cBioPortal Data Loading Steps

11/25/2024 Yizhen Chen

Table of Contents

[Data Loading Steps Overview 3](#_Toc183429163)

[Step 1: System Environment Setup 3](#_Toc183429164)

[Build and Run cBioPortal Application 3](#_Toc183429165)

[Software Requirements 3](#_Toc183429166)

[Build System 3](#_Toc183429167)

[Build cBioPortal Core environment (data loading script) 4](#_Toc183429168)

[Load-prepared data 4](#_Toc183429169)

[Create database 4](#_Toc183429170)

[Download prepared gene panel data 5](#_Toc183429171)

[Load custom data 5](#_Toc183429172)

[Prepare CCDI data 5](#_Toc183429173)

[Download CCDI Hub Data 5](#_Toc183429174)

[Reformatted CCDI hub data 5](#_Toc183429175)

[Cancer Study 5](#_Toc183429176)

[Cancer Type 6](#_Toc183429177)

[Clinical Data-meta-file-sample 7](#_Toc183429178)

[Load Study data files 9](#_Toc183429179)

# Data Loading Steps Overview

1. Setup System Environment
2. Load Seed Gene data
3. Load custom data

# Step 1: System Environment Setup

## Build and Run cBioPortal Application

### Software Requirements

Java version 21

MySQL

Apache Maven

### Build System

Get the latest code

git clone https://github.com/cBioPortal/cbioportal.git

cd cbioportal

git checkout master

**Prepare the global configuration file**

cd src/main/resources

cp application.properties.EXAMPLE application.properties

**update the database setting in application.proterties file**

**make sure the datasource.url contains**

**useSSL = false;**

allowPublicKeyRetrieval=true

allowLoadLocalInfile=true

# database

spring.datasource.url=jdbc:mysql://127.0.0.1:3306/cbioportal?useSSL=false&allowPublicKeyRetrieval=true&allowLoadLocalInfile=true

spring.datasource.username=cbio\_user

spring.datasource.password=123456

spring.datasource.driver-class-name=com.mysql.jdbc.Driver

spring.jpa.database-platform=org.hibernate.dialect.MySQL5InnoDBDialect

**Build with Maven**

**mvn -DskipTests clean install**

**Run the cBioProtal**

java -jar target/cbioportal-exec.jar

java -jar target/cbioportal-exec.jar -Dauthenticate=false

### Build cBioPortal Core environment (data loading script)

**Build cbioPortal loading script**

git clone https://github.com/cBioPortal/cbioportal-core.git

cd cbioportal-core

git checkout main

**Configure the cbioportal-core**

Create application.properties, save to cbioportal-core src/main/resource

Content of application.properties is database connection, same with cBioportal

**Set the PORTAL\_HOME environment variable**

PORTA\_HOME point to cbioportal application directory

export PORTAL\_HOME=/Users/johndoe/cbioportal

**Compile and build**

mvn -DskipTests clean install

**create Python environment**

**Create a Python virtual environment (first-time setup):**

python -m venv .venv

**Activate the virtual environment:**

source .venv/bin/activate

**Install required Python dependencies (first-time setup or when dependencies have changed):**

pip install -r requirements.txt

# Load-prepared data

### Create database

Use the script to create database.

[**https://github.com/cBioPortal/cbioportal/blob/master/src/main/resources/db-scripts/cgds.sql**](https://github.com/cBioPortal/cbioportal/blob/master/src/main/resources/db-scripts/cgds.sql)

### Download prepared gene panel data

Follow the instruction here

[**https://github.com/cBioPortal/datahub/tree/master/seedDB**](https://github.com/cBioPortal/datahub/tree/master/seedDB)

# Load custom data

### Prepare CCDI data

The CCDI Hub contains data elements such as studies, samples, and participants, which can be formatted to align with cBioPortal's file structure, including Cancer Study, Cancer Type, and Clinical Data. As of March 2016, the Clinical Data file in cBioPortal has been divided into two separate files: a patient clinical file and a sample clinical file.

CCDI data can be downloaded from the Hub and reformatted to meet the requirements of cBioPortal.

### Download CCDI Hub Data

Download the descriptive information about a study from CCDI explorer page.

<https://ccdi.cancer.gov/explore>

### Reformatted CCDI hub data

Case study on Molecular Characterization Initiative (MCI) Study

#### Cancer Study

Name: meta\_study.txt

This file contains metadata about the cancer study. The file contains the following fields:

|  |  |  |  |
| --- | --- | --- | --- |
| field | value | map to CCDI MODEL | comment |
| type\_of\_cancer | MIXED | N/A | The cancer type abbreviation, e.g., "brca". This should be the same cancer type as specified in the meta\_cancer\_type.txt file, if available. The type can be "mixed" for studies with multiple cancer types. |
| cancer\_study\_identifier | phs002790 | Study.study\_id |  |
| description | \*\*\*\*\*\* | Study. study\_description |  |
| add\_global\_case\_list | TRUE |  | set to 'true' if you would like the "All samples" case list to be generated automatically for you. |

#### Cancer Type

Name: meta\_cancer\_type.txt

If the type\_of\_cancer specified in the meta\_study.txt does not yet exist in the type\_of\_cancer database table, a meta\_cancer\_type.txt file is also mandatory.

The file is comprised of the following fields:

1. **genetic\_alteration\_type**: CANCER\_TYPE
2. **datatype**: CANCER\_TYPE
3. **data\_filename**: your datafile

|  |  |  |  |
| --- | --- | --- | --- |
| field | value | map to CCDI MODEL | comment |
| **genetic\_alteration\_type** | CANCER\_TYPE | N/A |  |
| **datatype** | CANCER\_TYPE | N/A |  |
| data\_filename | cancer\_type.txt | N/A |  |
|  |  |  |  |

Name: cancer\_type.txt

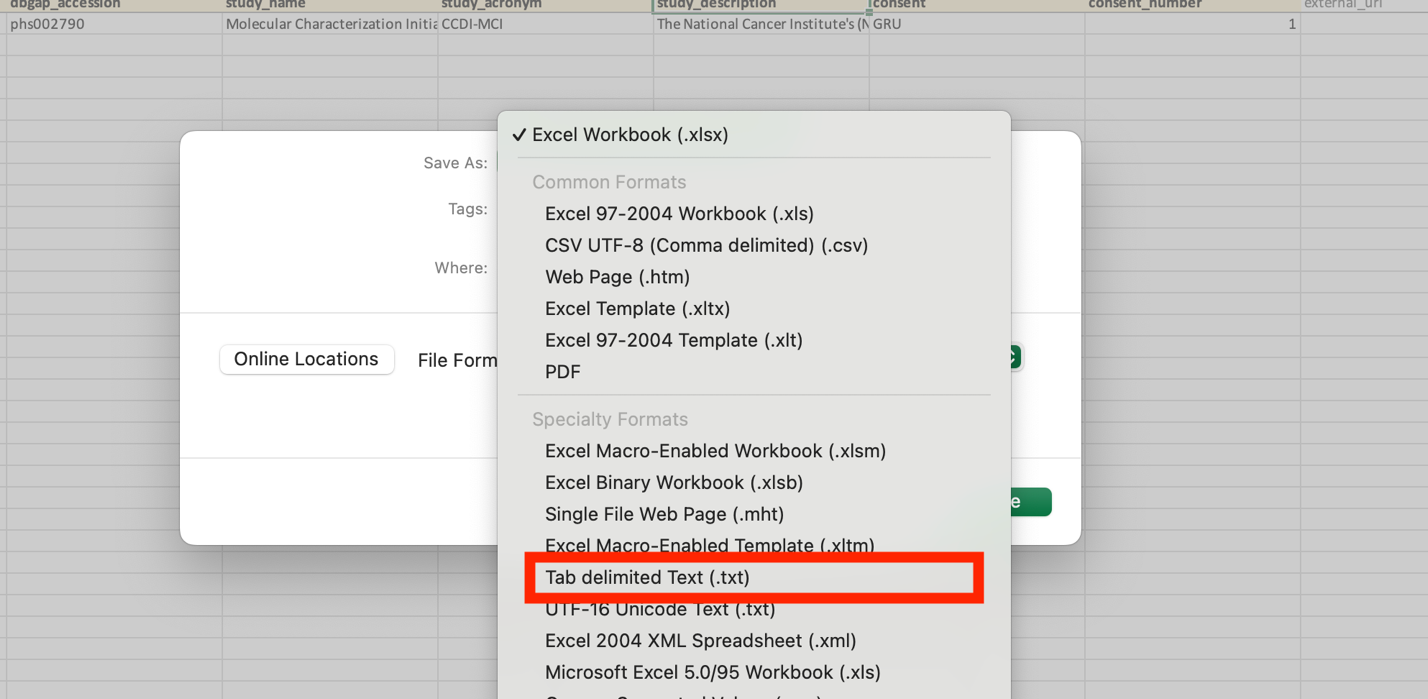
The file is comprised of the following columns in the order specified:

1. **type\_of\_cancer**: The cancer type abbreviation, e.g., "brca".
2. **name**: The name of the cancer type, e.g., "Breast Invasive Carcinoma".
3. **dedicated\_color**: CSS color name of the color associated with this cancer study, e.g., "HotPink". See [this list](https://www.w3.org/TR/css3-color/#svg-color) for supported names, and follow the [awareness ribbons](https://en.wikipedia.org/wiki/List_of_awareness_ribbons) color schema. This color is associated with the cancer study on various web pages within the cBioPortal.
4. **parent\_type\_of\_cancer**: The type\_of\_cancer field of the cancer type of which this is a subtype, e.g., "Breast". ℹ️ : you can set parent to tissue, which is the reserved word to place the given cancer type at "root" level in the "studies oncotree" that will be generated in the homepage (aka query page) of the portal.

It is a txt file and the content is delimited by tab

Here is example: brca<TAB>Breast Invasive Carcinoma<TAB>HotPink<TAB>Breast

I would recommend use Excel sheet to fill in the content and save as Tab delimited Text to avoid any formatting issues.



#### Clinical Data-meta-file-sample

The clinical file is split into a patient clinical file and a sample clinical file.

The *sample* file is required, whereas the *patient* file is optional.

**Sample meta file**

|  |  |  |  |
| --- | --- | --- | --- |
| field | value | map to CCDI MODEL | comment |
| cancer\_study\_identifier | phs002790 | N/A | Referes to cancer\_study\_identifier on meta\_study file |
| **genetic\_alteration\_type** | CLINICAL | N/A |  |
| dataType | SAMPLE\_ATTRIBUTES | N/A |  |
| data\_filename | data\_clinical\_sample.txt | N/A |  |

**Sample data file**

Name: data\_clinical\_sample.txt

Similar to the cancer\_type.txt and content is delimited by tab

You can put any sample data you like to the data\_clinical\_sample.

**Clinical sample columns**

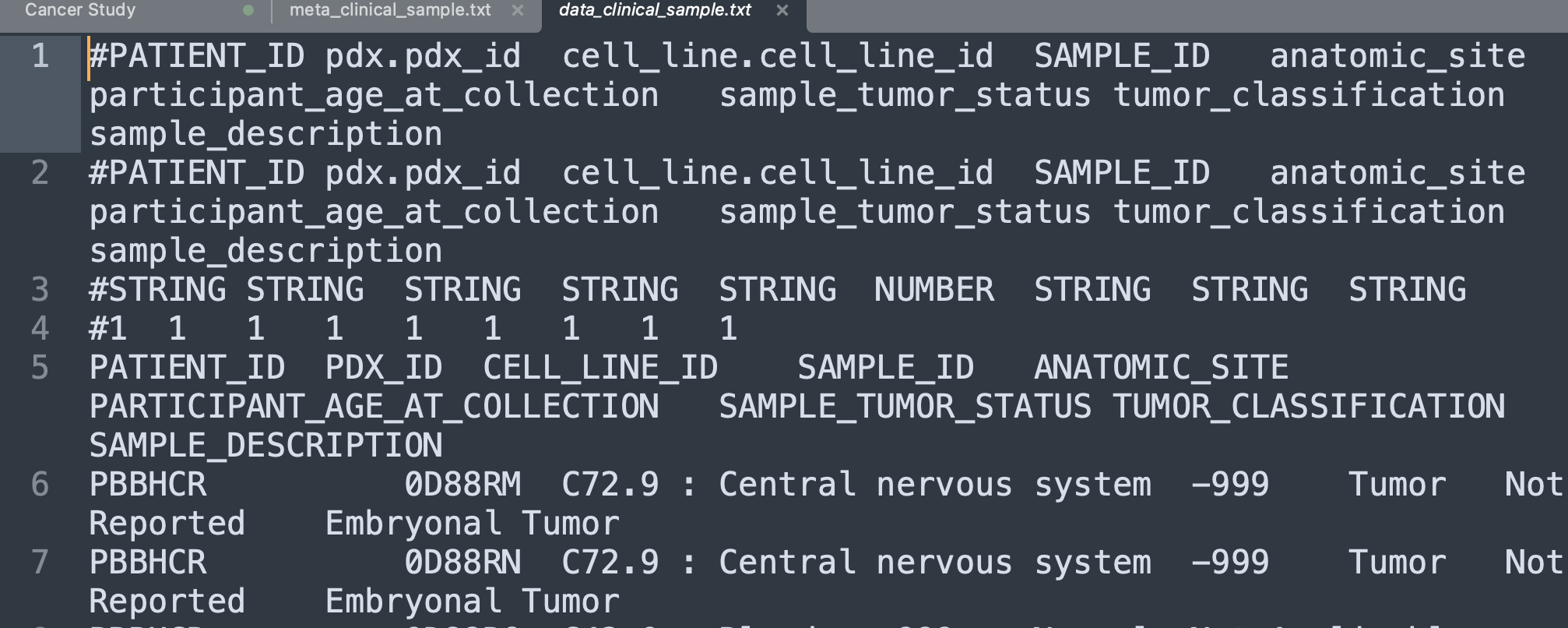
The file containing the sample attributes has two **required** columns:

* **PATIENT\_ID (required)**: A patient ID. This field can only contain numbers, letters, points, underscores and hyphens.
* **SAMPLE\_ID (required)**: A sample ID. This field can only contain numbers, letters, points, underscores and hyphens.

The first four rows of the clinical data file contain tab-delimited metadata about the clinical attributes. These rows **have to start with a '#' symbol**. Each of these four rows contain different type of information regarding each of the attributes that are defined in the fifth row:

* Row 1: **The attribute Display Names**: The display name for each clinical attribute
* Row 2: **The attribute Descriptions**: Long(er) description of each clinical attribute
* Row 3: **The attribute Datatype**: The datatype of each clinical attribute (must be one of: STRING, NUMBER, BOOLEAN)
* Row 4: **The attribute Priority**: A number which indicates the importance of each attribute
* Row 5: The **attribute name for the database**: This name should be in **upper case.**
* Row 6: This is the first row that contains actual **data**.

Example



**Patient data file (Participant data)**

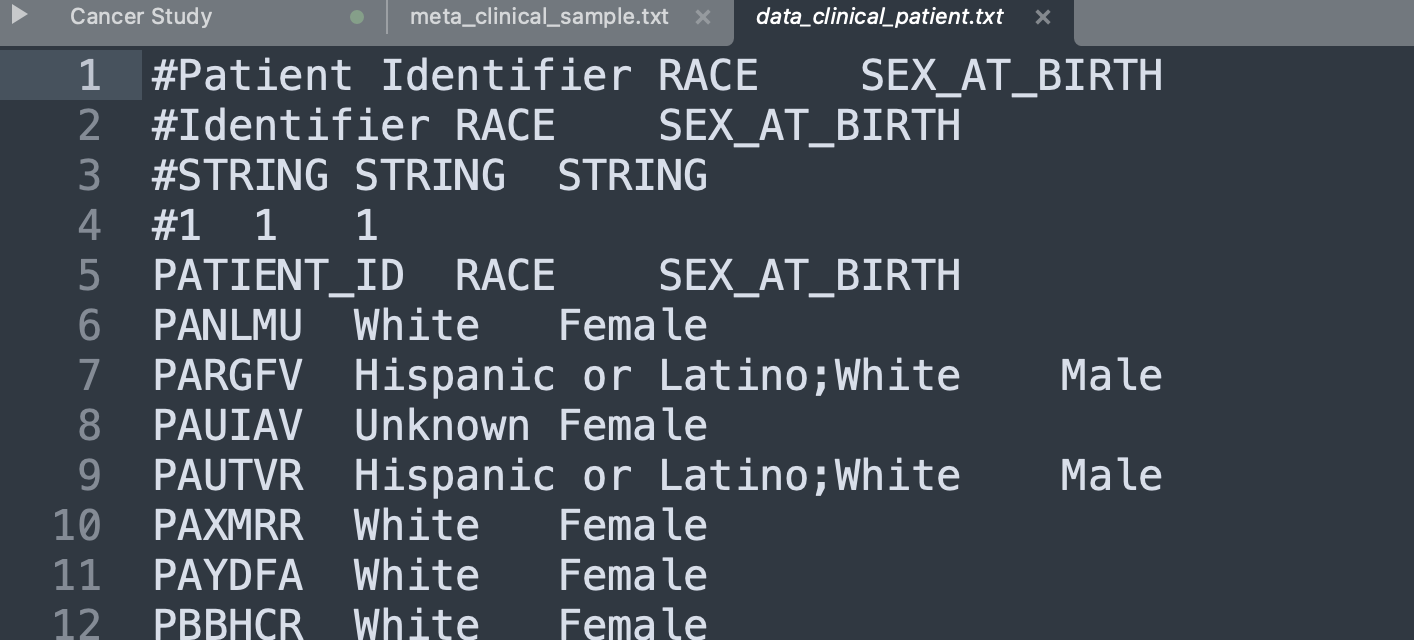
Name :data\_clinical\_patient.txt

Similar to the sample data file, you can include any participant data in the data\_clinical\_patient file, following the same structure as the sample data file.

**Clinical patient columns**

The file containing the patient attributes has one **required** column:

* **PATIENT\_ID (required)**: a unique patient ID. This field allows only numbers, letters, points, underscores and hyphens.



### Load Study data files

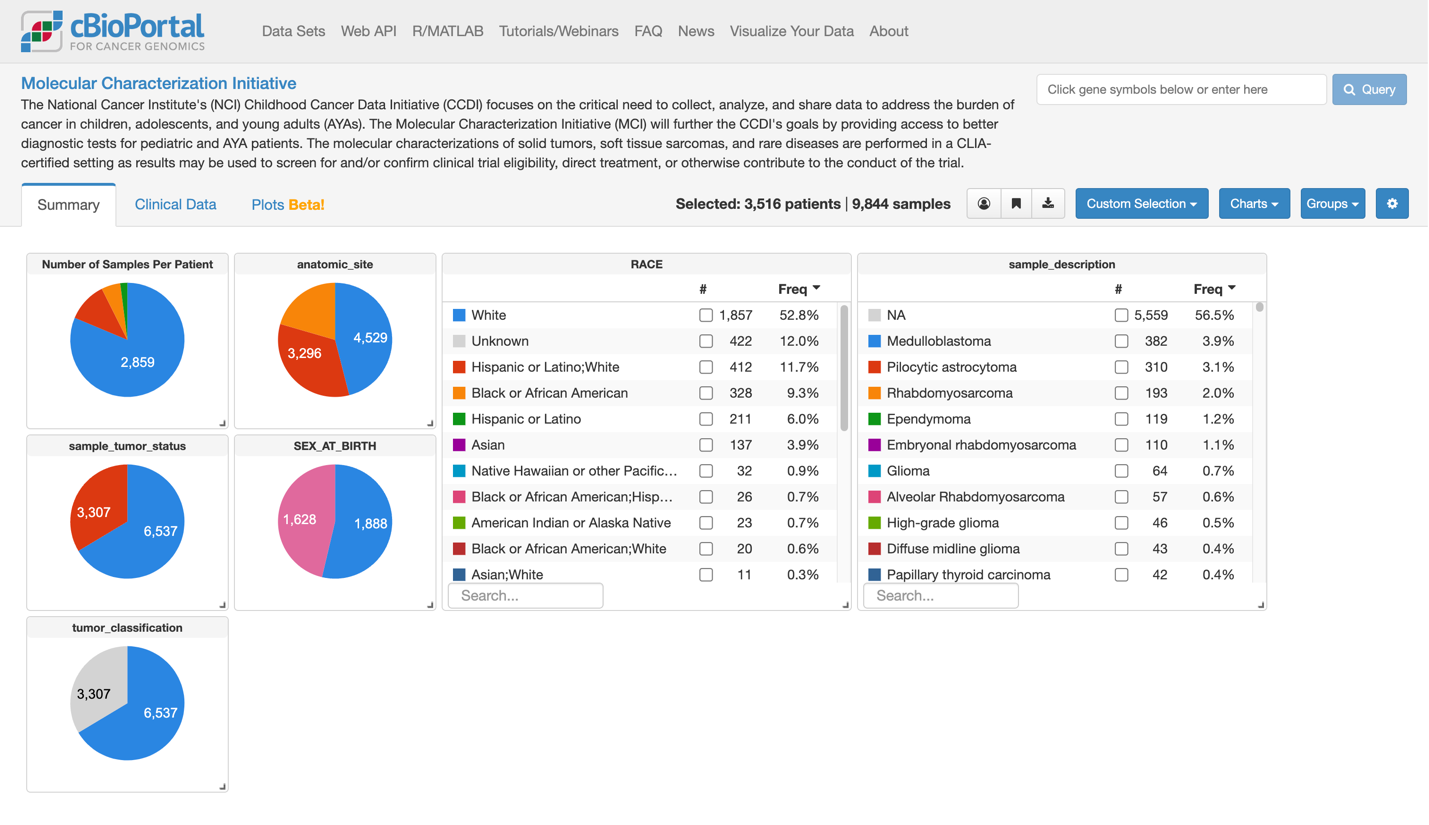
To import the study:

cd cbioportal-core/src/main/scripts/importer

and then run the following command:

./metaImport.py -s ./data\_files\_dir/ -n -o

After loading the study data, please restart the app



Reference:

Document <https://docs.cbioportal.org/deployment/deploy-without-docker/software-requirements/>

GIthub:

<https://github.com/cBioPortal/cbioportal>

<https://github.com/cBioPortal/cbioportal-core>