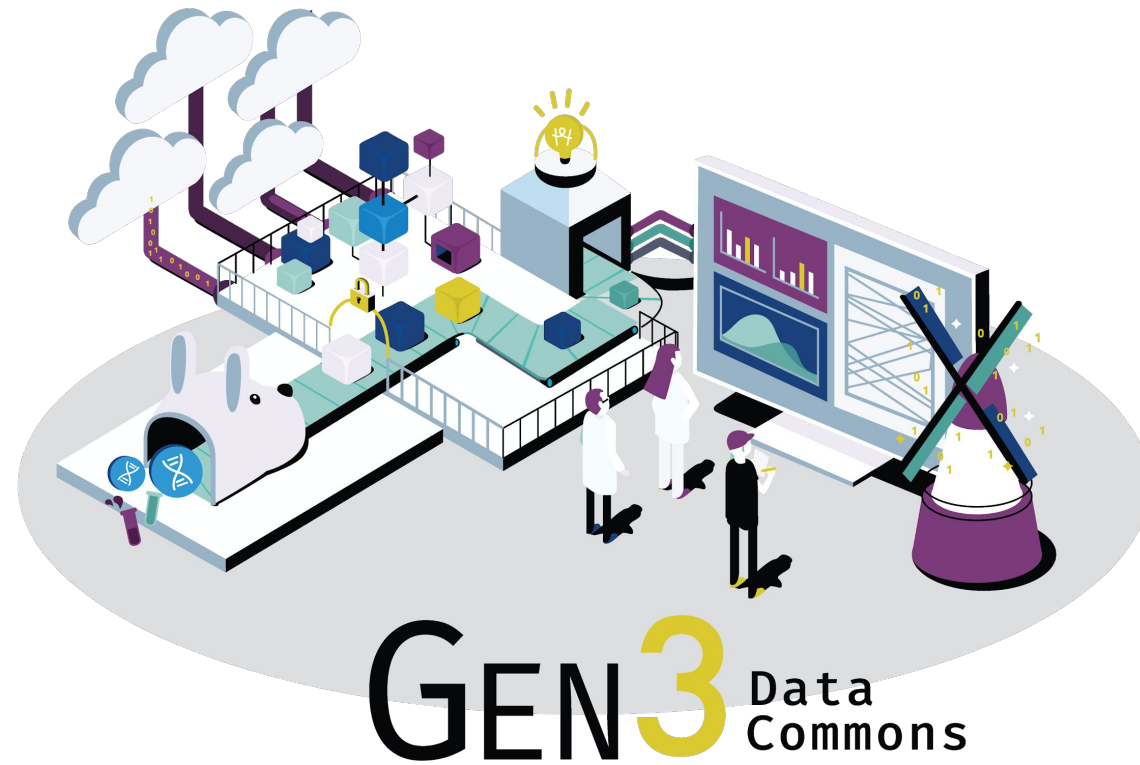


Introduction to Gen3 Data Commons

Michael Fitzsimons, PhD
Director of User Services and Outreach
Center for Translational Data Science
University of Chicago

Agenda

1. Overview of Gen3
2. Who are Gen3 users?
3. Gen3 Features
4. Demo
5. Questions



Data commons co-locate data, storage and computing infrastructure with commonly used software services, tools & apps for analyzing and **sharing data** to create a resource for the research community.*

*Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016.

Overview of Setting Up a Gen3 Data Commons

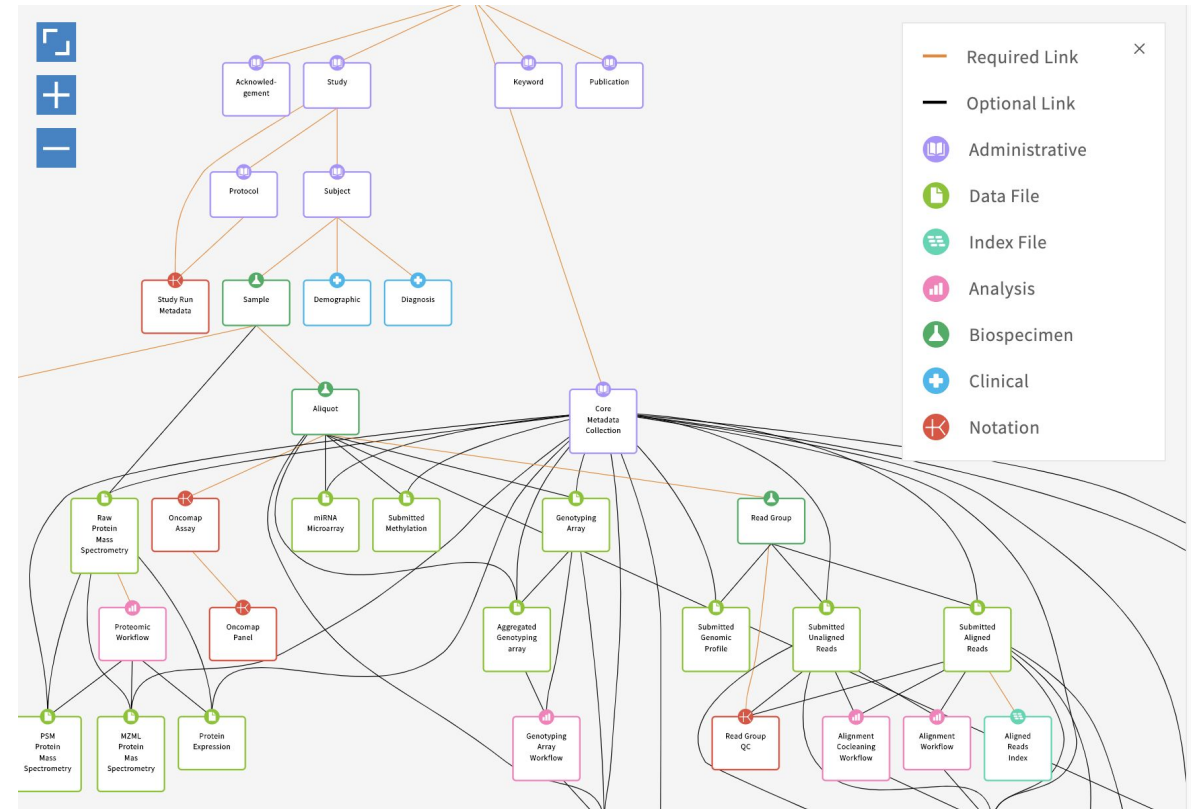
1. Define a data model
2. Use the Gen3 software to auto generate a commons with an API
3. Load data into the commons via the API or through the Data Submission Portal
4. Explore the data with the Data Exploration Portal
5. Analyze the data using Jupyter notebooks
6. Encourage your community to develop their own apps over your commons API

Who are the Gen3 Users?

1. **Users:** Researchers, Scientists, Clinicians who contribute or consume data from a data commons
2. **Developers:** Create Gen3 applications, resources & services
3. **Operators:** Projects that want to operate a Gen3 data commons or interoperate with other Gen3 data commons

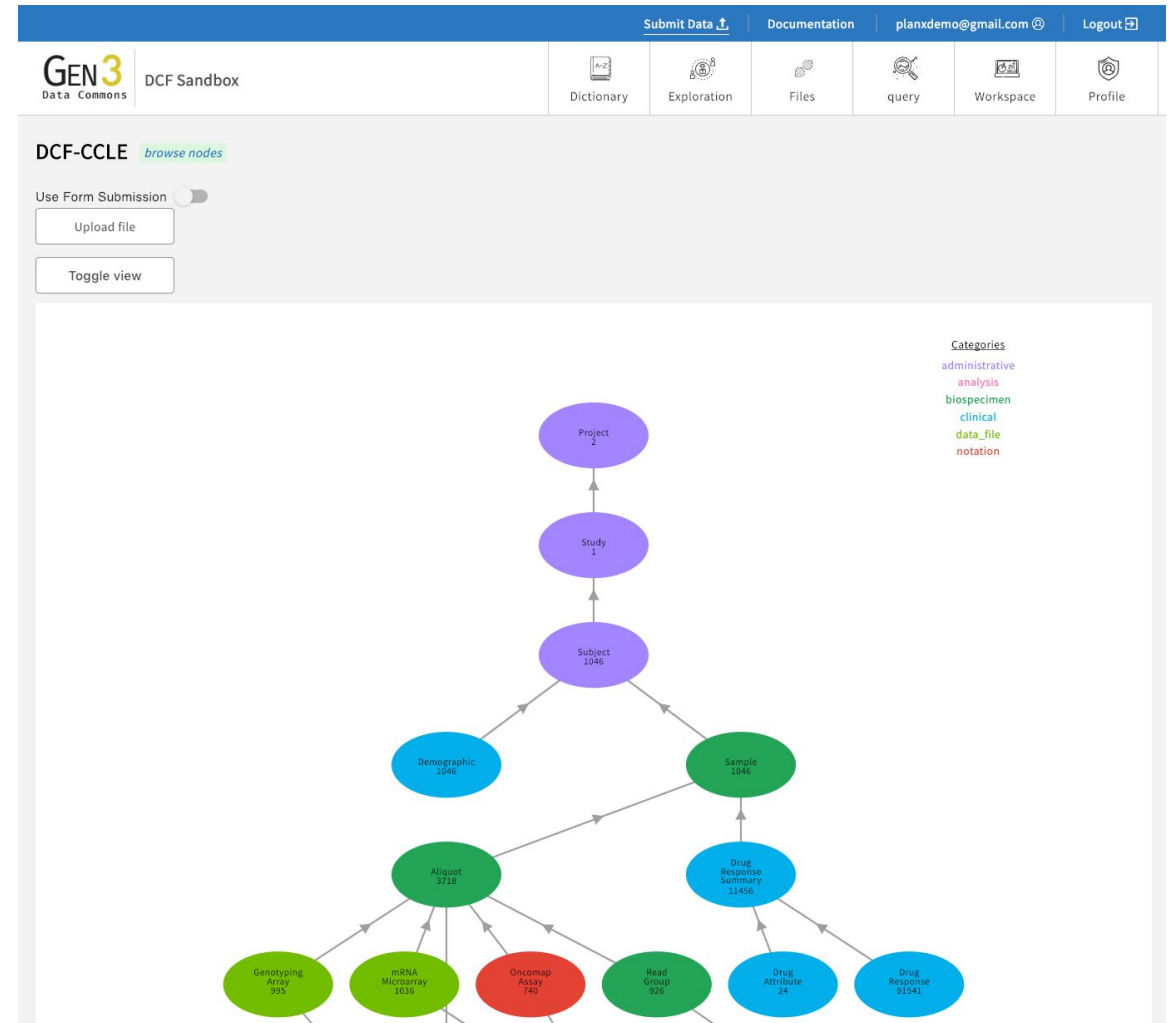
Data Model

- Graph model captures relationship between subjects, clinical, biospecimen, and molecular data
- Data dictionary defines rules for the structured data
- Leverages external terminology standards (eg. NCIt)



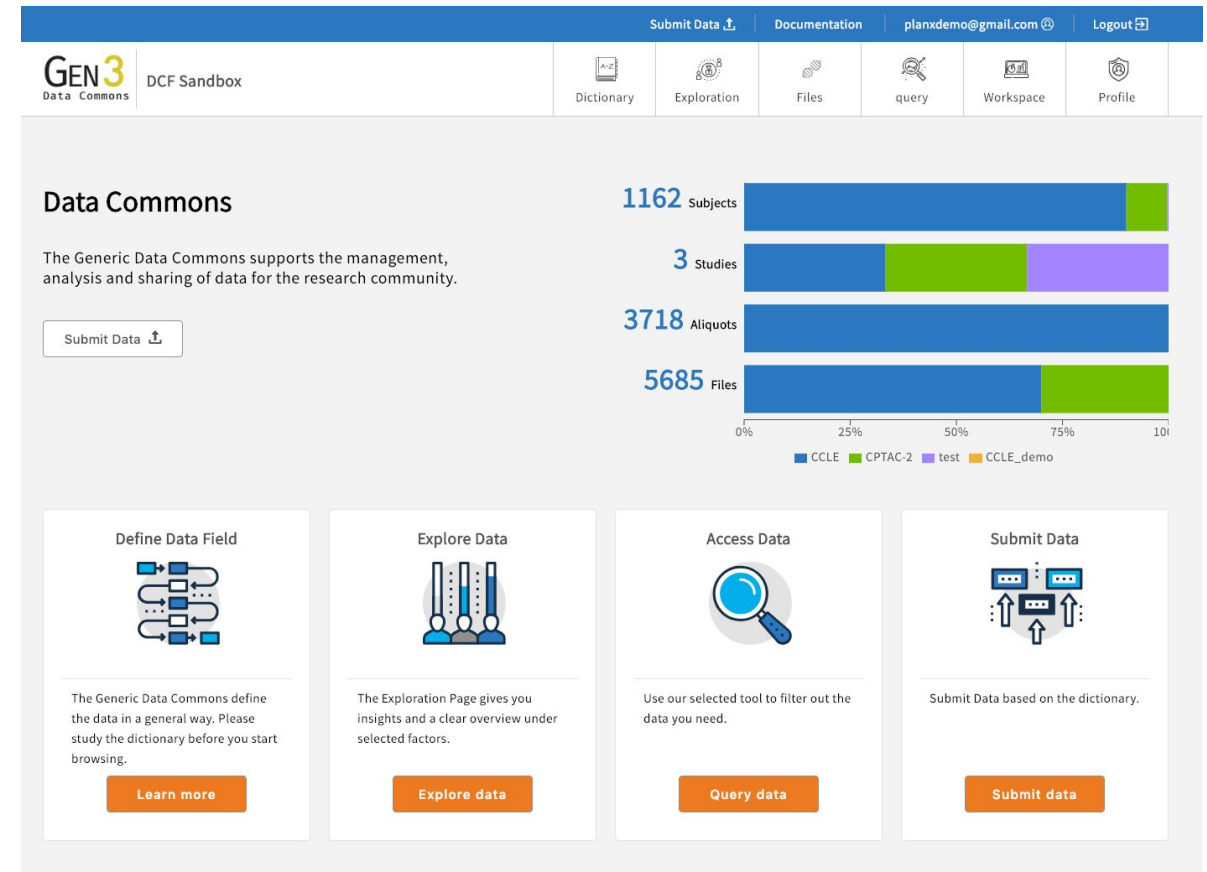
Data Submission

- Sheepdog imports data compliant with the data model and assigns a digital ID to each record and object
- Data curated into graphical data model
- Submission of structured data can be performed through the API or a UI



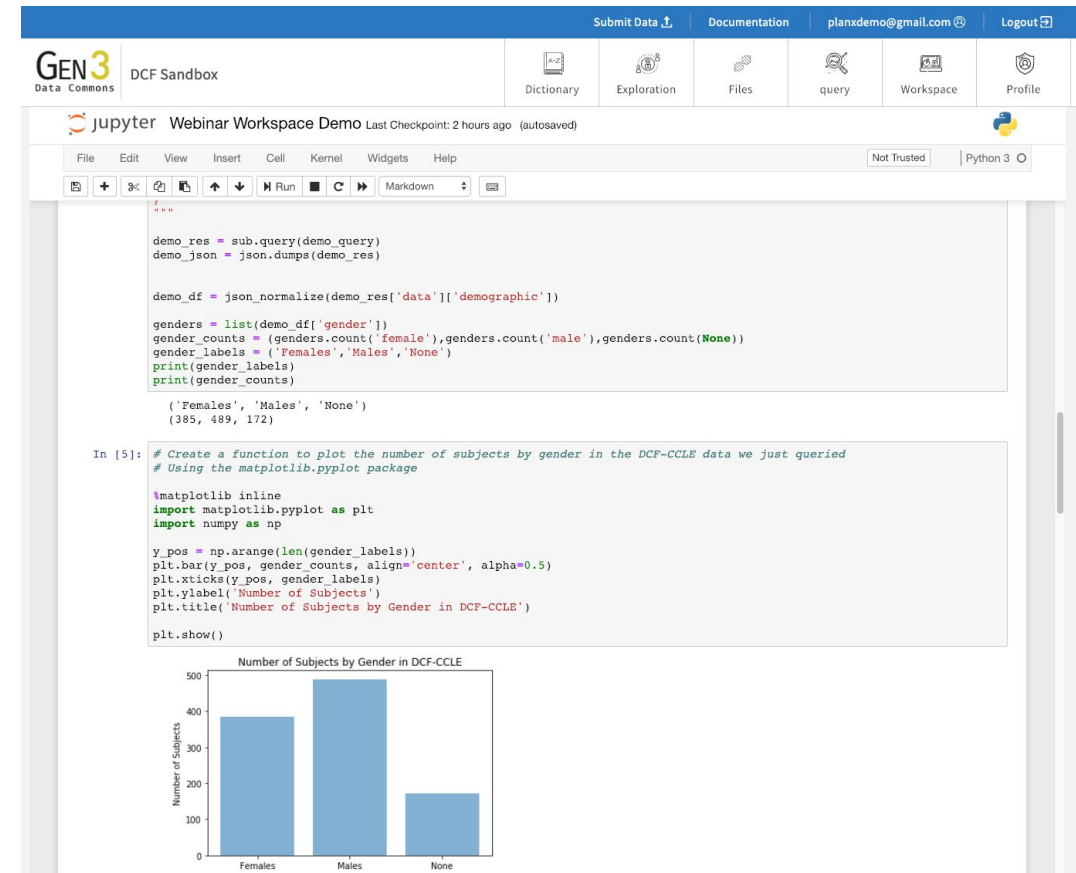
Data Portal – Explore Data & Create Cohorts for Analysis

- Windmill is a data portal for data submission, data search and query, data exploration and data analysis
- Example of app built over a Gen3 data commons



Analyze Your Data

- Lightweight workspaces for analysis and visualization
- Jupyter notebooks powered by Python or R
- Looking to expand to RStudio and Galaxy



The screenshot displays a Jupyter Notebook interface within a web browser. The top navigation bar includes 'GEN3 Data Commons', 'DCF Sandbox', and user information 'planxdemo@gmail.com'. The notebook title is 'Webinar Workspace Demo'. The code cell contains Python code for data processing and visualization:

```
demo_res = sub.query(demo_query)
demo_json = json.dumps(demo_res)

demo_df = json_normalize(demo_res['data']['demographic'])

genders = list(demo_df['gender'])
gender_counts = (genders.count('female'), genders.count('male'), genders.count(None))
gender_labels = ('Females', 'Males', 'None')
print(gender_labels)
print(gender_counts)

('Females', 'Males', 'None')
(385, 489, 172)
```

The second code cell defines a function to plot the data:

```
In [5]: # Create a function to plot the number of subjects by gender in the DCF-CCLC data we just queried
# Using the matplotlib.pyplot package

%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np

y_pos = np.arange(len(gender_labels))
plt.bar(y_pos, gender_counts, align='center', alpha=0.5)
plt.xticks(y_pos, gender_labels)
plt.ylabel('Number of Subjects')
plt.title('Number of Subjects by Gender in DCF-CCLC')

plt.show()
```

The output of the second cell is a bar chart titled 'Number of Subjects by Gender in DCF-CCLC'. The y-axis is labeled 'Number of Subjects' and ranges from 0 to 500. The x-axis has three categories: 'Females', 'Males', and 'None'. The bars show approximately 385 subjects for Females, 489 for Males, and 172 for None.

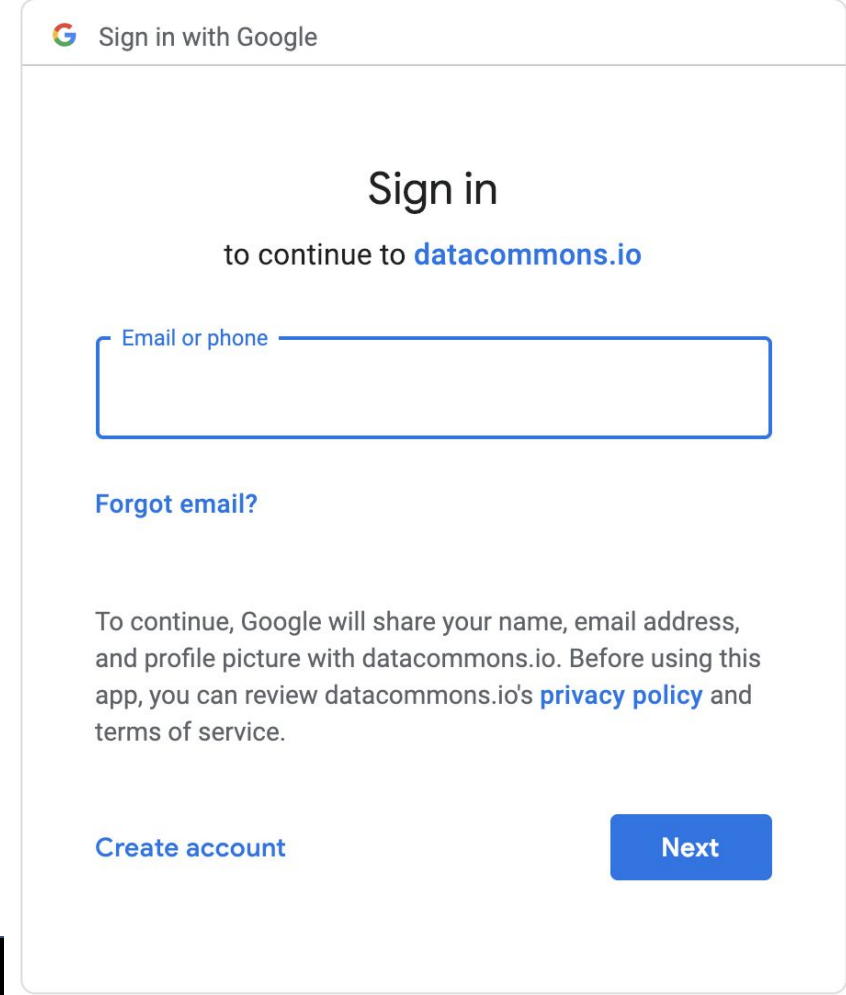
Gender	Number of Subjects
Females	385
Males	489
None	172

Digital ID Service

- IndexD provides permanent data GUIDs (globally unique IDs) for data objects
- Supports multiple URLs for files
- Minimal metadata is md5 hash and object size
- IndexD supports GA4GH DRS API and its own API for read/write

Authentication and Authorization

- Fence separates controlled access resources from the outside world and allows only trusted entities to enter
- Identity providers supported by Fence:
 - Google, eRA Commons, InCommons, eduGain, ...
- Fence utilizes OpenID Connect (OIDC) to generate tokens for clients and directly to a user



The screenshot shows a web interface for signing in with Google. At the top left, there is a Google logo and the text "Sign in with Google". The main heading is "Sign in" followed by "to continue to datacommons.io". Below this is a text input field with the placeholder "Email or phone". Underneath the input field is a link "Forgot email?". A paragraph of text explains that Google will share the user's name, email address, and profile picture with datacommons.io, and provides a link to the "privacy policy" and "terms of service". At the bottom left, there is a link "Create account", and at the bottom right, there is a blue button labeled "Next".

Gen3 APIs

- OpenID Connect (OIDC) for Auth
- IndexD API - GA4GH DRS standard
- GraphQL API for querying submitted data
- Supports app and commons development

Query graph

GraphQL



Prettify

History

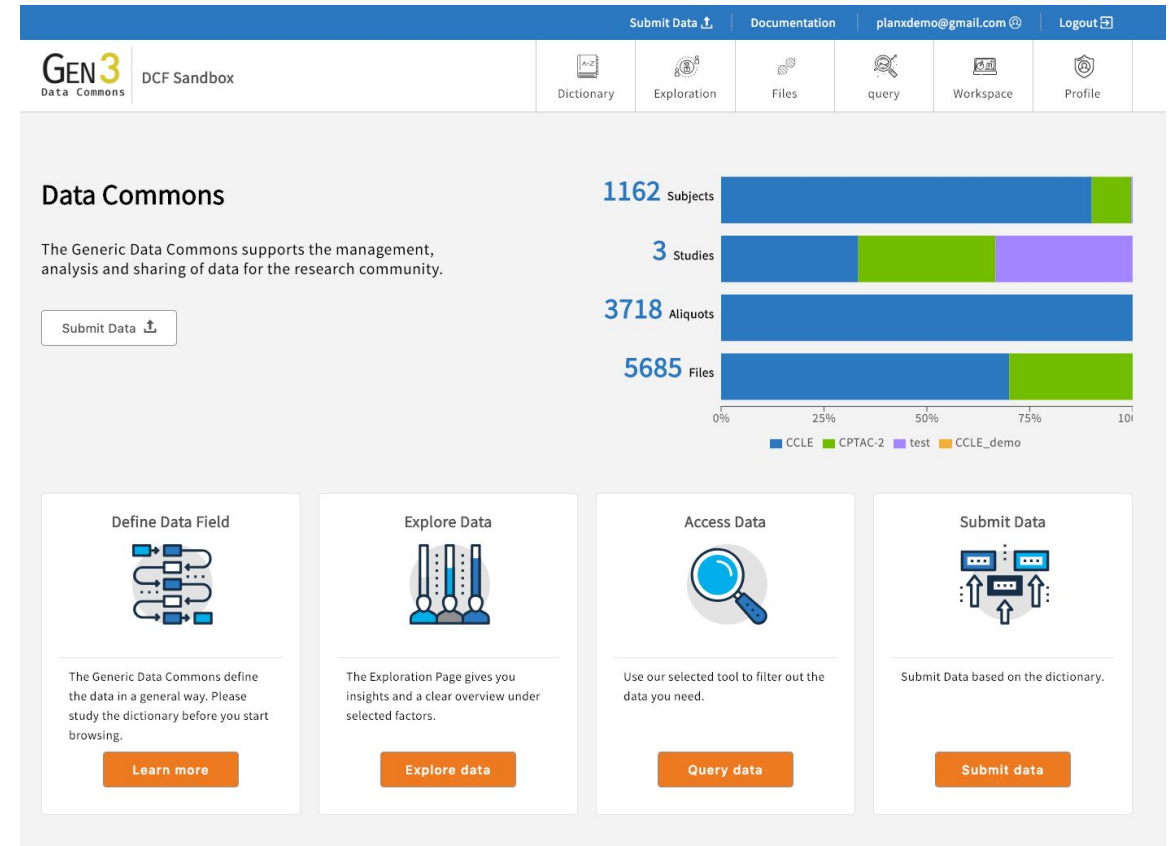
```
1 {  
2   subject(project_id: "DCF-CCLE") {  
3     id  
4     submitter_id  
5     demographics {  
6       gender  
7     }  
8   }  
9 }  
10
```

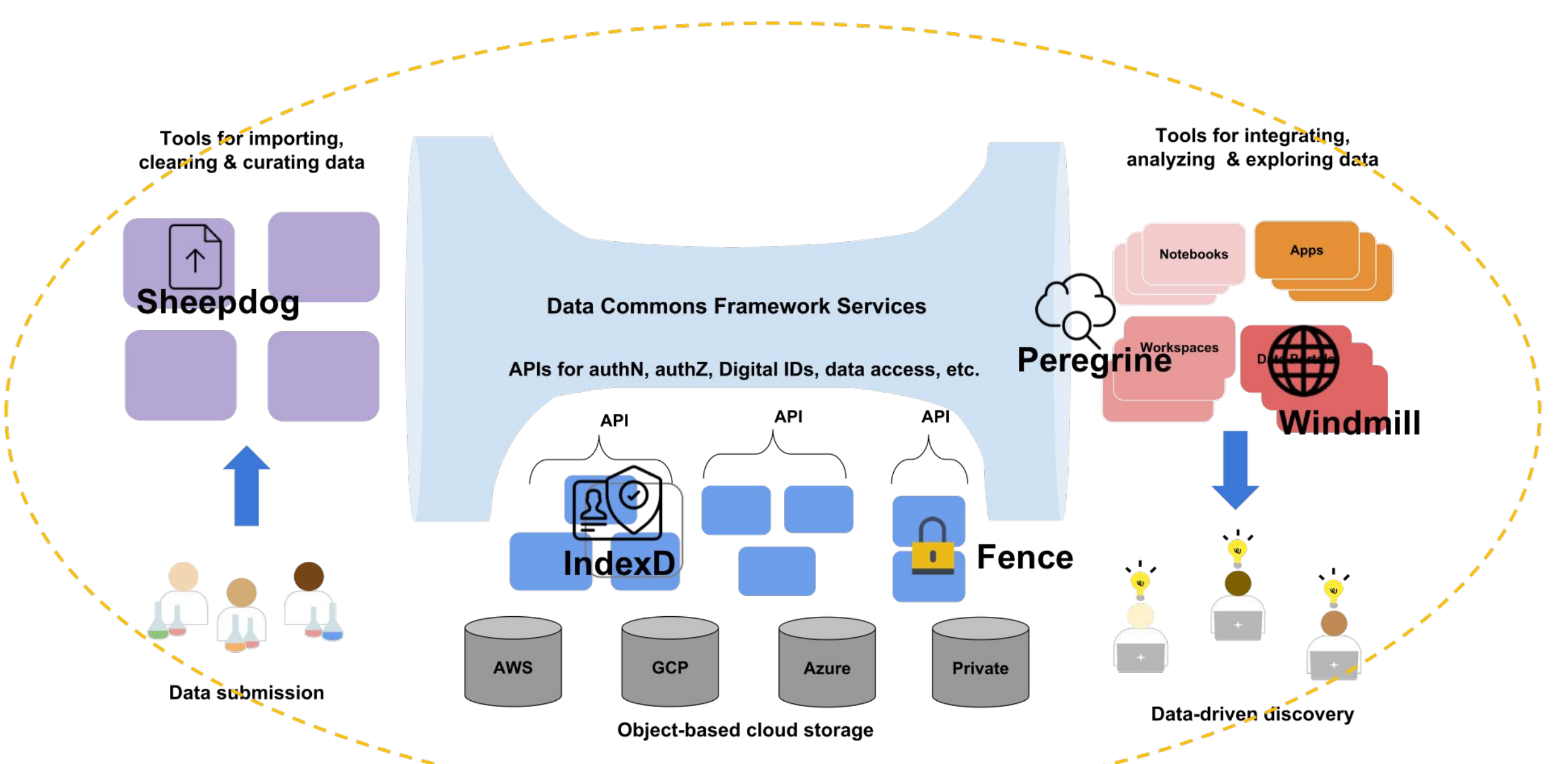
QUERY VARIABLES

```
{  
  "data": {  
    "subject": [  
      {  
        "demographics": [  
          {  
            "gender": "female",  
            "year_of_birth": null  
          }  
        ],  
        "id": "2dd84f5d-28cc-455a-b71e-38919c30055f",  
        "submitter_id": "ZR7530_BREAST_subject"  
      },  
      {  
        "demographics": [  
          {  
            "gender": "female",  
            "year_of_birth": null  
          }  
        ]  
      }  
    ]  
  }  
}
```

Summary of Gen3 Features

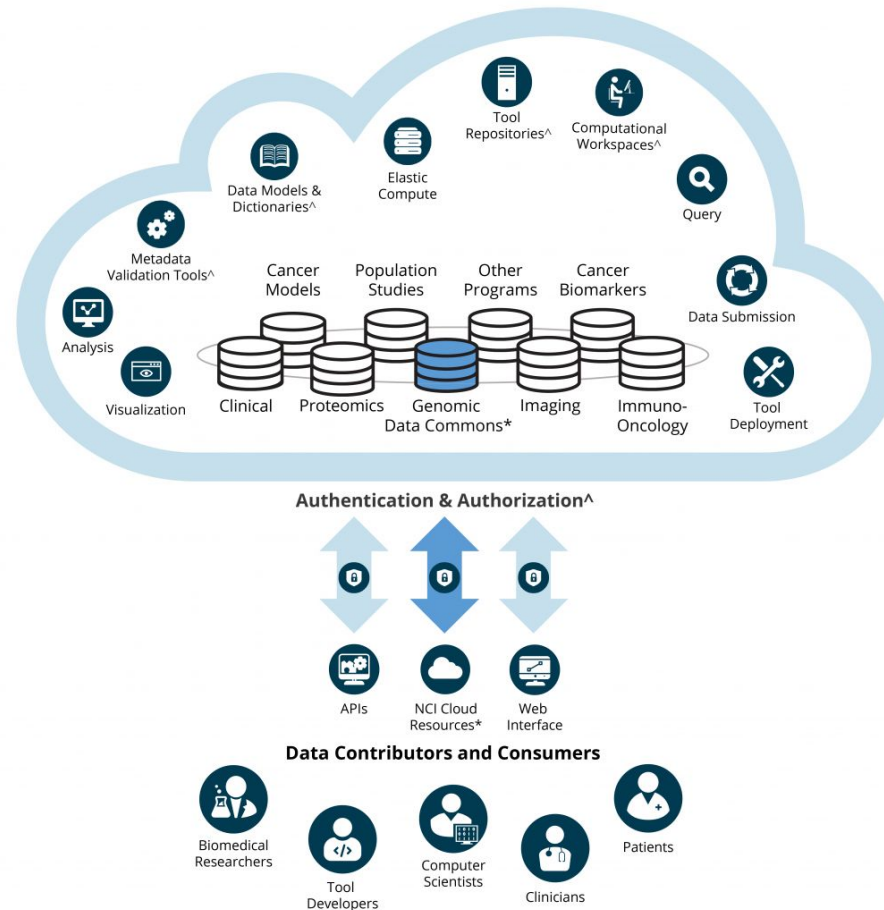
- Authentication/Authorization
- Data Model
- Data Submission
- Digital ID Service
- API
- Data Portal
- Analysis





Cancer Research Data Commons

NCI Cancer Research Data Commons (CRDC)



* The Genomics Data Commons and NCI Cloud Resources are in production and available to the community
^ Components of the Data Commons Framework

Demo

Submit Data Documentation

GEN3
Data Commons | DCF Sandbox

Dictionary Exploration Files query Workspace Profile

DCF Sandbox

EXPLORE, ANALYZE, AND SHARE RESEARCH DATA

The Data Common Frameworks (DCF) supports the management, analysis and sharing of many different types of biomedical data for the research community with the goal of accelerating research in the molecular basis for disease and matching targeted therapies that factor in each patient's unique biology.

[Login from Google](#)

If you have any questions about access or the registration process, please contact support@datacommons.io.

Dictionary v2.5.0 Submission v1.1.7 Portal v2.4.3

GEN3 Data Commons Center for Translational Data Science AT THE UNIVERSITY OF CHICAGO

How can I learn more?



github.com/uc-cdis



gen3.org



Slack Gen3 Community (ask us for an invite!)



dcf-support@datacommons.io

ctds.uchicago.edu

Contribute to the Gen3 Open Source Community

<https://github.com/UC-cdis>

Selected Data Commons Using Gen3



AnVIL