CDS User Guide

Contents

The purpose of this document is to help researchers understand the process of submitting NCI-funded research data to the Cancer Data Service (CDS), as well as accessing and analyzing research data shared by CDS through the CDS Portal.

Overview

CDS is one of several data commons within the Cancer Research Data Commons (CRDC) infrastructure for storing and sharing cancer research data generated by NCI-funded programs. CDS currently hosts a variety of data types from NCI projects and programs such as the Human Tumor Atlas Network (HTAN), Division of Cancer Control and Population Sciences (DCCPS), and Childhood Cancer Data Initiative (CCDI), as well as data from independent research projects.

CDS provides data-sharing capabilities for additional studies that fall under the following categories:

- Studies with data that do not match any existing CRDC data commons
- Studies with data that do not fit current data type criteria for any CRDC data commons

Searching the data housed in CDS is easy to do through the portal's Data Explore page, which has a filter function to find data by categories. Users can build cohorts from metadata on the Data Explore page and export those cohorts to Seven Bridges-CGC (SB-CGC), from Velsera to access the files for further analysis in this NCI-funded secure cloud environment, without the expense of downloading data to a local compute environment.

The metadata on CDS is searchable by all and can build cohorts to export to Velsera. Access to the controlled data files. Users requires approval from the NIH Data Access Committee through the dbGaP process.

Open-access data is publicly accessible and does not require any approvals. Being a cloud repository, CDS does not facilitate direct downloads of data, owing to high data download charges.

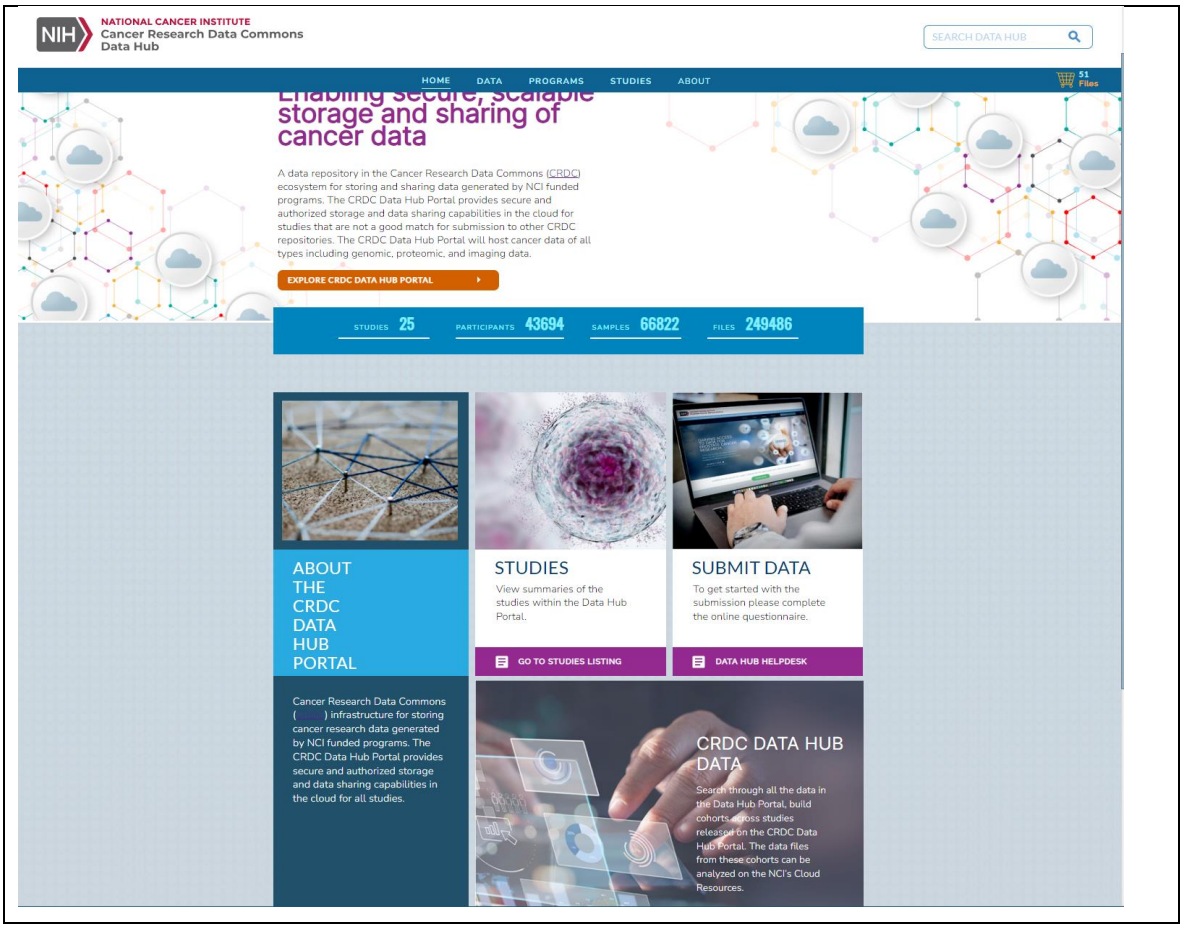
Access to data in CDS for review and/or analysis is through the NCI's Cloud Resource, Velsera Cancer Genomics Cloud (Velsera/CGC) (formerly Seven Bridges). See below for a description of how to select data and import to Velsera/CGC. More instructions on how to access and upload files can be found in the section below or on the cart page.

Explore Data on the CDS Portal

Start on the CDS Portal home page to learn how to navigate the services CDS provides.

CDS Portal Home Page

The following is the CDS Portal home page. A description of its features follows the image.

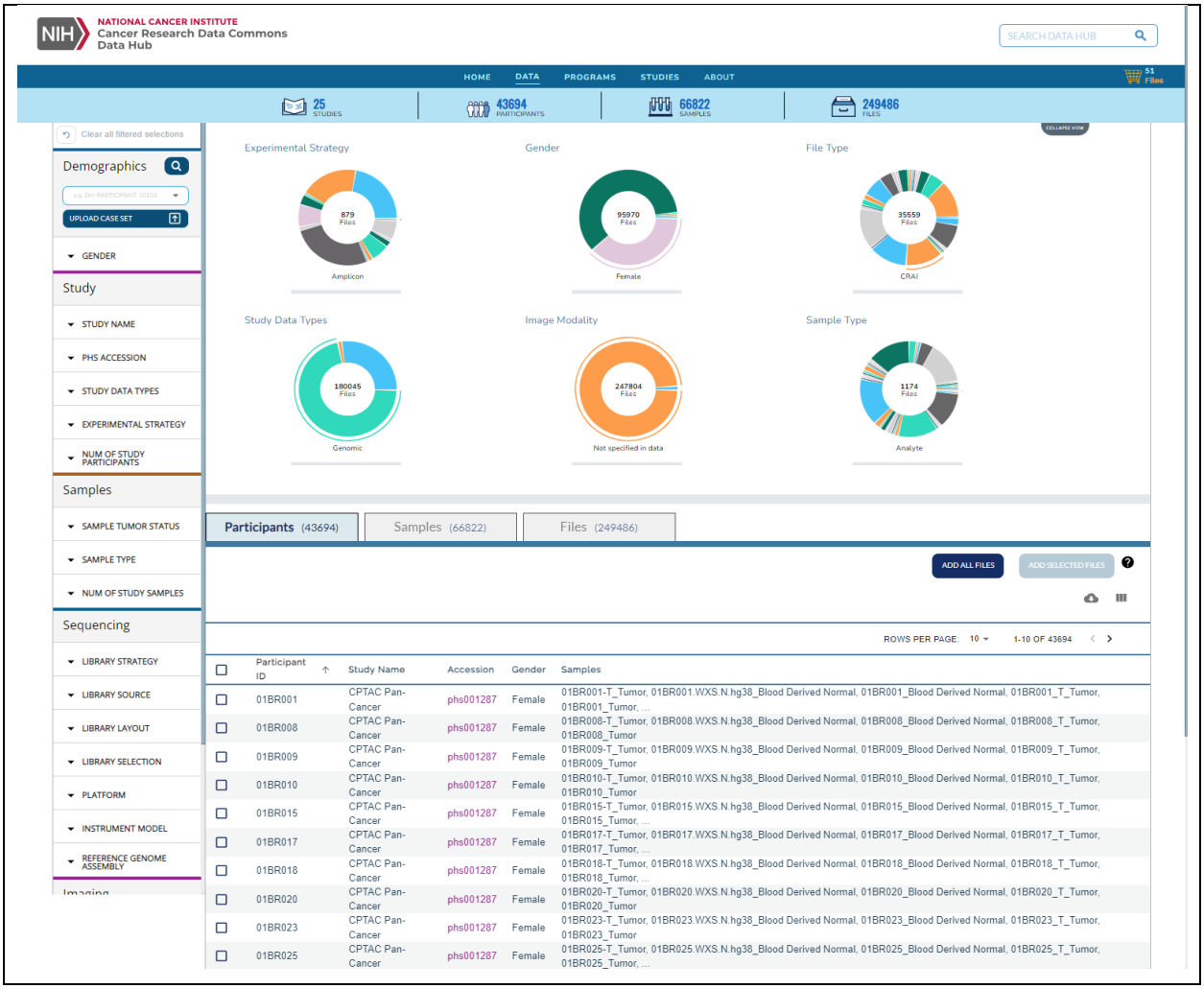


1. **HOME** – Use this page to generally explore the CDS.
2. **DATA** – This tab opens the Data page where users can search studies using various filters.
3. **PROGRAMS** – This tab opens the Program listing page, where users can learn more about the programs that submit data to CDS, as well as associated studies within the program.
4. **STUDIES** – This tab opens the studies listing page where the user can see the studies that have submitted data to CDS
5. **ABOUT** – The About tab is a drop-down list that opens several pages with information about CDS, including:
   - A link to the submission requests
   - GraphQL API interface,
   - Data Model,
   - CDS data and software releases
6. Stats bar across the top of the page
   - **STUDIES** – A count of all the studies in CDS
   - **PARTICIPANTS** – A count of all the participants in CDS

- **SAMPLES** – A count of all the samples in CDS
- **FILES** – A count of all the files in CDS

CDS Data Page

The following is the CDS Data page. A description of its features follows the image.



1. The following data filters are listed on the left side of the page.

   Note: CDS is file-centric, so the counts shown for each filter also represent the number of files.

   - **STUDY** – If this study included controlled access data, the title will correspond to the dbGaP Study Name **EXPERIMENTAL STRATEGY** – The type of study or experimental data generated for testing or research purposes.
   - **SAMPLE TUMOR STATUS** – Normal or Tumor Sample Pathology Indicator
   - **SAMPLE TYPE** - Text term to describe the source of a biospecimen used for a laboratory test

- **GENDER** – Text designations that identify gender. Gender is described as the assemblage of properties that distinguish people based on their societal roles.
- **FILE TYPE** – A defined organization or layout representing and structuring data in a computer file.
- **PHS ACCESSION** – dbGaP study accession number
- **STUDY DATA TYPES** – The type of study or experimental data generated for testing or research purposes
- **IMAGE MODALITY** - The method in which the images are generated
- **LIBRARY STRATEGY** – The overall strategy for the collection of double-stranded DNA fragments flanked by oligonucleotide sequence adapters to enable their analysis by high-throughput sequencing.
- **LIBRARY SOURCE** – The source of a sample collection of double-stranded DNA fragments analyzed by high-throughput sequencing.
- **LIBRARY SELECTION** – The type of systematic actions performed to select or enrich DNA fragments used in analysis by high-throughput sequencing.
- **LIBRARY LAYOUT** – The read strategy or method that was used for sequencing and analysis of a nucleotide library.
- **PLATFORM** – The words used to describe the instrument used to carry out a high-throughput sequencing experiment.
- **INSTRUMENT MODEL** – The words used to describe the specific model of the instrument used to carry out a high-throughput sequencing experiment.
- **REFERENCE GENOME ASSEMBLY** – One or more characters are used to identify the published NCBI genetic sequence that is used as a reference against which other sequences are compared.
- **PRIMARY DIAGNOSIS** – Text term used to describe the patient's histologic diagnosis, as described by the World Health Organization's (WHO) International Classification of Diseases for Oncology (ICD-O)
- **NUM OF STUDY PARTICIPANTS** – Number of participants in the selected study/studies
- **NUM OF STUDY SAMPLES** – Number of samples in the selected study/studies

2. Facets (the four circles – donuts – on the top half of the screen)
    - **Experimental Strategy** – a visual representation of the various types of experimental strategies in the CDS Portal
    - **Gender** – visual representation of the files for various genders in the portal
    - **File Type** – a visual representation of all the count of file types in CDS
    - **Study Data Types** – a visual representation of the count of all the study data types in CDS
    - **Image Modality** – The method in which the images are generated

- **Sample Type** – Text term to describe the source of a biospecimen used for a laboratory test

3. Data Table (the light blue header across the top of the screen)
    - **Participants** – The participants tab contains the number of participants in the study/studies
    - **Samples** – The number of samples in the study/studies
    - **Files** – the counts of the files

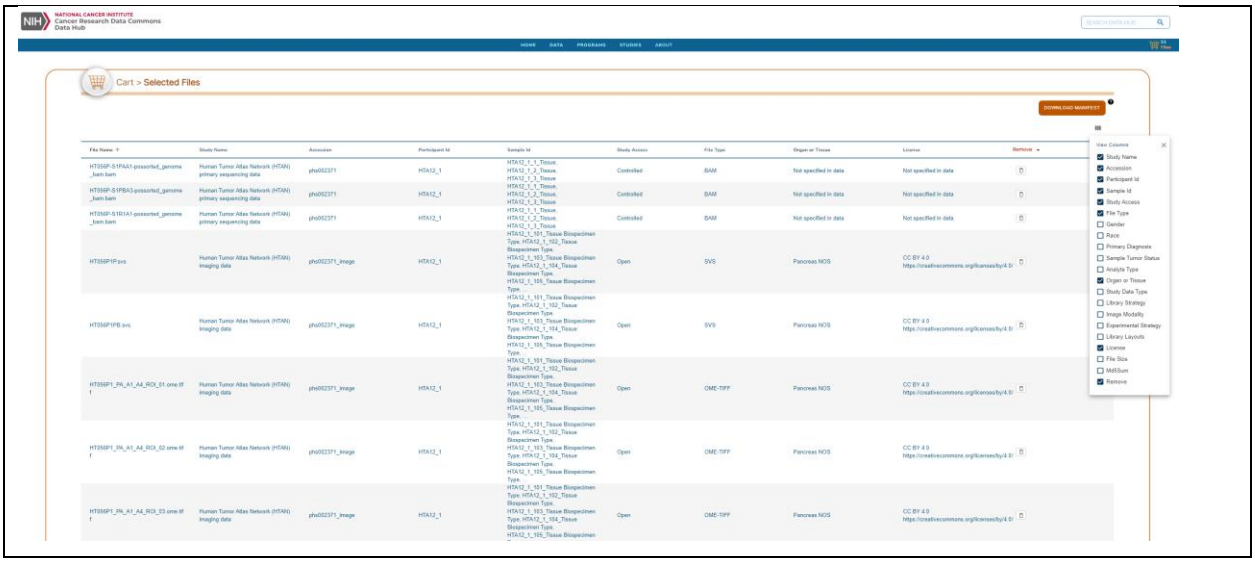
Adding Files to the Cart for Analysis

Files of interest for analysis can be selected and added to the cart as follows:

1. Select the study or PHS of interest.
2. Select all filters/metadata elements of interest.
3. The table below the cart (lower-right section of the screen) populates with the selected files.
4. Click the box for the files to be added to the cart.
5. Click **ADD SELECTED FILES**, which loads the selected files to the cart.
6. Click the cart icon on the top right of the page, which opens the cart page.

CDS Cart Page

Using the Cart page, users can download a data manifest to import and use on the SBG-GCG cloud resource.



1. While on the Cart page, create a description for your file in the text box on the bottom left of the screen (this is optional).
2. Users can immediately export the manifest directly to Velsera or click the **DOWNLOAD MANIFEST** button on the top-right corner of the Cart page.
3. Open the downloaded folder on your computer. The manifest will contain the elements which are selected on the cart page from the dropdown. Please note that the following metadata elements are default and will autopopulate:
   - DRS URI (File ID)
   - Note that this DRS File ID will be needed when you import the manifest to the SBG-CGC
   - Name
   - Participant ID

- Md5sumUser Comments

This manifest can be imported for analysis on SBG-CGC by following the steps below:

Exporting to SBG-CGC/Velsera

1. Create an account or log in at the following URL: https://www.cancergenomicscloud.org/.
2. Create a new project or select an existing project suitable for digital access to the files listed in the file manifest.
3. Navigate to the SBG-CGC dashboard's navigation bar, choose "Files," and click the "+ Add files" button in the dropdown menu.
4. From the dropdown menu, select "GA4GH Data Repository Service (DRS)."
5. Choose "From a manifest file" and click "Browse manifest" to import the files with the DRS URI attached, generated from the CDS portal. (Typically found in your downloads folder or on the desktop for easy access.)
6. Utilize the free-text search box to add relevant tags or comments associated with the batch of imported files and click "Import."
7. Check the Data Use Agreement box to confirm compliance with the data guidelines set by the SBG-CGC. Then, click submit to access your files in the created project.
8. Use free-text to create tags for associating with your data.
9. DRS (Data Repository Service) identifiers will now be displayed for each file from the imported file manifest within the selected SBG-CGC project.
10. These files are accessible in the SBG-CGC Genome Browser and can be selected as inputs for downstream analysis.

How to use GraphQL interface:

While the filtering and visualization capabilities of the CDS portal help researchers find basic information about the CDS housed data, they can also use GraphQL interface to extract more detailed information about studies, participants and samples.

Users can access the GraphQL interface using Jupyter or directly querying the interface. Below are a few queries which can be run directly from the GraphQL interface on CDS:

Query #1: Look at the CDS portal schema

```
{
```

```
__schema{
 types{
   name,
   description,
   fields{
     name
   }
 }
}

}
```

Query #2: To see all relational data for a specific study

```
{
 study (phs_accession: "phs002371"){
   study_name
   study_acronym
   study_description
   short_description
   study_external_url
   primary_investigator_name
   primary_investigator_email
   co_investigator_name
   co_investigator_email
   phs_accession
   bioproject_accession
   index_date
   cds_requestor
   funding_agency
   funding_source_program_name
   grant_id
   clinical_trial_system
   clinical_trial_identifier
   clinical_trial_arm
   number_of_participants
   number_of_samples
   study_data_types
   file_types_and_format
   size_of_data_being_uploaded
   size_of_data_being_uploaded_unit
   size_of_data_being_uploaded_original
   size_of_data_being_uploaded_original_unit
   acl
   study_access
 }

}
```

Query #3: Retrieve file information in CDS

```
{
 file{
   file_id,
   file_name,
   file_description
 }
```

```
}
```

Query #4: Finding all files in a specific study

```
{
  study(phs_accession: "phs002371"){
    files{
      file_id,
      file_name,
      file_description
    }
  }
}
```

Query #5: Retrieve all the diagnosis information including participant ID, gender, race, ethnicity

```
 {
diagnosis{
  age_at_diagnosis,
  participant {
    participant_id
    race
    gender
    ethnicity
    dbGaP_subject_id
  },
  disease_type
}
}
```

Query #6: Retrieve an association between, file, participant, and sample

```
 {
participant{
  participant_id,
  samples{
    sample_id,
    files{
      file_id,
      file_name,
      file_description
    }
  }
}
}
```

Query #7: Retrieve a file and sample association for a specified participant

```
{
participant(participant_id:"PBBJUH"){
  dbGaP_subject_id,
  samples{
    sample_id,
    files{
      file_id,
```

```
      file_name,
      file_description
    }
  }
}

}
```