CDIS Biomedical Data Commons

Computational Life Science Seminar Series
October 18, 2017

Michael Fitzsimons
Center for Data Intensive Science
Agenda

• What is a Data Commons?
• Data Commons at CDIS
• NCI GDC Data Commons
  – GDC Data Portal Demo
  – GDC API Demo
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.*


OCC Open Science Data Cloud (2010)

Bionimbus Protected Data Cloud* (2013)

NCI Genomic Data Commons* (2016)

OCC-NOAA Environmental Data Commons (2016)


Kids First Data Resource (2017)

Brain Commons (2017)

*Operated under a subcontract from NCI / Leidos Biomedical to the University of Chicago with support from the OCC.

OCC Open Science Data Cloud (2010)

Bionimbus Protected Data Cloud* (2013)

NCI Genomic Data Commons* (2016)

OCC-NOAA Environmental Data Commons (2016)


Kids First Data Resource (2017)

Brain Commons (2017)

*Operated under a subcontract from NCI / Leidos Biomedical to the University of Chicago with support from the OCC.
GDC Data Portal Demo
API URL Structure

- https://api.gdc.cancer.gov/endpoint

The available endpoints and their functionality are listed below:

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>status</td>
<td>Status</td>
<td>Get the API status and version information</td>
</tr>
<tr>
<td>projects</td>
<td>Search &amp; Retrieval</td>
<td>Search all data generated by a project</td>
</tr>
<tr>
<td>cases</td>
<td>Search &amp; Retrieval</td>
<td>Find all files related to a specific case, or sample donor.</td>
</tr>
<tr>
<td>files</td>
<td>Search &amp; Retrieval</td>
<td>Find all files with specific characteristics such as file_name, md5sum, data format and others</td>
</tr>
<tr>
<td>annotations</td>
<td>Search &amp; Retrieval</td>
<td>Search annotations added to data after curation</td>
</tr>
<tr>
<td>data</td>
<td>Download</td>
<td>Used to download GDC data</td>
</tr>
<tr>
<td>manifest</td>
<td>Download</td>
<td>Generates manifests for use with GDC Data Transfer Tool</td>
</tr>
<tr>
<td>slicing</td>
<td>BAM Slicing</td>
<td>Allows remote slicing of BAM format objects</td>
</tr>
<tr>
<td>submission</td>
<td>Submission</td>
<td>Returns the available resources at the top level above programs</td>
</tr>
</tbody>
</table>
The available endpoints and their functionality are listed below:

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>genes</td>
<td>Data Analysis</td>
<td>Allows users to access information about a gene</td>
</tr>
<tr>
<td>ssms</td>
<td>Data Analysis</td>
<td>Allows users to access information about each somatic mutation</td>
</tr>
<tr>
<td>ssm_occurrences</td>
<td>Data Analysis</td>
<td>An ssm from a specific case</td>
</tr>
<tr>
<td>analysis/top_cases_counts_by_genes</td>
<td>Data Analysis</td>
<td>Returns the number of cases with a mutation in each gene listed in the gene_ids parameter for each project</td>
</tr>
<tr>
<td>Analysis/top_mutated_genes_by_product</td>
<td>Data Analysis</td>
<td>Returns a list of genes that have the most mutations within a given project</td>
</tr>
<tr>
<td>analysis/top_mutated_cases_by_gene</td>
<td>Data Analysis</td>
<td>Generates information about the cases that are most affected by mutations in a given number of genes</td>
</tr>
<tr>
<td>analysis/mutated_cases_count_by_product</td>
<td>Data Analysis</td>
<td>Returns counts for the number of cases that have associated ssm data in each project</td>
</tr>
<tr>
<td>analysis_survival</td>
<td>Data Analysis</td>
<td>Survival plots can be generated in the Data Portal for different subsets of data, based upon many query factors such as variants, disease type and projects</td>
</tr>
</tbody>
</table>
Sample Requests

- **Files Endpoint** - https://api.gdc.cancer.gov/files/4f6e2e7a-b617-4444-8a48-83059301f4c3
  - Returns basic information about a specific file

- **Genes Endpoint** - https://api.gdc.cancer.gov/genes/ENSG00000141510
  - Returns basic information about a specific gene

Additional request parameters can be supplied to create more advanced queries.
The filters parameter is a percent encoded JSON that specifies the search terms for the query.

The filters parameter is a percent encoded JSON that specifies the search terms for the query.

```
{  
  "op": "and",  
  "content": [  
    {  
      "op": "in",  
      "content": {  
        "field": "cases.project.program.name",  
        "value": [  
          "TCGA"  
        ]  
      }  
    },  
    {  
      "op": "in",  
      "content": {  
        "field": "cases.project.primary_site",  
        "value": [  
          "Breast"  
        ]  
      }  
    }  
  ]  
}
```

Encode

```
%7B%22op%22:%22and%22,%22content%22:%22%7B%22%5B%7B%22op%22:%22in%22,%22content%22:%7B%22%5D%7D%7D,cases.project.program.name%22,%22value%22:%7B%22%5D%7D%7D,%7B%22op%22:%22in%22,%22content%22:%7B%22%5D%7D%7D,cases.project.primary_site%22,%22value%22:%7B%22%5D%7D%7D
```
API Request Parameters


The filters parameter is a percent encoded JSON that specifies the search terms for the query.

```

{ "op": "and", "content": [
    { "op": "in", "content": { "field": "cases.project.program.name", "value": [ "TCGA" ] } },
    { "op": "in", "content": { "field": "cases.project.primary_site", "value": [ "Breast" ] } }
  ]}

%7B%22op%22:%22and%22,%22content%22:%5B%7B%22%22in%22,%22cases.project.program.name%22:%22TCGA%22%5D%7B%22in%22,%22cases.project.primary_site%22:%22Breast%22%5D%7D

Additional request parameters can be supplied to create more advanced queries.
```
Complex DAVE requests

- Show me the patients with mutations in a specific gene
    {
      "op": "and",
      "content": [
        {
          "op": "in",
          "content": {
            "field": "genes.gene_id",
            "value": ["ENSG00000141510"]
          }
        }
      ]
    }
    &fields=case_id,submitter_id&size=100000&pretty=true
Complex DAVE requests

- Show me survival information for patients with and without mutations in a specific gene.

- Can create queries in other languages of your choice (e.g. python in ipython notebook).
Questions?
Object-based storage with access control lists

Scalable lightweight workflow

Database services

Community data products

Portals for accessing & submitting data

Data Commons Framework Services (Digital ID, Metadata, Authentication, Auth., etc.) that support multiple data commons.

Data Commons 1

Data Commons 2

APIs

Workspaces

Apps & Notebooks

Apps

Notebooks

Workspaces

Workspaces

Apps