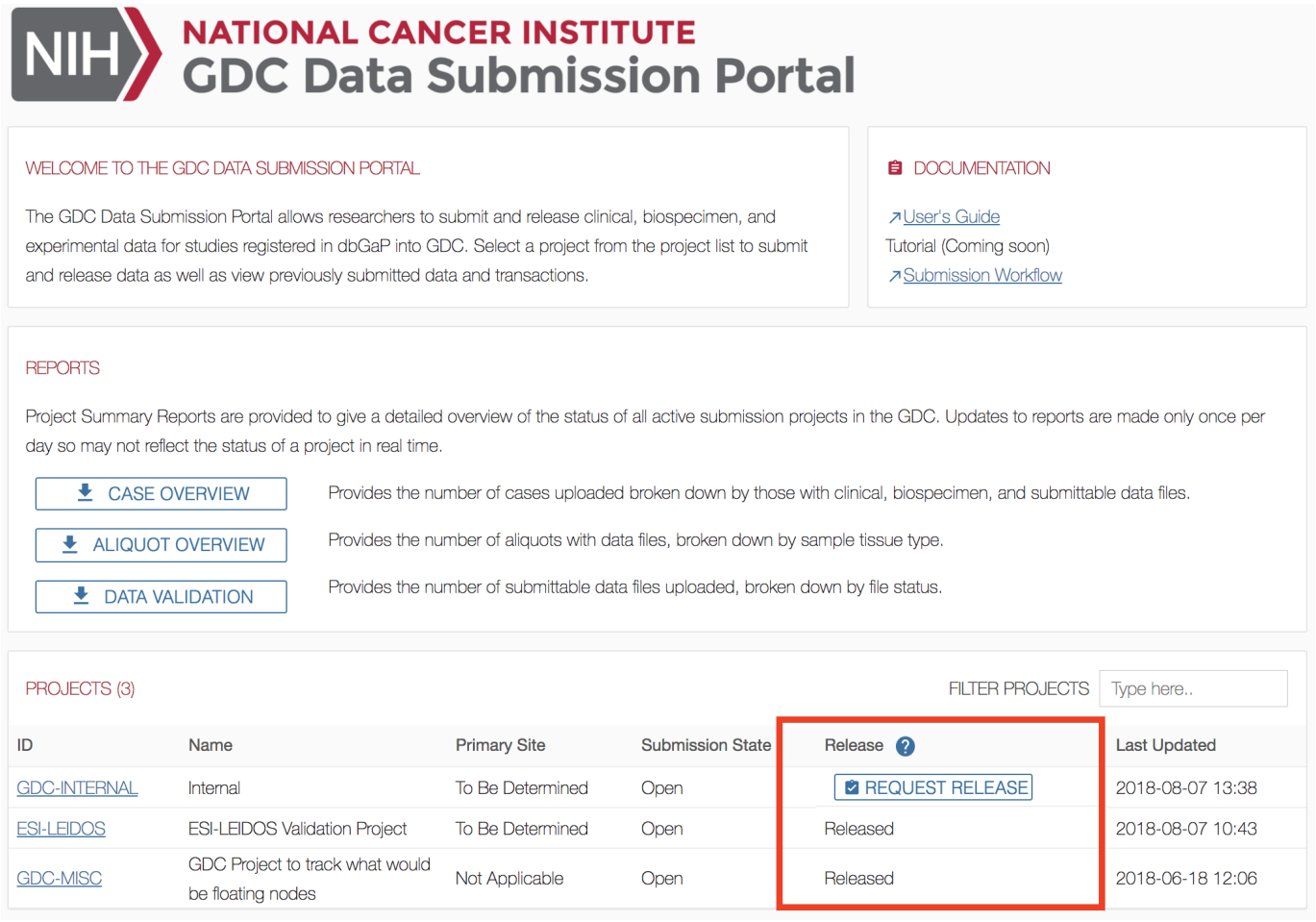
If users have access to other derived data files, like called variants or expression levels, there is another level of QC that is possible.

If you have access to this data you could also investigate the following: * Are expected variants present for a given tumor-normal pair? Note, due to differences between the GDC and user workflows (e.g. reference genome, variant calling pipelines, variant filtering, etc.) the exact list of variants may differ significantly between MAFs generated by users and those generated by the GDC. * Does gene expression correlate with previously generated expression data from the same aliquot? Note, the GDC performs non-stranded expression quantification for HTSeq workflows. To review strand-specific results please review STAR output.

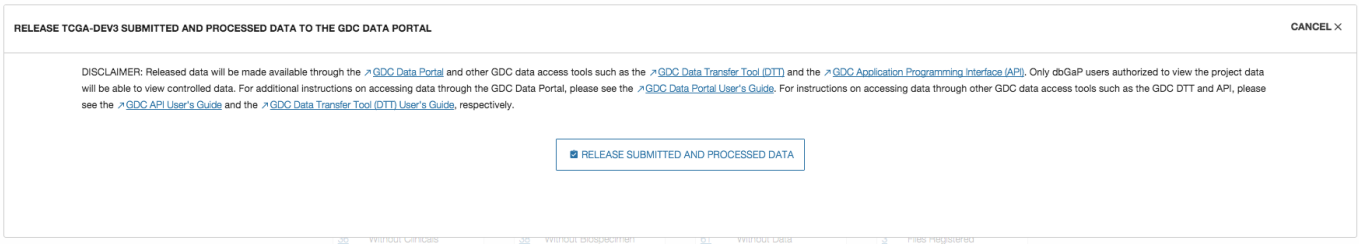
Once these user reviews have been completed, the user will need to contact the GDC and inform them that the project is ready for release.

1.3.9 Release

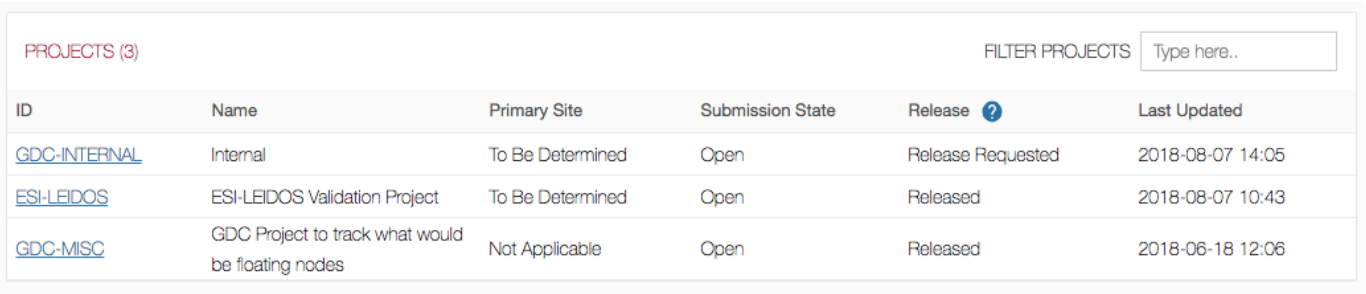
Project release occurs after the data has been harmonized, and allows users to access this data with the GDC Data Portal and other GDC Data Access Tools. The GDC will release data according to GDC Data Sharing Policies. Data must be released within six months after GDC data processing has been completed, or the submitter may request earlier release using the "Request Release" function. A project can only be released once.



When the user clicks on the action `REQUEST RELEASE`, the following Release popup is displayed:



After the user clicks on `RELEASE SUBMITTED AND PROCESSED DATA`, the project release state becomes "Release Requested":



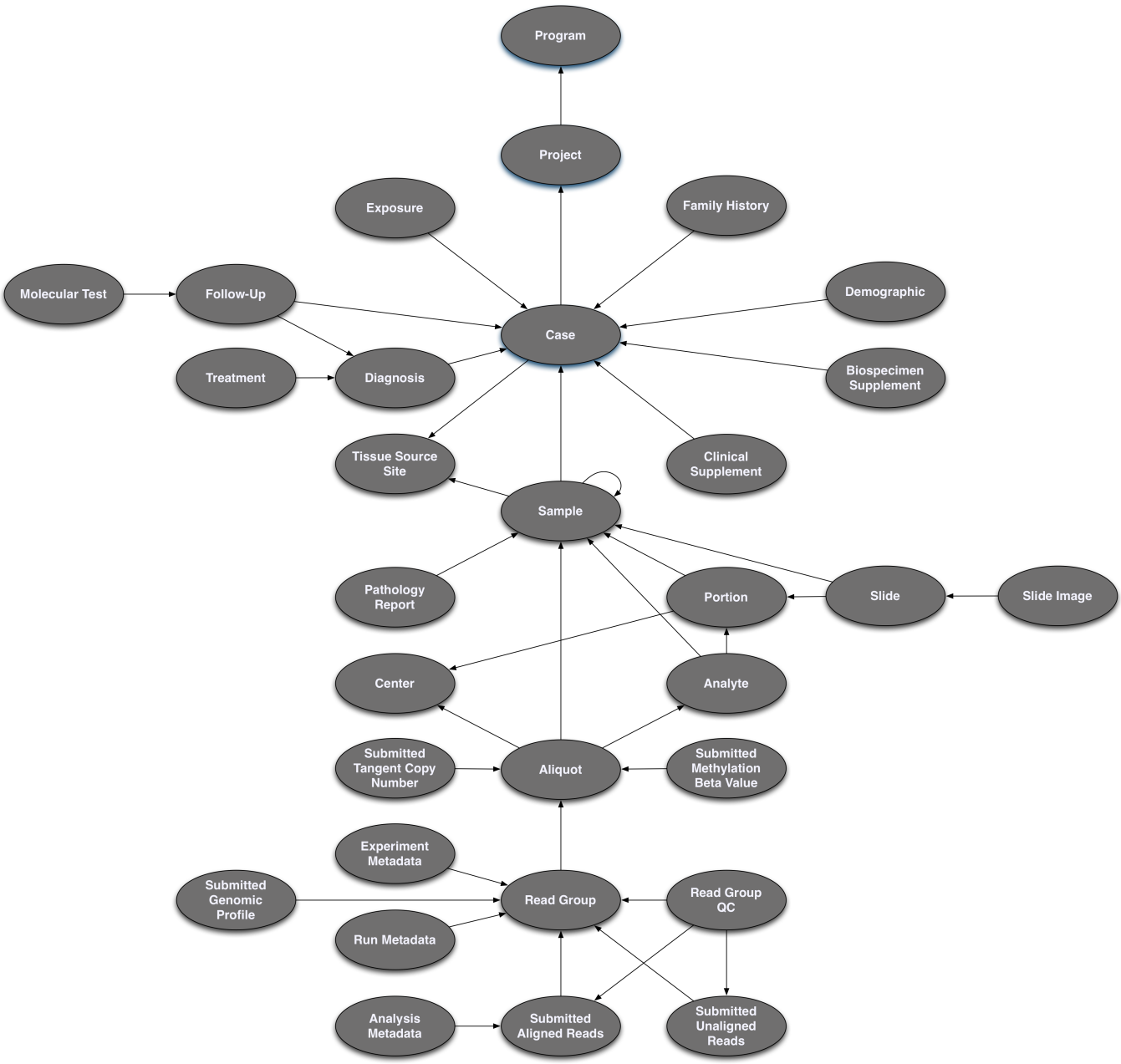
Note: Released cases and/or files can be redacted from the GDC. For more information, visit the GDC Policies page (under GDC Data Sharing Policies).

1.4 Data Submission Walkthrough

This guide details step-by-step procedures for different aspects of the GDC Data Submission process and how they relate to the GDC Data Model and structure. The first sections of this guide break down the submission process and associate each step with the Data Model. Additional sections are detailed below for strategies on expediting data submission, using features of the GDC Data Submission Portal, and best practices used by the GDC.

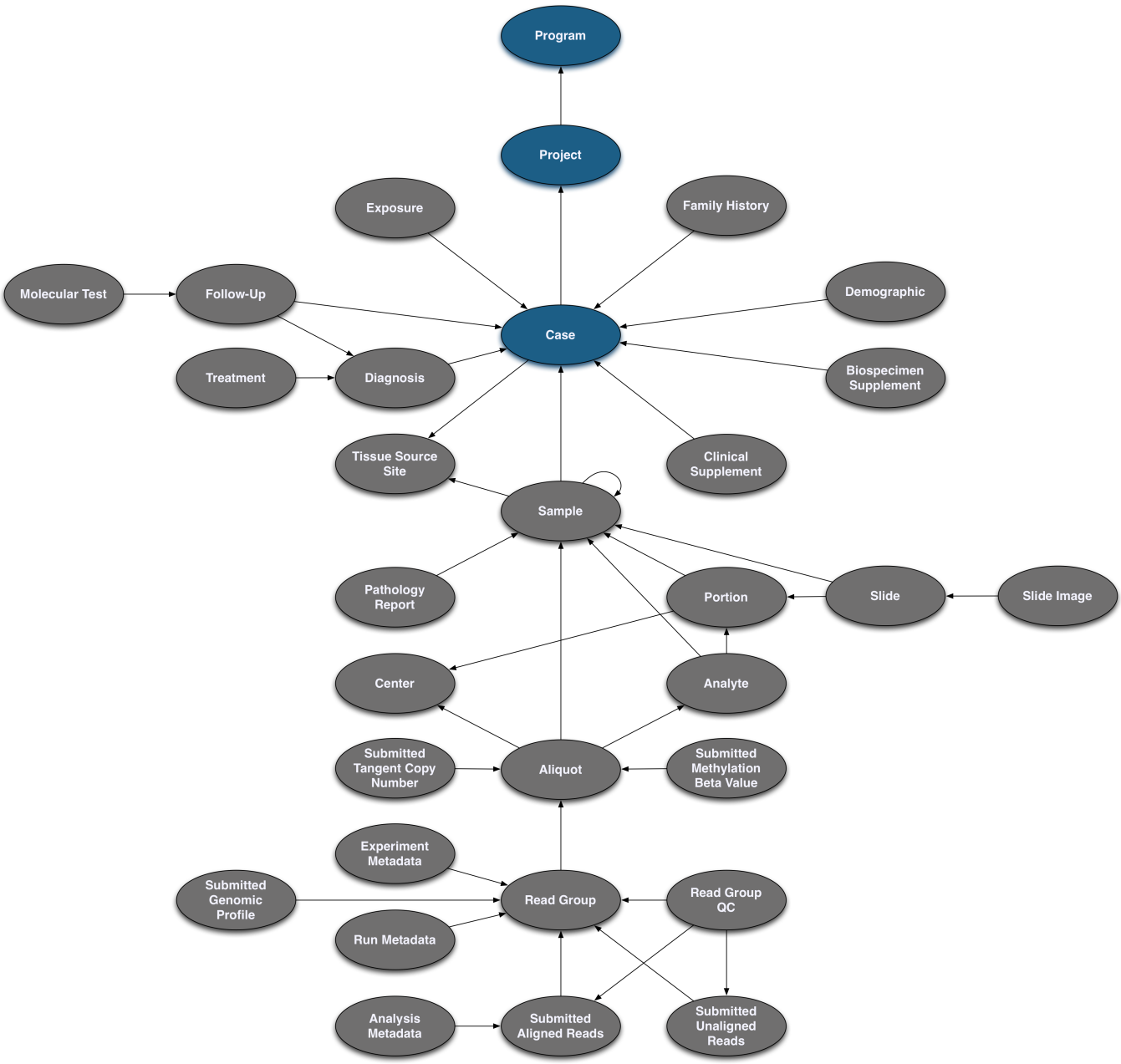
1.4.1 GDC Data Model Basics

Pictured below is the submittable subset of the GDC Data Model: a roadmap for GDC data submission. Each oval node in the graphic represents an entity: a logical unit of data related to a specific clinical, biospecimen, or file facet in the GDC. An entity includes a set of fields, the associated values, and information about its related node associations. All submitted entities require a connection to another entity type, based on the GDC Data Model, and a `submitter_id` as an identifier. This walkthrough will go through the submission of different entities. The completed (submitted) portion of the entity process will be highlighted in **blue**.

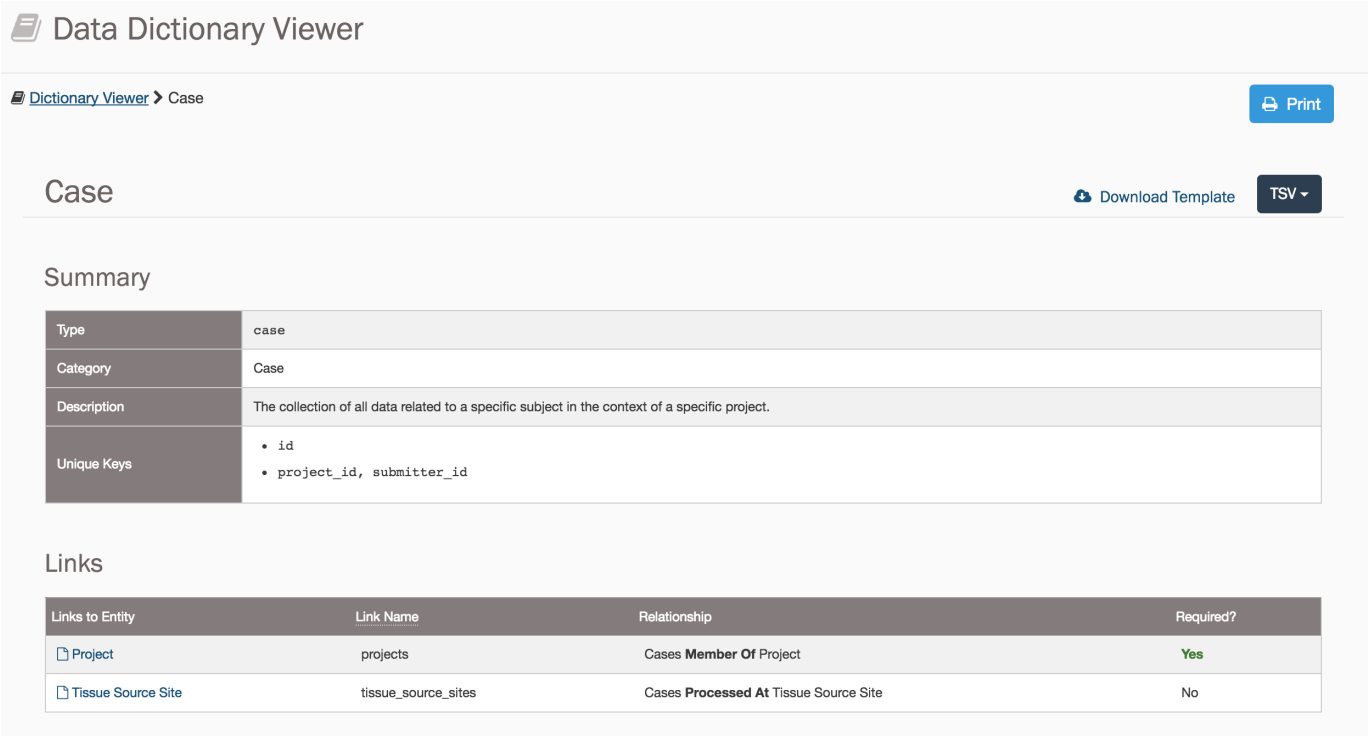


1.4.2 Case Submission

The `case` is the center of the GDC Data Model and usually describes a specific patient. Each `case` is connected to a `project`. Different types of clinical data, such as `diagnoses` and `exposures`, are connected to the `case` to describe the case's attributes and medical information.



The main entity of the GDC Data Model is the `case`, each of which must be registered beforehand with dbGaP under a unique `submitter_id`. The first step to submitting a `case` is to consult the Data Dictionary, which details the fields that are associated with a `case`, the fields that are required to submit a `case`, and the values that can populate each field. Dictionary entries are available for all entities in the GDC Data Model.



Submitting a **Case** entity requires:

- **`submitter_id`:** A unique key to identify the `case`
- **`projects.code`:** A link to the `project`

The submitter ID is different from the universally unique identifier (UUID), which is based on the UUID Version 4 Naming Convention. The UUID can be accessed under the `<entity_type>_id` field for each entity. For example, the `case` UUID can be accessed under the `case_id` field. The UUID is either assigned to each entity automatically or can be submitted by the user. Submitter-generated UUIDs cannot be uploaded in `submittable_data_file` entity types. See the Data Model Users Guide for more details about GDC identifiers.

The `projects.code` field connects the `case` entity to the `project` entity. The rest of the entity connections use the `submitter_id` field instead.

The `case` entity can be added in JSON or TSV format. A template for any entity in either of these formats can be found in the Data Dictionary at the top of each page. Templates populated with `case` metadata in both formats are displayed below.

JSON **TSV**

```
{
    "type": "case",
    "submitter_id": "PROJECT-INTERNAL-000055",
    "projects": {
        "code": "INTERNAL"
    }
}
```

```
type  submitter_id  projects.code
case  PROJECT-INTERNAL-000055 INTERNAL
```

Note: JSON and TSV formats handle links between entities (`case` and `project`) differently. JSON includes the `code` field nested within `projects` while TSV appends `code` to `projects` with a period.

Uploading the Case Submission File

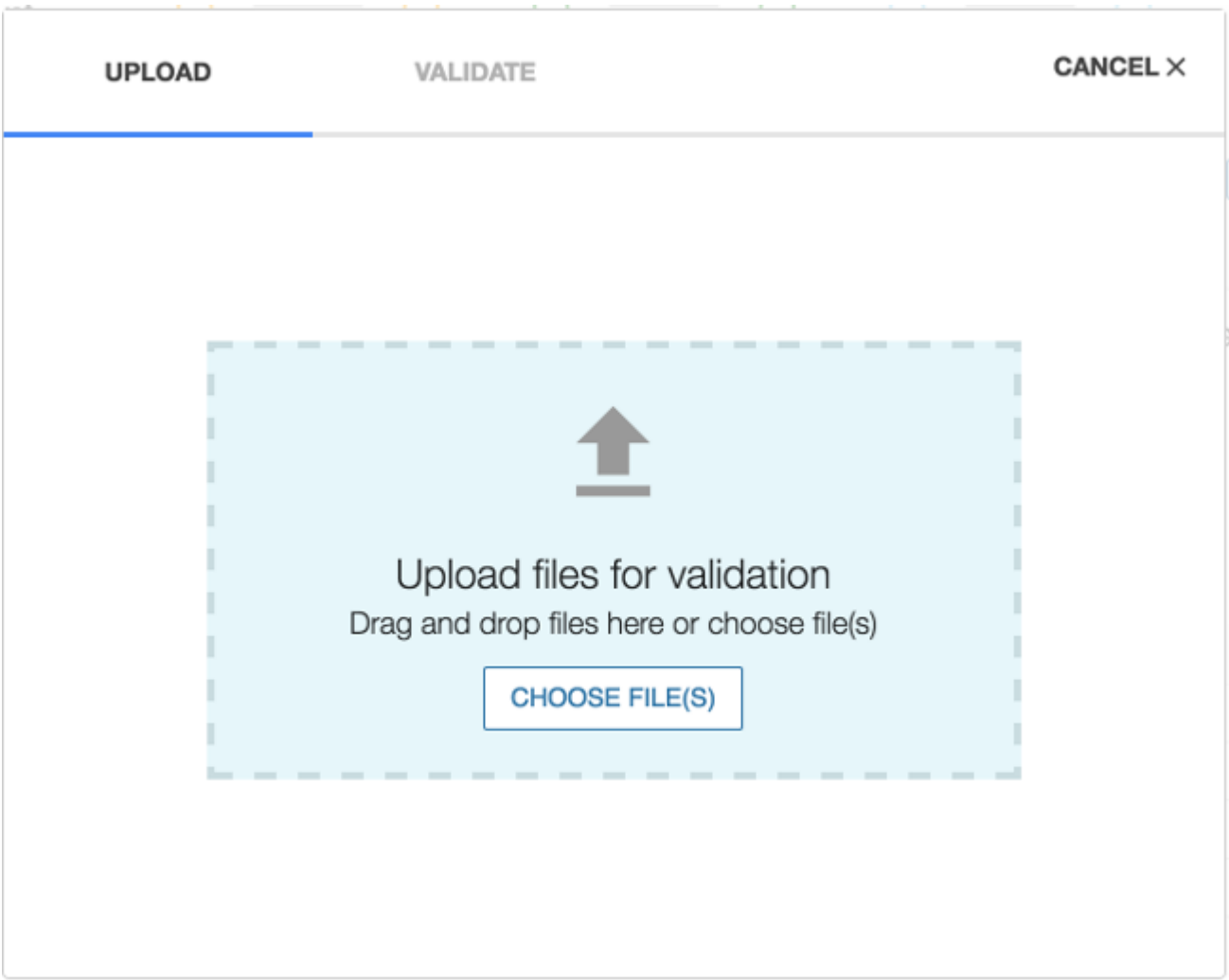
The file detailed above can be uploaded using the GDC Data Submission Portal and the GDC API as described below:

UPLOAD USING THE GDC DATA SUBMISSION PORTAL

An example of a `case` upload is detailed below. The GDC Data Submission Portal is equipped with a wizard window to facilitate the upload and validation of entities.

1. Upload Files

Choosing *'UPLOAD'* from the project dashboard will open the Upload Data Wizard.

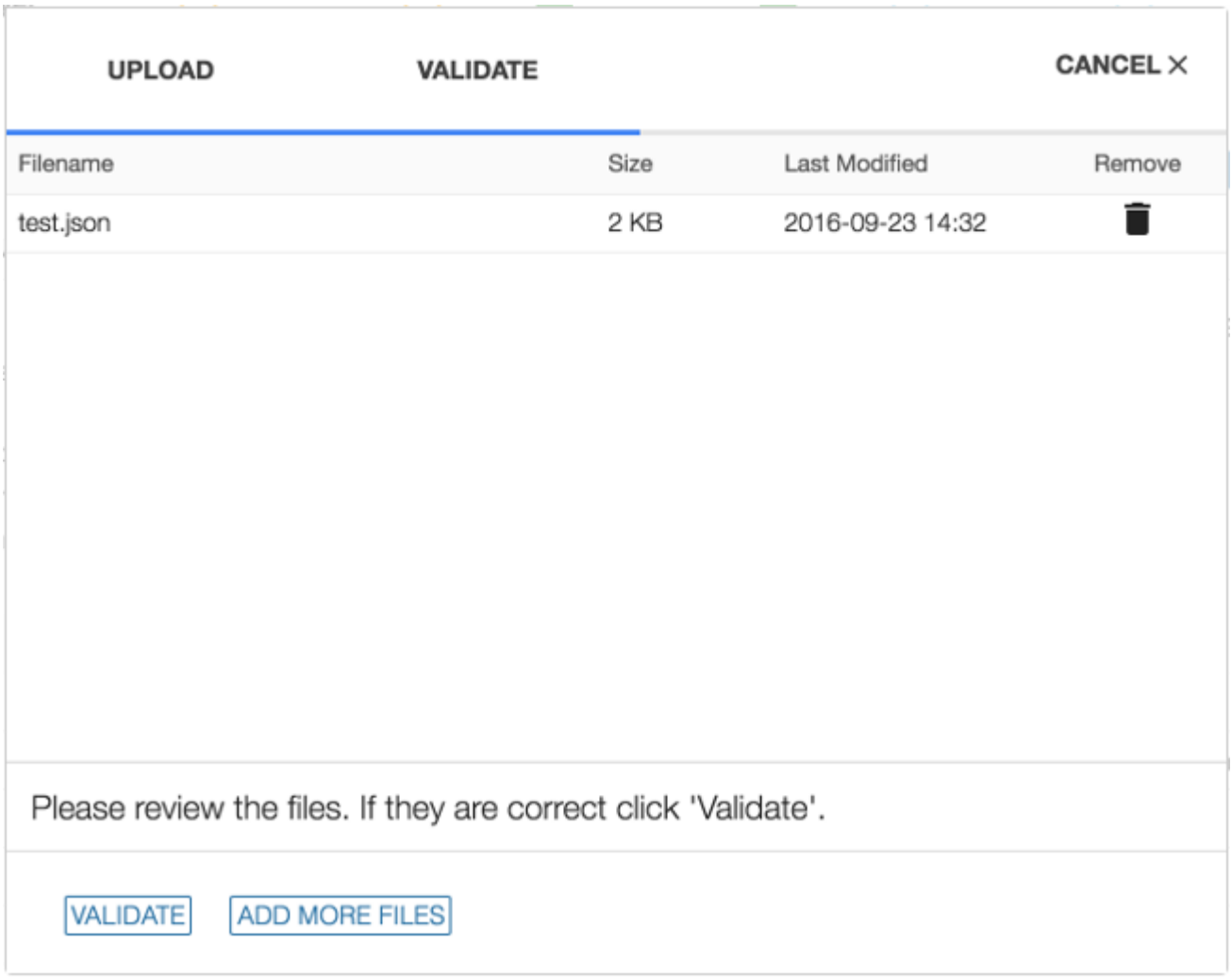


Files containing one or more entities can be added either by clicking on `CHOOSE FILE(S)` or using drag and drop. Files can be removed from the Upload Data Wizard by clicking on the garbage can icon that is displayed next to the file after the file is selected for upload.

2. Validate Entities

The **Validate Entities** stage acts as a safeguard against submitting incorrectly formatted data to the GDC Data Submission Portal. During the validation stage, the GDC API will validate the content of uploaded entities against the Data Dictionary to detect potential errors. Invalid entities will not be processed and must be corrected by the user and re-uploaded before being accepted. A validation error report provided by the system can be used to isolate and correct errors.

When the first file is added, the wizard will move to the Validate section and the user can continue to add files. When all files have been added, choosing `VALIDATE` will run a test to check if the entities are valid for submission.



3. Commit or Discard Files

If the upload contains valid entities, a new transaction will appear in the latest transactions panel with the option to `COMMIT` or `DISCARD` the data. Entities contained in these files can be committed (applied) to the project or discarded using these two buttons.

If the upload contains invalid files, a transaction will appear with a FAILED status. Invalid files will need to be either corrected and re-uploaded or removed from the submission. If more than one file is uploaded and at least one is not valid, the validation step will fail for all files.

ID	Type	Step	DateTime ↓	User	Status		
42714	Upload	Validate	2016-09-09 14:56		SUCCEEDED	COMMIT	DISCARD
42711	Upload	Validate	2016-09-09 14:56		SUCCEEDED	COMMIT	DISCARD

UPLOAD USING THE GDC API

The API has a much broader range of functionality than the Data Wizard. Entities can be created, updated, and deleted through the API. See the API Submission User Guide for a more detailed explanation and for the rest of the functionalities of the API. Generally, uploading an entity through the API can be performed using a command similar to the following:

Shell

```
curl --header "X-Auth-Token: $token" --request POST --data @CASE.json https://api.gdc.cancer.gov/v0/submission/GDC/INTERNAL/_dry_run?async=true
```

CASE.json is detailed below.

JSON

```
{
    "type": "case",
    "submitter_id": "PROJECT-INTERNAL-000055",
    "projects": {
        "code": "INTERNAL"
    }
}
```

In this example, the `_dry_run` marker is used to determine if the entities can be validated, but without committing any information. If a command passed through the `_dry_run` works, the command will work when it is changed to `commit`. For more information please go to Dry Run Transactions.

Note: Submission of TSV files is also supported by the GDC API.

Next, the file can either be committed (applied to the project) through the Data Submission Portal as before, or another API query can be performed that will commit the file to the project. The transaction number in the URL (467) is printed to the console during the first step of API submission and can also be retrieved from the Transactions tab in the Data Submission Portal.

Shell

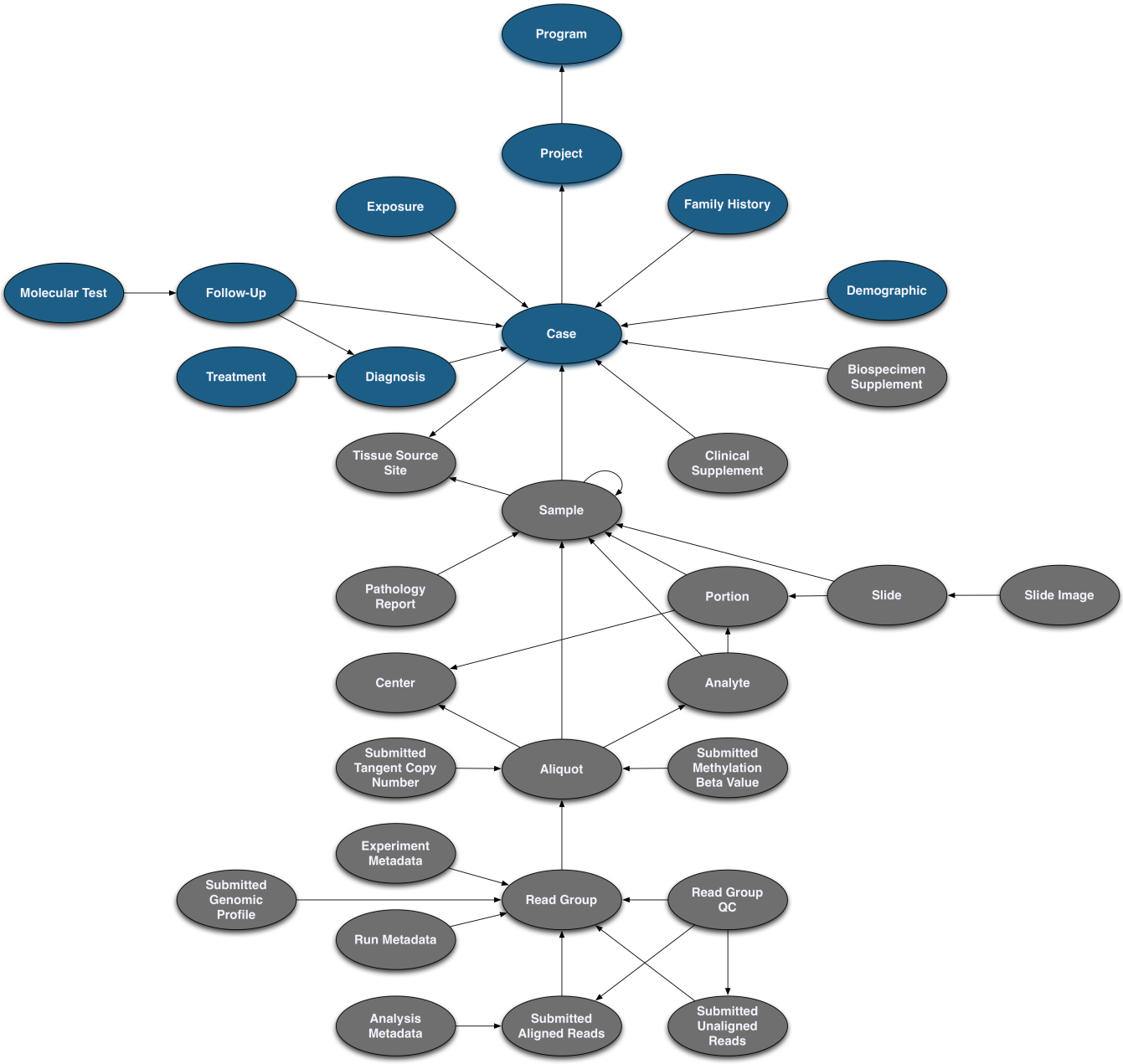
```
curl --header "X-Auth-Token: $token" --request POST https://api.gdc.cancer.gov/v0/submission/GDC/INTERNAL/transactions/467/commit?async=true
```

1.4.3 Clinical Data Submission

Typically, a submission project will include additional information about a `case` such as `demographic`, `diagnosis`, or `exposure` data.

Clinical Data Requirements

For the GDC to release a project there is a minimum number of clinical properties that are required. Minimal GDC requirements for each project includes age, gender, and diagnosis information. Other requirements may be added when the submitter is approved for submission to the GDC.



Submitting a Demographic Entity to a Case

The `demographic` entity contains information that characterizes the `case` entity.

Submitting a **Demographic** entity requires:

- **`submitter_id`:** A unique key to identify the `demographic` entity.
- **`cases.submitter_id`:** The unique key that was used for the `case` that links the `demographic` entity to the `case`.
- **`ethnicity`:** An individual's self-described social and cultural grouping, specifically whether an individual describes themselves as Hispanic or Latino. The provided values are based on the categories defined by the U.S. Office of Management and Business and used by the U.S. Census Bureau.
- **`gender`:** Text designations that identify gender. Gender is described as the assemblage of properties that distinguish people on the basis of their societal roles.
- **`race`:** An arbitrary classification of a taxonomic group that is a division of a species. It usually arises as a consequence of geographical isolation within a species and is characterized by shared heredity, physical attributes and behavior, and in the case of humans, by common history, nationality, or geographic distribution. The provided values are based on the categories defined by the U.S. Office of Management and Business and used by the U.S. Census Bureau.

JSON **TSV**

```
{
    "type": "demographic",
    "submitter_id": "PROJECT-INTERNAL-000055-DEMOGRAPHIC-1",
    "cases": {
        "submitter_id": "PROJECT-INTERNAL-000055"
    },
    "ethnicity": "not hispanic or latino",
    "gender": "male",
    "race": "asian",
}
```

```
type    cases.submitter_id  ethnicity   gender  race
demographic PROJECT-INTERNAL-000055 not hispanic or latino  male    asian
```

Submitting a Diagnosis Entity to a Case

Submitting a **Diagnosis** entity requires:

- **`submitter_id`:** A unique key to identify the `diagnosis` entity.
- **`cases.submitter_id`:** The unique key that was used for the `case` that links the `diagnosis` entity to the `case`.
- **`age_at_diagnosis`:** Age at the time of diagnosis expressed in number of days since birth.
- **`days_to_last_follow_up`:** Time interval from the date of last follow up to the date of initial pathologic diagnosis, represented as a calculated number of days.
- **`days_to_last_known_disease_status`:** Time interval from the date of last follow up to the date of initial pathologic diagnosis, represented as a calculated number of days.
- **`days_to_recurrence`:** Time interval from the date of new tumor event including progression, recurrence and new primary malignancies to the date of initial pathologic diagnosis, represented as a calculated number of days.
- **`last_known_disease_status`:** The state or condition of an individual's neoplasm at a particular point in time.
- **`morphology`:** The third edition of the International Classification of Diseases for Oncology, published in 2000 used principally in tumor and cancer registries for coding the site (topography) and the histology (morphology) of neoplasms. The study of the structure of the cells and their arrangement to constitute tissues and, finally, the association among these to form organs. In pathology, the microscopic process of identifying normal and abnormal morphologic characteristics in tissues, by employing various cytochemical and immunocytochemical stains. A system of numbered categories for representation of data.
- **`primary_diagnosis`:** Text term for the structural pattern of cancer cells used to define a microscopic diagnosis.
- **`progression_or_recurrence`:** Yes/No/Unknown indicator to identify whether a patient has had a new tumor event after initial treatment.
- **`site_of_resection_or_biopsy`:** The third edition of the International Classification of Diseases for Oncology, published in 2000, used principally in tumor and cancer registries for coding the site (topography) and the histology (morphology) of neoplasms. The description of an anatomical region or of a body part. Named locations of, or within, the body. A system of numbered categories for representation of data.
- **`tissue_or_organ_of_origin`:** Text term that describes the anatomic site of the tumor or disease.
- **`tumor_grade`:** Numeric value to express the degree of abnormality of cancer cells, a measure of differentiation and aggressiveness.
- **`tumor_stage`:** The extent of a cancer in the body. Staging is usually based on the size of the tumor, whether lymph nodes contain cancer, and whether the cancer has spread from the original site to other parts of the body. The accepted values for tumor_stage depend on the tumor site, type, and accepted staging system. These items should accompany the tumor_stage value as associated metadata.
- **`vital_status`:** The survival state of the person registered on the protocol.

JSON **TSV**


```
{
    "type": "diagnosis",
    "submitter_id": "PROJECT-INTERNAL-000055-DIAGNOSIS-1",
    "cases": {
        "submitter_id": "GDC-INTERNAL-000099"
    },
    "age_at_diagnosis": 10256,
    "days_to_last_follow_up": 34,
    "days_to_last_known_disease_status": 34,
    "days_to_recurrence": 45,
    "last_known_disease_status": "Tumor free",
    "morphology": "8260/3",
    "primary_diagnosis": "ACTH-producing tumor",
    "progression_or_recurrence": "no",
    "site_of_resection_or_biopsy": "Lung, NOS",
    "tissue_or_organ_of_origin": "Lung, NOS",
    "tumor_grade": "Not Reported",
    "tumor_stage": "stage i",
    "vital_status": "alive"
}
```

```
type    submitter_id    cases.submitter_id  age_at_diagnosis    days_to_last_follow_up  days_to_last_known_disease_status   days_to_recurrence
last_known_disease_status   morphology  primary_diagnosis   progression_or_recurrence   site_of_resection_or_biopsy tissue_or_organ_of_origin   tumor_grade
tumor_stage vital_status
diagnosis   PROJECT-INTERNAL-000055-DIAGNOSIS-1 GDC-INTERNAL-000099 10256   34  34  45  Tumor free  8260/3  ACTH-producing tumor    no  Lung, NOS   Lung,
NOS   not reported    stage i alive
```

Submitting an Exposure Entity to a Case

Submitting an **Exposure** entity does not require any information besides a link to the `case` and a `submitter_id`. The following fields are optionally included:

- **`alcohol_history`:** A response to a question that asks whether the participant has consumed at least 12 drinks of any kind of alcoholic beverage in their lifetime.
- **`alcohol_intensity`:** Category to describe the patient's current level of alcohol use as self-reported by the patient.
- **`alcohol_days_per_week`:** Numeric value used to describe the average number of days each week that a person consumes an alchoolic beverage.
- **`years_smoked`:** Numeric value (or unknown) to represent the number of years a person has been smoking.
- **`tobacco_smoking_onset_year`:** The year in which the participant began smoking.
- **`tobacco_smoking_quit_year`:** The year in which the participant quit smoking.

JSON **TSV**

```
{
    "type": "exposure",
    "submitter_id": "PROJECT-INTERNAL-000055-EXPOSURE-1",
    "cases": {
        "submitter_id": "PROJECT-INTERNAL-000055"
    },
    "alcohol_history": "yes",
    "alcohol_intensity": "Drinker",
    "alcohol_days_per_week": 2,
    "years_smoked": 5,
    "tobacco_smoking_onset_year": 2007,
    "tobacco_smoking_quit_year": 2012
}
```

```
type    submitter_id    cases.submitter_id  alcohol_history alcohol_intensity   alcohol_days_per_week years_smoked  tobacco_smoking_onset_year
tobacco_smoking_quit_year
exposure    PROJECT-INTERNAL-000055-EXPOSURE-1  PROJECT-INTERNAL-000055 yes Drinker 2 5 2007    2012
```

Note: Submitting a clinical entity uses the same conventions as submitting a `case` entity (detailed above).

1.4.4 Biospecimen Submission

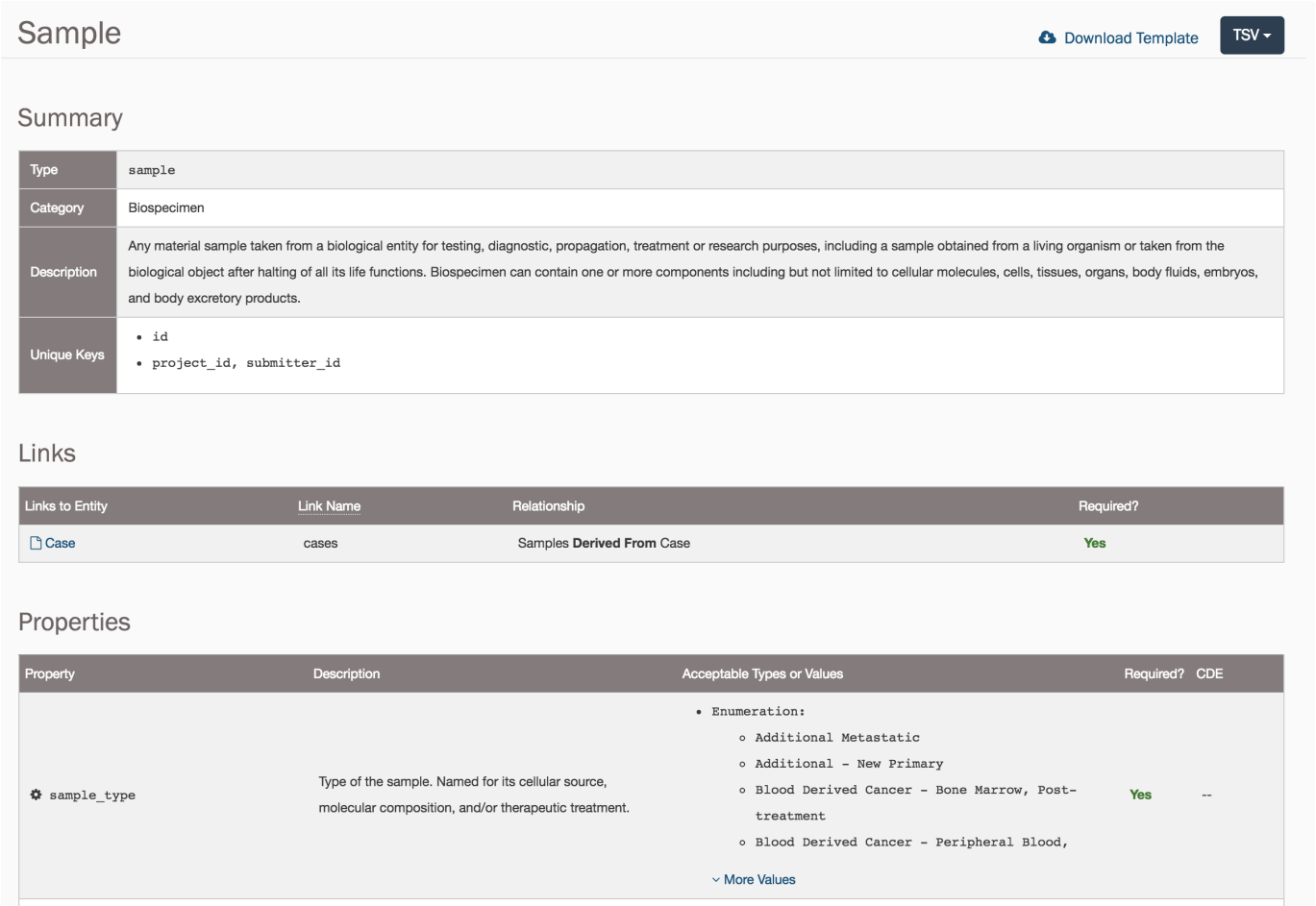
One of the main features of the GDC is the genomic data harmonization workflow. Genomic data is connected the case through biospecimen entities. The `sample` entity describes a biological piece of matter that originated from a `case`. Subsets of the `sample` such as `portions` and `analytes` can optionally be described. The `aliquot` originates from a `sample` or `analyte` and describes the nucleic acid extract that was sequenced. The `read_group` entity describes the resulting set of reads from one sequencing lane.

Sample Submission

A `sample` submission has the same general structure as a `case` submission as it will require a unique key and a link to the `case`. However, `sample` entities require four additional values:

- `tissue_type`
- `tumor_descriptor`
- `specimen_type`
- `preservation_method`

This peripheral data is required because it is necessary for the data to be interpreted. For example, an investigator using this data would need to know whether the `sample` came from tumor or normal tissue.



Sample

Download Template | TSV

Summary

Type	sample
Category	Biospecimen
Description	Any material sample taken from a biological entity for testing, diagnostic, propagation, treatment or research purposes, including a sample obtained from a living organism or taken from the biological object after halting of all its life functions. Biospecimen can contain one or more components including but not limited to cellular molecules, cells, tissues, organs, body fluids, embryos, and body excretory products.
Unique Keys	• id • project_id, submitter_id

Links

Links to Entity	Link Name	Relationship	Required?
Case	cases	Samples **Derived From** Case	Yes

Properties

Property	Description	Acceptable Types or Values	Required?	CDE
sample_type	Type of the sample. Named for its cellular source, molecular composition, and/or therapeutic treatment.	• Enumeration: ◦ Additional Metastatic ◦ Additional - New Primary ◦ Blood Derived Cancer - Bone Marrow, Post-treatment ◦ Blood Derived Cancer - Peripheral Blood, More Values	Yes	--

Submitting a **Sample** entity requires:

- **`submitter_id`:** A unique key to identify the `sample`.
- **`cases.submitter_id`:** The unique key that was used for the `case` that links the `sample` to the `case`.
- **`tissue_type`:** Text term that represents a description of the kind of tissue collected with respect to disease status or proximity to tumor tissue.
- **`tumor_descriptor`:** Text that describes the kind of disease present in the tumor specimen as related to a specific timepoint.
- **`specimen_type`:** The type of a material sample taken from a biological entity for testing, diagnostic, propagation, treatment or research purposes. This includes particular types of cellular molecules, cells, tissues, organs, body fluids, embryos, and body excretory substances.
- **`preservation_method`:** Text term that represents the method used to preserve the sample.

Note: The `case` must be "committed" to the project before a `sample` can be linked to it. This also applies to all other links between entities.

JSON **TSV**

```
{
    "type": "sample",
    "cases": {
        "submitter_id": "PROJECT-INTERNAL-000055"
    },
    "submitter_id": "Blood-00001SAMPLE_55",
    "tissue_type": "Normal",
    "tumor_descriptor": "Not Applicable",
    "specimen_type": "Peripheral Blood NOS",
    "preservation_method": "Frozen"
}
```

```
type    cases.submitter_id  submitter_id    tissue_type specimen_type   tumor_descriptor    preservation_method
sample  PROJECT-INTERNAL-000055 Blood-00001SAMPLE_55    Normal  Peripheral Blood NOS    Not Applicable  Frozen
```

Portion, Analyte and Aliquot Submission

Submitting a **Portion** entity requires:

- **`submitter_id`:** A unique key to identify the `portion`.
- **`samples.submitter_id`:** The unique key that was used for the `sample` that links the `portion` to the `sample`.

JSON **TSV**

```
{
    "type": "portion",
    "submitter_id": "Blood-portion-000055",
    "samples": {
        "submitter_id": "Blood-00001SAMPLE_55"
    }
}
```

```
type    submitter_id    samples.submitter_id
portion Blood-portion-000055    Blood-00001SAMPLE_55
```

Submitting an **Analyte** entity requires:

- **`submitter_id`:** A unique key to identify the `analyte`.
- **`portions.submitter_id`:** The unique key that was used for the `portion` that links the `analyte` to the `portion`.
- **`analyte_type`:** Text term that represents the kind of molecular specimen analyte.

JSON **TSV**

```
{
    "type": "analyte",
    "portions": {
        "submitter_id": "Blood-portion-000055"
    },
    "analyte_type": "DNA",
    "submitter_id": "Blood-analyte-000055"
}
```

```
type    portions.submitter_id   analyte_type    submitter_id
analyte Blood-portion-000055    DNA Blood-analyte-000055
```

Submitting an **Aliquot** entity requires:

- **`submitter_id`:** A unique key to identify the `aliquot`.
- **`analytes.submitter_id`:** The unique key that was used for the `analyte` that links the `aliquot` to the `analyte`.

JSON **TSV**

```
{
    "type": "aliquot",
    "submitter_id": "Blood-00021-aliquot55",
    "analytes": {
        "submitter_id": "Blood-analyte-000055"
    }
}
```

```
type    submitter_id    analytes.submitter_id
aliquot Blood-00021-aliquot55   Blood-analyte-000055
```

Note: `aliquot` entities can be directly linked to `sample` entities via the `samples.submitter_id`. The `portion` and `analyte` entities are not required for submission.

Read Group Submission

Information about sequencing reads is necessary for downstream analysis, thus the `read_group` entity requires more fields than the other Biospecimen entities (`sample` , `portion` , `analyte` , `aliquot`).

Submitting a **Read Group** entity requires:

- **`submitter_id`:** A unique key to identify the `read_group`.
- **`aliquots.submitter_id`:** The unique key that was used for the `aliquot` that links the `read_group` to the `aliquot`.
- **`experiment_name`:** Submitter-defined name for the experiment.
- **`is_paired_end`:** Are the reads paired end? (Boolean value: `true` or `false`).
- **`library_name`:** Name of the library.
- **`library_strategy`:** Library strategy.
- **`platform`:** Name of the platform used to obtain data.
- **`read_group_name`:** The name of the `read_group`.
- **`read_length`:** The length of the reads (integer).
- **`sequencing_center`:** Name of the center that provided the sequence files.
- **`library_selection`:** Library Selection Method.
- **`target_capture_kit`:** Description that can uniquely identify a target capture kit. Suggested value is a combination of vendor, kit name, and kit version.

JSON **TSV**

```
{
    "type": "read_group",
    "submitter_id": "Blood-00001-aliquot_lane1_barcodeACGTAC_55",
    "experiment_name": "Resequencing",
    "is_paired_end": true,
    "library_name": "Solexa-34688",
    "library_strategy": "WXS",
    "platform": "Illumina",
    "read_group_name": "205DD.3-2",
    "read_length": 75,
    "sequencing_center": "BI",
    "library_selection": "Hybrid Selection",
    "target_capture_kit": "Custom MSK IMPACT Panel - 468 Genes",
    "aliquots":
        {
            "submitter_id": "Blood-00021-aliquot55"
        }
}
```

```
type    submitter_id    experiment_name is_paired_end   library_name    library_selection   library_strategy    platform    read_group_name read_length
sequencing_center   target_capture_kit  aliquots.submitter_id
read_group  Blood-00001-aliquot_lane1_barcodeACGTAC_55  Resequencing    true    Solexa-34688    Hybrid Selection    WXS Illumina    205DD.3-2   75  BI
Custom MSK IMPACT Panel - 468 Genes Blood-00021-aliquot55
```

Note: Submitting a biospecimen entity uses the same conventions as submitting a `case` entity (detailed above).

1.4.5 Experiment Data Submission

Several types of experiment data can be uploaded to the GDC. The `submitted_aligned_reads` and `submitted_unaligned_reads` files are associated with the `read_group` entity, while the array-based files such as the `submitted_tangent_copy_number` are associated with the `aliquot` entity. Each of these file types are described in their respective entity submission and are uploaded separately using the GDC API or the GDC Data Transfer Tool.

Before the experiment data file can be submitted, the GDC requires that the user provides information about the file as a `submittable_data_file` entity. This includes file-specific data needed to validate the file and assess which analyses should be performed. Sequencing data files can be submitted as `submitted_aligned_reads` or `submitted_unaligned_reads`.

Submitting a **Submitted Aligned-Reads** (**Submitted Unaligned-Reads**) entity requires:

- **`submitter_id`:** A unique key to identify the `submitted_aligned_reads`.
- **`read_groups.submitter_id`:** The unique key that was used for the `read_group` that links the `submitted_aligned_reads` to the `read_group`.
- **`data_category`:** Broad categorization of the contents of the data file.
- **`data_format`:** Format of the data files.
- **`data_type`:** Specific content type of the data file. (must be "Aligned Reads").
- **`experimental_strategy`:** The sequencing strategy used to generate the data file.
- **`file_name`:** The name (or part of a name) of a file (of any type).
- **`file_size`:** The size of the data file (object) in bytes.
- **`md5sum`:** The 128-bit hash value expressed as a 32 digit hexadecimal number used as a file's digital fingerprint.

JSON **TSV**

```
{
    "type": "submitted_aligned_reads",
    "submitter_id": "Blood-00001-aliquot_lane1_barcodeACGTAC_55.bam",
    "data_category": "Raw Sequencing Data",
    "data_format": "BAM",
    "data_type": "Aligned Reads",
    "experimental_strategy": "WGS",
    "file_name": "test.bam",
    "file_size": 38,
    "md5sum": "aa6e82d11ccd8452f813a15a6d84faf1",
    "read_groups": [
        {
            "submitter_id": "Primary_Tumor_RG_86-1"
        }
    ]
}
```

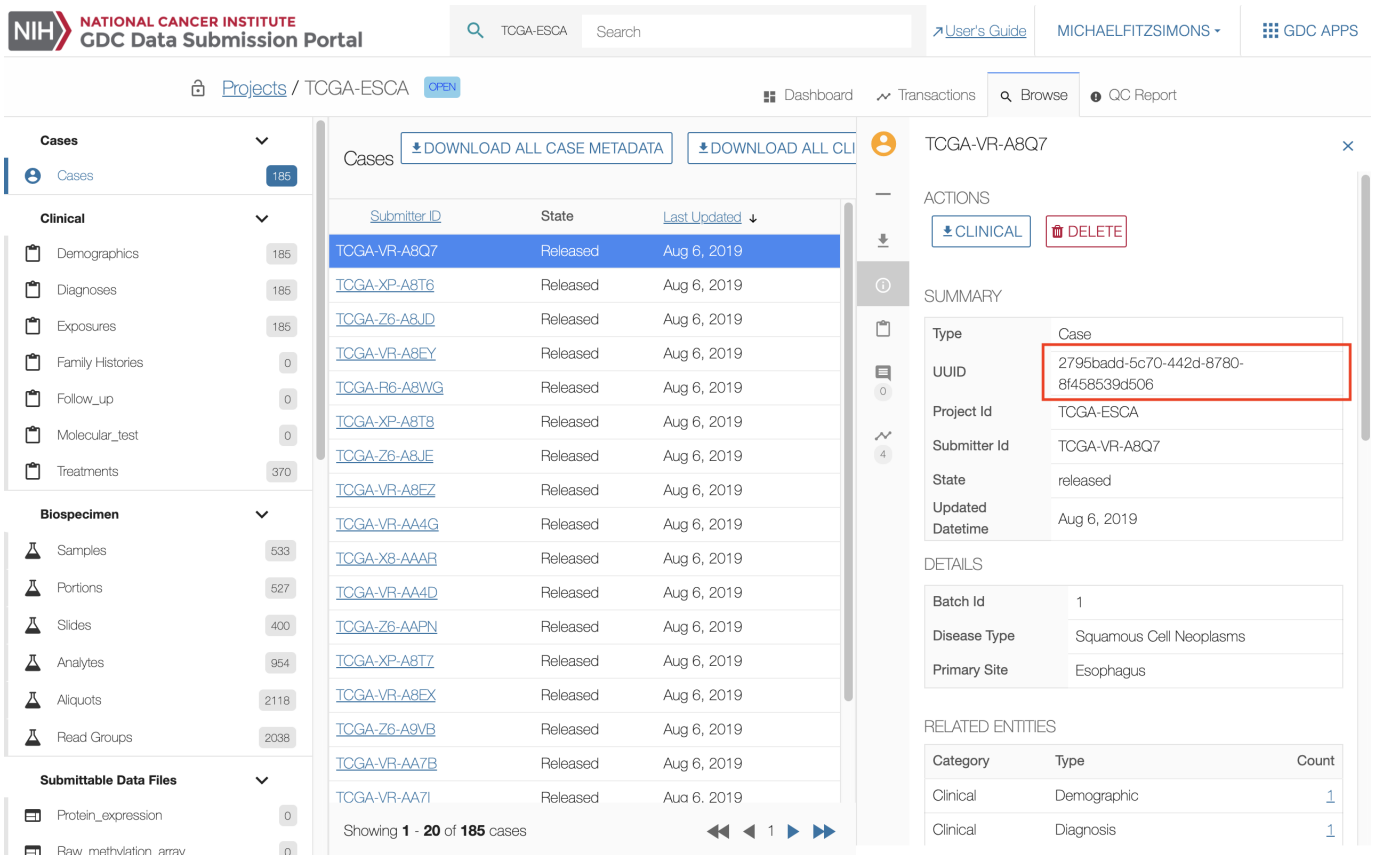
```
type    submitter_id    data_category   data_format data_type   experimental_strategy   file_name   file_size   md5sum  read_groups.submitter_id#1
submitted_aligned_reads Blood-00001-aliquot_lane1_barcodeACGTAC_55.bam  Raw Sequencing Data BAM Aligned Reads   WGS test.bam    38
aa6e82d11ccd8452f813a15a6d84faf1    Primary_Tumor_RG_86-1
```

Note: For details on submitting experiment data associated with more than one `read_group` entity, see the Tips for Complex Submissions section.

Uploading the Submittable Data File to the GDC

The submittable data file can be uploaded when it is registered with the GDC. A submittable data file is registered when its corresponding entity (e.g. `submitted_unaligned_reads`) is uploaded and committed. It is important to note that the Harmonization process does not occur on these submitted files until the user clicks the `Request Submission` button. Uploading the file can be performed with either the GDC Data Transfer Tool or the GDC API. Other types of data files such as clinical supplements, biospecimen supplements, and pathology reports are uploaded to the GDC in the same way. Supported data file formats are listed at the GDC Data Dictionary.

GDC Data Transfer Tool: A file can be uploaded using its UUID (which can be retrieved from the GDC Submission Portal or API) once it is registered.



The following command can be used to upload the file:

Shell

```
gdc-client upload --project-id PROJECT-INTERNAL --identifier a053fad1-adc9-4f2d-8632-923579128985 -t $token -f $path_to_file
```

Additionally a manifest can be downloaded from the Submission Portal and passed to the Data Transfer Tool. This will allow for the upload of more than one `submittable_data_file`:

Shell

```
gdc-client upload -m manifest.yml -t $token
```

API Upload: A `submittable_data_file` can be uploaded through the API by using the `/submission/$PROGRAM/$PROJECT/files` endpoint. The following command would be typically used to upload a file:

Shell

```
curl --request PUT --header "X-Auth-Token: $token" https://api.gdc.cancer.gov/v0/submission/PROJECT/INTERNAL/files/6d45f2a0-8161-42e3-97e6-e058ac18f3f3 -d
$path_to_file
```

For more details on how to upload a `submittable_data_file` to a project see the API Users Guide and the Data Transfer Tool Users Guide.

1.4.6 Annotation Submission

The GDC Data Portal supports the use of annotations for any submitted entity or file. An annotation entity may include comments about why particular patients or samples are not present or why they may exhibit critical differences from others. Annotations include information that cannot be submitted to the GDC through other existing nodes or properties.

If a submitter would like to create an annotation, please contact the GDC Support Team (support@nci-gdc.datacommons.io).

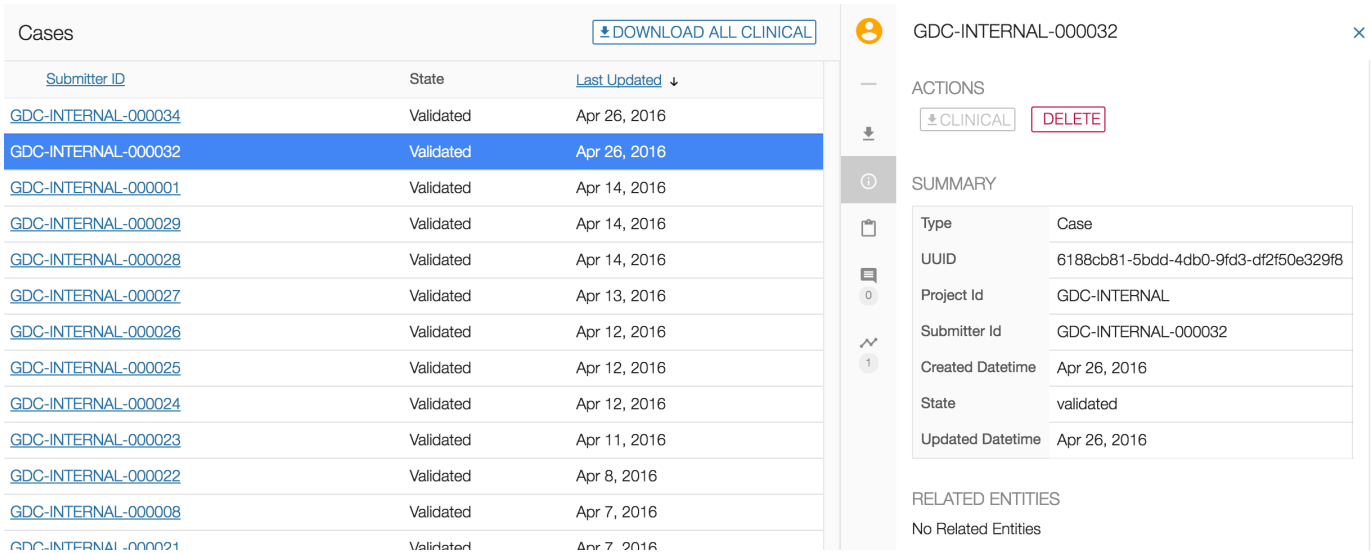
1.4.7 Deleting Submitted Entities

The GDC Data Submission Portal allows users to delete submitted entities from the project when the project is in an "OPEN" state. Files cannot be deleted while in the "SUBMITTED" state. This section applies to entities that have been committed to the project. Entities that have not been committed can be removed from the project by choosing the `DISCARD` button. Entities can also be deleted using the API. See the API Submission Documentation for specific instructions.

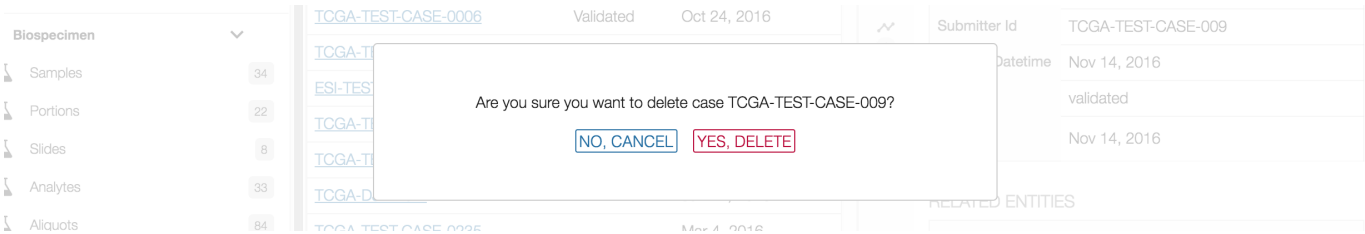
NOTE: Entities associated with files uploaded to the GDC object store cannot be deleted until the associated file has been deleted. Users must utilize the GDC Data Transfer Tool to delete these files first.

Simple Deletion

If an entity was uploaded and has no related entities, it can be deleted from the Browse tab. Once the entity to be deleted is selected, choose the `DELETE` button in the right panel under "ACTIONS".

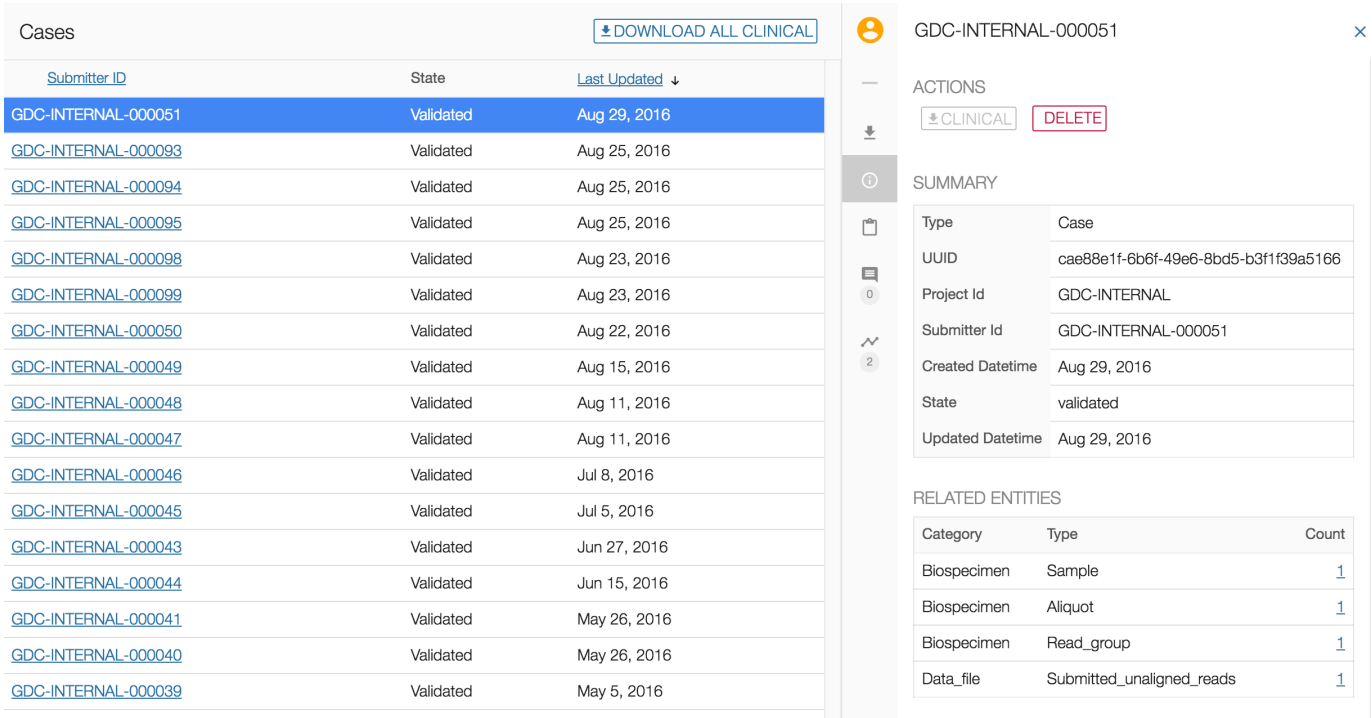


A message will then appear asking if you are sure about deleting the entity. Choosing the `YES, DELETE` button will remove the entity from the project, whereas choosing the `NO, CANCEL` button will return the user to the previous screen.



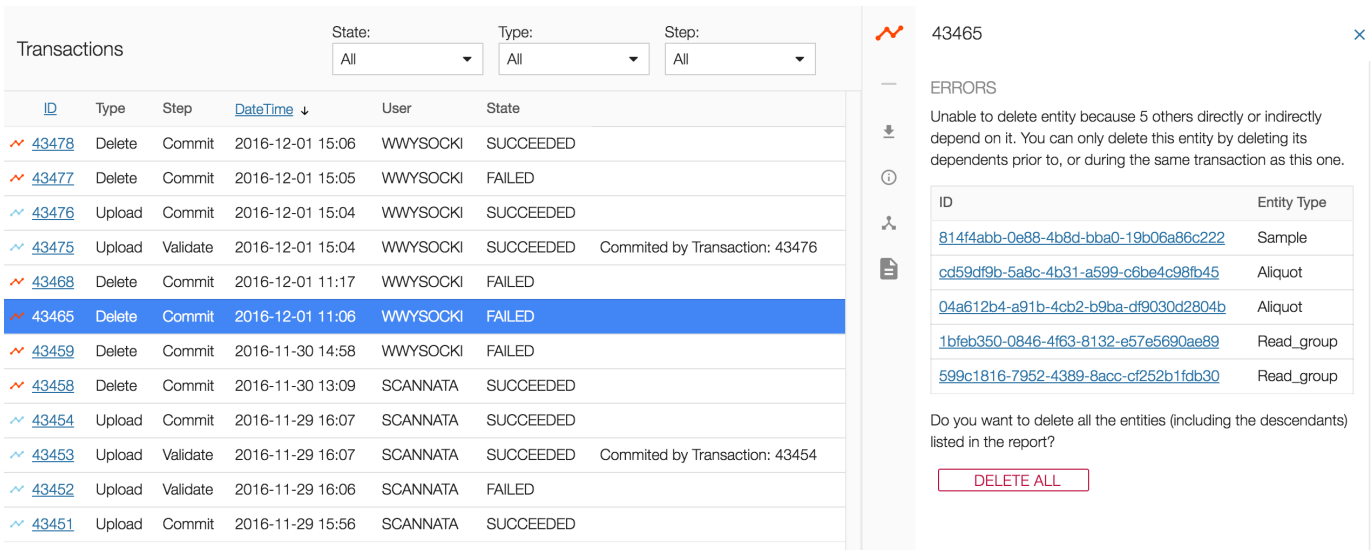
Deletion with Dependents

If an entity has related entities, such as a `case` with multiple `samples` and `aliquots`, deletion takes one extra step.



Follow the Simple Deletion method until the end. This action will appear in the Transactions tab as "Delete" with a "FAILED" state.

Choose the failed transaction and the right panel will show the list of entities related to the entity that was going to be deleted.



Selecting the `DELETE ALL` button at the bottom of the list will delete all of the related entities, their descendants, and the original entity.

Submitted Data File Deletion

The `submittable_data_files` that were uploaded erroneously are deleted separately from their associated entity using the GDC Data Transfer Tool. See the section on Deleting Data Files in the Data Transfer Tool users guide for specific instructions.

1.4.8 Updating Uploaded Entities

Before harmonization occurs, entities can be modified to update, add, or delete information. These methods are outlined below.

Updating or Adding Fields

Updated or additional fields can be applied to entities by re-uploading them through the GDC Data Submission portal or API. See below for an example of a case upload with a `primary_site` field being added and a `disease_type` field being updated.

Before **After**

```
{
"type":"case",
"submitter_id":"GDC-INTERNAL-000043",
"projects":{
  "code":"INTERNAL"
},
"disease_type": "Myomatous Neoplasms"
}
```

```
{
"type":"case",
"submitter_id":"GDC-INTERNAL-000043",
"projects":{
  "code":"INTERNAL"
},
"disease_type": "Myxomatous Neoplasms",
"primary_site": "Pancreas"
}
```

Guidelines:

- The newly uploaded entity must contain the `submitter_id` of the existing entity so that the system updates the correct one.
- All newly updated entities will be validated by the GDC Dictionary. All required fields must be present in the newly updated entity.
- Fields that are not required do not need to be re-uploaded and will remain unchanged in the entity unless they are updated.

Deleting Optional Fields

It may be necessary to delete fields from uploaded entities. This can be performed through the API and can only be applied to optional fields. It also requires the UUID of the entity, which can be retrieved from the submission portal or using a GraphQL query.

In the example below, the `primary_site` and `disease_type` fields are removed from a `case` entity:

Shell **Before** **After**

```
curl --header "X-Auth-Token: $token_string" --request DELETE  --header "Content-Type: application/json" "https://api.gdc.cancer.gov/v0/submission/EXAMPLE/
PROJECT/entities/7aab7578-34ff-5651-89bb-57aefdc4c4f8?fields=primary_site,disease_type"
```

```
{
"type":"case",
"submitter_id":"GDC-INTERNAL-000043",
"projects":{
  "code":"INTERNAL"
},
"disease_type": "Germ Cell Neoplasms",
"primary_site": "Pancreas"
}
```

```
{
"type":"case",
"submitter_id":"GDC-INTERNAL-000043",
"projects":{
  "code":"INTERNAL"
}
}
```

Versioning

Changes to entities will create versions. For more information on this, please go to Uploading New Versions of Data Files.

1.4.9 Strategies for Submitting in Bulk

Each submission in the previous sections was broken down by component to demonstrate the GDC Data Model structure. However, the submission of multiple entities at once is supported and encouraged. Here two strategies for submitting data in an efficient manner are discussed.

Registering a BAM File: One Step

Registering a BAM file (or any other type) can be performed in one step by including all of the entities, from `case` to `submitted_aligned_reads`, in one file. See the example below:

JSON

```
[{
    "type": "case",
    "submitter_id": "PROJECT-INTERNAL-000055",
    "projects": {
        "code": "INTERNAL"
    }
},
{
    "type": "sample",
    "cases": {
        "submitter_id": "PROJECT-INTERNAL-000055"
    },
    "submitter_id": "Blood-00001SAMPLE_55",
    "tissue_type": "Normal",
    "tumor_descriptor": "Not Applicable",
    "specimen_type": "Peripheral Blood NOS",
    "preservation_method": "Frozen"
},
{
    "type": "portion",
    "submitter_id": "Blood-portion-000055",
    "samples": {
        "submitter_id": "Blood-00001SAMPLE_55"
    }
},
{
    "type": "analyte",
    "portions": {
        "submitter_id": "Blood-portion-000055"
    },
    "analyte_type": "DNA",
    "submitter_id": "Blood-analyte-000055"
},
{
    "type": "aliquot",
    "submitter_id": "Blood-00021-aliquot55",
    "analytes": {
        "submitter_id": "Blood-analyte-000055"
    }
},
{
    "type": "read_group",
    "submitter_id": "Blood-00001-aliquot_lane1_barcodeACGTAC_55",
    "experiment_name": "Resequencing",
    "is_paired_end": true,
    "library_name": "Solexa-34688",
    "library_selection":"Hybrid Selection",
    "library_strategy": "WXS",
    "platform": "Illumina",
    "read_group_name": "205DD.3-2",
    "read_length": 75,
    "sequencing_center": "BI",
    "aliquots":
        {
            "submitter_id": "Blood-00021-aliquot55"
        }
},
{
    "type": "submitted_aligned_reads",
    "submitter_id": "Blood-00001-aliquot_lane1_barcodeACGTAC_55.bam",
    "data_category": "Raw Sequencing Data",
    "data_format": "BAM",
    "data_type": "Aligned Reads",
    "experimental_strategy": "WGS",
    "file_name": "test.bam",
    "file_size": 38,
    "md5sum": "aa6e82d11ccd8452f813a15a6d84faf1",
    "read_groups": [
        {
            "submitter_id": "Blood-00001-aliquot_lane1_barcodeACGTAC_55"
        }
    ]
}]
```

All of the entities are placed into a JSON list object:

```
[{"type": "case","submitter_id": "PROJECT-INTERNAL-000055","projects": {"code": "INTERNAL"}}}, entity-2, entity-3]
```

The entities need not be in any particular order as they are validated together.

Note: Tab-delimited format is not recommended for 'one-step' submissions due to an inability of the format to accommodate multiple 'types' in one row.

Submitting Numerous Cases

The GDC understands that submitters will have projects that comprise more entities than would be reasonable to individually parse into JSON formatted files. Additionally, many investigators store large amounts of data in a tab-delimited format (TSV). For instances like this, we recommend parsing all entities of the same type into separate TSVs and submitting them on a type-basis.

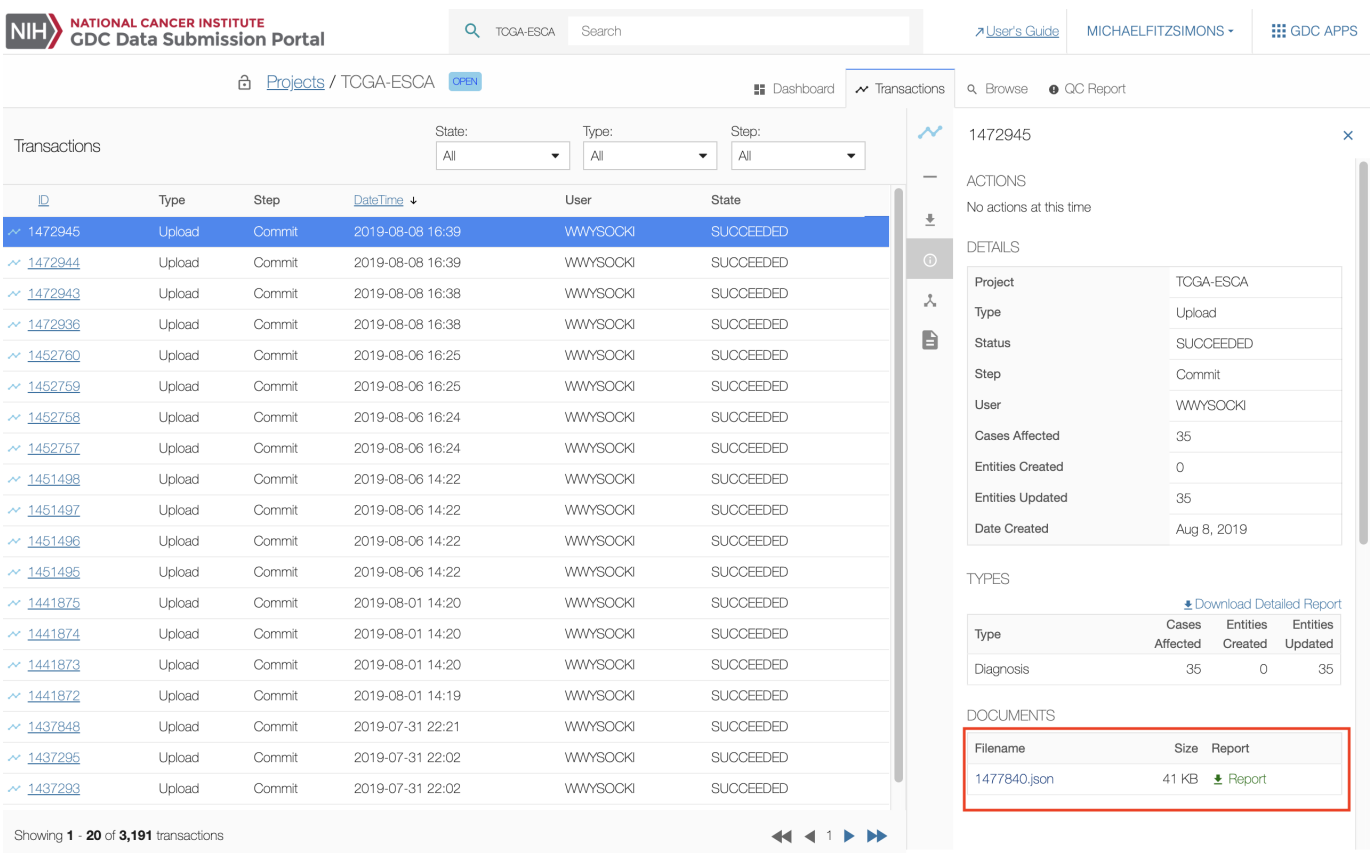
For example, a user may want to submit 100 Cases associated with 100 `samples`, 100 `portions`, 100 `analytes`, 100 `aliquots`, and 100 `read_groups`. Constructing and submitting 100 JSON files would be tedious and difficult to organize. The solution is submitting one `case` TSV containing the 100 `cases`, one `sample` TSV containing the 100 `samples`, so on and so forth. Doing this would only require six TSVs and these files can be formatted in programs such as Microsoft Excel or Google Spreadsheets.

See the following example TSV files:

- Cases.tsv
- Samples.tsv
- Portions.tsv
- Analytes.tsv
- Aliquots.tsv
- Read-Groups.tsv

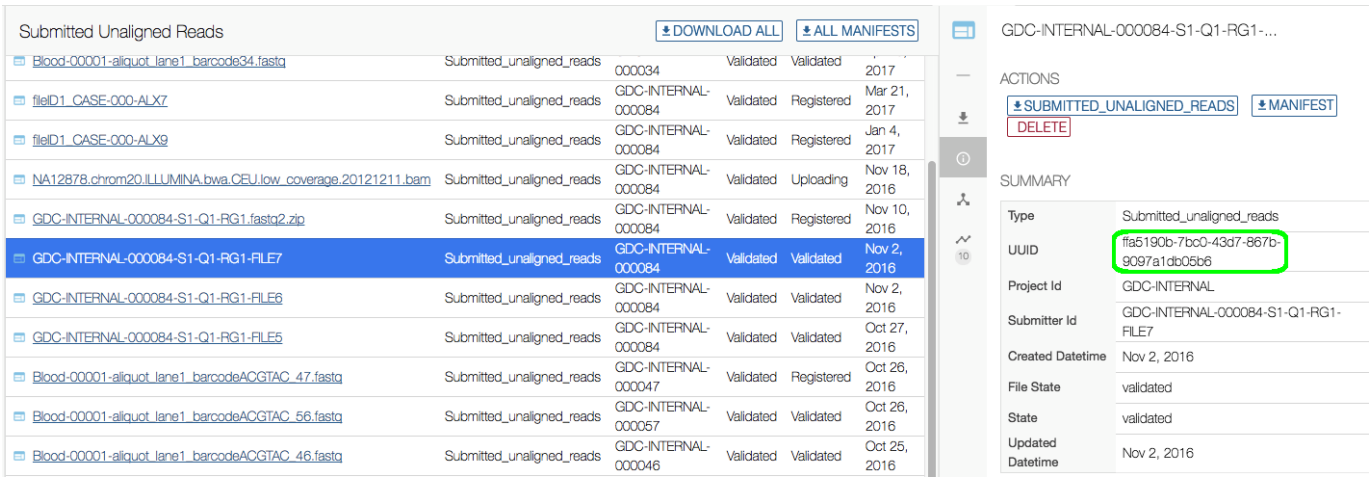
Download Previously Uploaded Metadata Files

The transaction page lists all previous transactions in the project. The user can download metadata files uploaded to the GDC workspace in the details section of the screen by selecting one transaction and scrolling to the "DOCUMENTS" section.



Download Previously Uploaded Data Files

The only supported method to download data files previously uploaded to the GDC Submission Portal that have not been release yet is to use the API or the Data Transfer Tool. To retrieve data previous upload to the submission portal you will need to retrieve the data file's UUID. The UUIDs for submitted data files are located in the submission portal under the file's Summary section as well as the manifest file located on the file's Summary page.



Once the UUID(s) have been retrieved, the download process is the same as it is for downloading data files at the GDC Portal using UUIDs.

Note: When submittable data files are uploaded through the Data Transfer Tool they are not displayed as transactions.

1.5 Data Submission Portal Release Notes

Version	Date
v3.2.0	June 30, 2025
v2.6.0	July 8, 2022
v2.5.1	August 14, 2020
v2.5.0	July 2, 2020
v2.4.1	March 9, 2020
v2.4.0	November 6, 2019
v2.3.0	June 5, 2019
v2.2.0	February 20, 2019
v2.1.0	November 7, 2018
v2.0.0	August 23, 2018
v1.9.0	May 21, 2018
v1.8.0	February 15, 2018
v1.7.0	November 16, 2017
v1.6.0	August 22, 2017
v1.5.1	March 16, 2017
v1.3.0	October 31, 2016
v1.2.2	September 23, 2016
v1.1.0	May 20th, 2016
v0.3.24.1	February 26, 2016
v0.3.21	January 27, 2016
v0.2.18.3	November 30, 2015

1.5.1 Regular Maintenance Records 3.2.0

- **GDC Product**: GDC Data Submission Portal
- **Date**: June 30, 2025

New Features and Changes

- Removed legacy archive from "GDC Apps" cube
- Added UTM codes to the "GDC Apps" cube
- Expanded banner functionality
- Operating system updates and maintenance

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- If a project ID has a character that is not alphanumeric, a dash, or an underscore, submission portal users may experience errors.
- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.2 Release 2.6.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: July 8, 2022

New Features and Changes

- Text on the QC tab was changed to clarify which errors require attention from the submitter.

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- If a project ID has a character that is not alphanumeric, a dash, or an underscore, submission portal users may experience errors.
- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.3 Release 2.5.1

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: August 14, 2020

New Features and Changes

- Enhancements were made to the submission API to increase performance and usability.

Bugs Fixed Since Last Release

- None

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.4 Release 2.5.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: July 2, 2020

New Features and Changes

- None.

Bugs Fixed Since Last Release

- Fixed bug where the Details pane in the QC Report was displaying crowded, non-uniform buttons.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.5 Release 2.4.1

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: March 9, 2020

New Features and Changes

- Removed duplicate queries from various pages in the Submission Portal to optimize data retrieval and rendering.

Bugs Fixed Since Last Release

- Fixed bug where the right-hand detail pane in the Transactions and QC Report tabs was being cut off and not scrollable in the viewport for Windows environments (all browsers).
- Fixed bug in the PDF file downloaded from the QC Report tab's Project Summary, where text was being cut off when browsing in Firefox or Microsoft Edge.
- Fixed bug where the TSV and JSON download buttons completely disappear and cannot be scrolled to in the Project Data Download modal, if it is shrunk beyond a certain threshold.
- Fixed bug in the Manifest download button that was trying to capture certain incorrect or unnecessary file states.
- Fixed incorrect DTT hyperlink in the GDC Apps menu.
- Fixed bug where the banner warning users that ERA Commons login was currently not working, would only appear after the user logged in, thus defeating the purpose of the warning in the first place.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.6 Release 2.4.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: November 6, 2019

New Features and Changes

- Added new QC Report tab that allows users to view and download QC errors detected on the current set of unsubmitted data. Users must examine these errors and fix them appropriately before re-uploading the data and requesting harmonization. New donut added to the Dashboard tab to display a quick breakdown of CRITICAL vs WARNING errors across the project.

Bugs Fixed Since Last Release

- Fixed the Project Data download button in the Project Overview, so that the JSON option is selectable in the modal.
- Fixed the trash can icon for the Delete button in the Submitter Detail pane, so that it is fully visible and no longer cutoff
- Fixed a Section 508 Accessibility violation in the Submitter Detail pane.
- Increased the global transaction polling interval to 15 seconds across the portal to improve performance.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.7 Release 2.3.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: June 5, 2019

New Features and Changes

- Added Aligned Reads, Gene Expressions, miRNA Expression to the Harmonized Data Files list in the Browse tab.

Bugs Fixed Since Last Release

- Fixed logic on the Submittable Data Files donut to more accurately display number of submittable files that have been validated. Specifically, both the file state and corresponding note state must be validated. Also updated the corresponding tooltip text.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.8 Release 2.2.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: February 20, 2019

New Features and Changes

- Renamed the "Request Submission" button to "Request Harmonization" to make the purpose of this action more clear.

Bugs Fixed Since Last Release

- Fixed the right scroll bar in the records list on the Browse page so that it works in Firefox.
- Fixed a dead link to the Submission Portal User Guide on the Dashboard.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.9 Release 2.1.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: November 7, 2018

New Features and Changes

- Updated the project columns to include a Release column in addition to the Batch Submit column.

Bugs Fixed Since Last Release

- Fixed quick search so that projects with a dash in the name will no longer break the search.
- PO reports will now return the latest data for each project that has completed running.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.10 Release 2.0.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: August 23, 2018

New Features and Changes

- The submission process has been updated to request submission and release.
- Users can "Request Submission" once data has been reviewed. Previously the button that said "Submit" now says "Request Submission".
- Users can "Request Release" once data has processed by the GDC. Previously the button that said "Release" now says "Request Release".

Bugs Fixed Since Last Release

- Fixed Download All Manifest button being much larger than other buttons.
- Fixed bug where all tables said Showing x of xx projects instead of the correct entity.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.11 Release 1.9.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: May 21, 2018

New Features and Changes

- Added the ability to download all case metadata in the Browse Cases view

Bugs Fixed Since Last Release

- Time outs when loading submission portal project list
- Missing PO reports for CPTAC-3 project
- Windows - Scroll bar interfering with browser scroll bar
- Donut "Cases with submittable data files" always shows 0
- For Biospecimen entities, the "Download all" button does not take filtering into account
- For Clinical entities, the "Download all" button was downloading all clinical entities instead of selected clinical type

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.12 Release 1.8.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: February 15, 2018

New Features and Changes

- Added the ability to support banners with hyperlinks

Bugs Fixed Since Last Release

- Fixed 508 compliance issues

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.13 Release 1.7.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: November 16, 2017

New Features and Changes

- None

Bugs Fixed Since Last Release

- Fixed bug where error would be produced even while project was successfully submitted

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.14 Release 1.6.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: August 22, 2017

New Features and Changes

- Added ability to see metadata for particular harmonized data files in the Submission Portal

Bugs Fixed Since Last Release

None

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.15 Release 1.5.1

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: March 16, 2017

New Features and Changes

- Added ability to delete an entity. Read more about this here
- Added Project Reports in the projects list page. Read more about this here.
- To avoid confusion, renamed "Status" to "State" in the Browse section
- Added tooltip over Hierarchy title when reviewing an entity
- Restrict the upload window to only supported data formats (JSON and TSV)

Bugs Fixed Since Last Release

- File status column should not be displayed for any clinical or biospecimen entities but only for submittable data files.
- Diagnosis / Treatment detail: Submitter ID (of the child / parent) is missing in the Details -> Hierarchy view.
- In some situations tooltip entries remain on-screen. Workaround is to refresh the page.
- In Browse tab, "Submittable Data Files" filter, clicking on "Download All" currently returns case and clinical informations instead of returning file informations. Workaround is to download information from the file the details panel.
- In Dashboard, the donut chart for number of cases with submittable data files is always empty. A workaround is to visit the Browse, detailed case view section to see, case by case, if it has submittable data files.
- In Transactions tab, after clicking on Commit or Discard, status is not automatically refreshed.
- Added the API version in the Data Submission Portal footer on the project list page.
- Inconsistent behavior when clicking on a Transaction ID on the Dashboard.
- Empty transactions created when submitting files in an incorrect format.
- JSON file downloaded from the Data Submission Portal cannot be used to resubmit data.
- "Submitted data files" donut chart and "Download Manifest" button do not get refreshed after committing a transaction.
- Release information on the Dashboard creates confusion.
- Missing Boolean fields from details panel.
- No message displayed if no results are found via the top menu search.
- "Invalid Date" on IE11 and Firefox ESR 45.x
- Download option truncated in the details panel.
- Download All from the browse section returns too many records.

Known Issues and Workarounds

- When creating entities in the Submission Portal, occasionally an extra transaction will appear with status error. This does not seem to impact that actual transaction, which is recorded as occurring successfully.

1.5.16 Release 1.3.0

- **GDC Product**: GDC Data Submission Portal

- **Release Date**: October 31, 2016

New Features and Changes

Not Applicable

Bugs Fixed Since Last Release

- Adding a call to the backend to ensure a refreshed token is being downloaded when user clicks on "Download Token".
- Fixed an issue with some buttons not working in Firefox
- Disabled Download Clinical button if the file has no clinical data
- Disabled Download Manifest button while the page is loading
- Fixed an issue with some dropdowns being cut off

Known Issues and Workarounds

- Project submission and release is currently disabled.
- Diagnosis / Treatment detail: Submitter ID (of the child / parent) is missing in the Details -> Hierarchy view.
- File status column should not be displayed for any clinical or biospecimen entities but only for submittable data files.
- In some situations tooltip entries remain on-screen. Workaround is to refresh the page.
- In Browse tab, "Submittable Data Files" filter, clicking on "Download All" currently returns case and clinical informations instead of returning file informations. Workaround is to download information from the file the details panel.
- In Dashboard, the donut chart for number of cases with submittable data files is always empty. A workaround is to visit the Browse, detailed case view section to see, case by case, if it has submittable data files.
- In Transactions tab, after clicking on Commit or Discard, status is not automatically refreshed. Workaround is to refresh the page after clicking on Commit or Discard. This does not affect the transaction section of the project dashboard.
- Reports are currently not available in the Data Submission Portal and will be added back in an upcoming version:
- Data Validation Report: The rows in the report are sometimes duplicated and #Files in error are not showing up in the report. The user should go to Project > Browse > Submitted Files to see the files in error and the error type.
- The Scientific Pre-alignment QC Report is not available.

1.5.17 Release 1.2.2

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: September 23, 2016

New Features and Changes

This version contains major improvements to the GDC Data Submission Portal in both usability, performance and reliability.

Some known issues and workarounds listed in previous release notes have been made redundant due to this refactoring effort, thus are not listed anymore.

Please refer to the GDC Data Submission Portal User's Guide for more details about the features.

- Submission-related actions have been made Asynchronous.
- Fully revamped the dashboard layout and features to clarify the submission process and give easier access to key features.
- Created a transactions list page with options to take actions on transactions (in particular committing an upload)
- Improved performance of the Browse tab.
- Added GDC Apps to the header section.

Bugs Fixed Since Last Release

- Data submitted to the project can be downloaded from each project page by clicking on "PROJECT DATA" from the project page.
- When uploading multiple files at once, validation will fail if a child entity is listed before its parent.
- In Browse > Case > Details, Experimental Data (renamed to Submittable Data Files) are not listed in "Related Entities" section.
- In the upload report, the number of affected cases is incorrect (show 0) when entities are created.

Known Issues and Workarounds

- Project submission and release is currently disabled.
- If case has no clinical data, the "Download Clinical" button is not disabled. The downloaded TSV will not contain Clinical Data.
- Download Manifest button is available while page is loading or when no files are in "registered" state. Clicking on Download will return a file with an error message.
- In Internet Explorer, GDC APPs and File dropdown are incorrectly aligned, making some elements only partially visible.
- Diagnosis / Treatment detail: Submitter ID (of the child / parent) is missing in the Details -> Hierarchy view.
- File status column should not be displayed for any clinical or biospecimen entities but only for submittable data files.
- In some situations tooltip entries remain on-screen. Workaround is to refresh the page.
- In Browse tab, "Submittable Data Files" filter, clicking on "Download All" currently returns case and clinical informations instead of returning file informations. Workaround is to download information from the file the details panel.
- In Dashboard, the donut chart for number of cases with submittable data files is always empty. A workaround is to visit the Browse, detailed case view section to see, case by case, if it has submittable data files.
- In Transactions tab, after clicking on Commit or Discard, status is not automatically refreshed. Workaround is to refresh the page after clicking on Commit or Discard. This does not affect the transaction section of the project dashboard.
- Reports are currently not available in the Data Submission Portal and will be added back in an upcoming version:
- Data Validation Report: The rows in the report are sometimes duplicated and #Files in error are not showing up in the report. The user should go to Project > Browse > Submitted Files to see the files in error and the error type.
- The Scientific Pre-alignment QC Report is not available.

1.5.18 Release 1.1.0

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: May 20th, 2016

New Features and Changes

- Updated login text

Bugs Fixed Since Last Release

- Improved 508 compliance of the landing page

Known Issues and Workarounds

- Some actions on the data submission portal are resource intensive and might result in timeout or errors. The team is currently working on a major update to the data submission portal expected to be available towards the end of the summer. This version will address performance issues and improve overall user experience.

1.5.19 Release 0.3.24.1

- **GDC Product**: GDC Data Submission Portal

- **Release Date**: February 26, 2016

New Features and Changes

- Fully revamped dashboard, improved widgets with more details.
- Improved the release process and provided guidance.
- Removed dictionary viewer (moved to Documentation site).
- Added reports.

Bugs Fixed Since Last Release

- Issues with transactions list page have been addressed.
- Search feature has been improved.
- Release processes (submit and release) have been finalized and implemented.

Known Issues and Workarounds

- In Browse > Submitted Files, "DOWNLOAD ALL" downloads cases and clinical data instead of submitted files. A workaround is to initiate the download from the Dashboard.
- In Browse > Case > Details, experimental data are not listed in "Related Entities" section.
- When uploading multiple files at once, validation will fail if a child entity is listed before its parent. A workaround is to ensure parent entities are listed before their children in the Upload wizard.
- Submission status column is inconsistent between Submitted Files and Read Group. Submission Status says "Validated" when it should say "Submitted" when users submit data to the GDC.
- In the upload report, the number of affected cases is incorrect (show 0) when entities are created.
- In Dashboard > Release, Download of submitted data returns data from the project workspace but not from snapshot (submitted files).
- In Browse > Read Groups, the data completeness property is incorrect when multiple files are in the bundle.
- In Browse > Diagnosis/Treatment > Details, the hierarchy section is missing elements.
- Data Validation Report: The rows in the report are sometimes duplicated and #Files in error are not showing up in the report. The user should go to Project > Browse > Submitted Files to see the files in error and the error type.
- The Scientific Pre-alignment QC Report is not available.
- The Submission Portal is not meant to support XML file submission, users have to submit files through the GDC API.

1.5.20 Release 0.3.21

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: January 27, 2016

New Features and Changes

- New landing page.
- Updated dashboard with new widgets, updated instructions and added a transactions list.
- Improved user experience in the browse section of the portal (new layout, revamped detailed section, updated icons, etc.).

Bugs Fixed Since Last Release

- Some sections of the GDC Data Submission Portal are not compatible with Internet Explorer 10 and 11 (Note: support has been dropped for IE 10).
- Case count coverage report is currently not available. Case overview report is available from the landing page.
- If a page does not return data, the Submission Portal does not indicate that there is no data.

- The projects dropdown show legacy projects. It should only show projects active for submission.

Known Issues and Workarounds

- Users may experience issues (e.g., unable to access a specific transaction ID) when navigating the transactions list page. The workaround is to refresh the page.
- Search feature partially implemented, some items are not searchable.
- Implementation of Submit and Release not finalized.
- XML file submission is returning a bad request error. The workaround is to submit files through the GDC API.
- Data Bundles - Lane Level Sequence: read group ID not unique, the message generated is not user friendly.

1.5.21 Release 0.2.18.3

- **GDC Product**: GDC Data Submission Portal
- **Release Date**: November 30, 2015

New Features and Changes

- Dashboard to provide high-level details about one or more projects.
- Submission wizard to guide user through a three stage submission:
- Upload file to the GDC.
- Validate uploaded files.
- Submit validated files to the GDC.
- Support submission of JSON, TSV and XML files.
- Tables and reports to identify key elements of the submission:
- List of Cases, Samples, Portions, Analytes, Aliquots, Lane Level Sequence.
- Cases missing clinical data.
- Cases missing samples.
- Aliquots missing data bundles.
- Submitted data bundle and associated QC metrics.
- Recent and all transactions.
- Case count coverage.
- Manifest to facilitate submission of molecular data via the GDC Data Transfer Tool.
- Per-project dictionary viewer.
- Download authentication token.

Bugs Fixed Since Last Release

- Initial Release - Not Applicable

Known Issues and Workarounds

- Some sections of the GDC Data Submission Portal are not compatible with Internet Explorer 10 and 11.
- Case count coverage report is currently not available.
- Unable to use the search feature (implementation of this feature is not complete).
- If the page does not return data, the Submission Portal does not mention that there is no data.
- Case is not releasable when submitting a partial lane level seq data bundle. Note: This feature is currently being revised.
- XML file submission is returning bad request error. A workaround is to submit files through the API.

- Data Bundles - Lane Level Sequence: read group ID not unique, message generated is not user friendly.
- The projects dropdown show legacy projects, it should only show projects active for submission.