NCI DCFS: Metadata API

Monday, August 15, 2022
3:00PM - 4:00PM (CDT)

NCI DCFS: Metadata API

Monday, August 15, 2022
3:00PM - 4:00PM (CDT)

NCI CRDC Framework Services

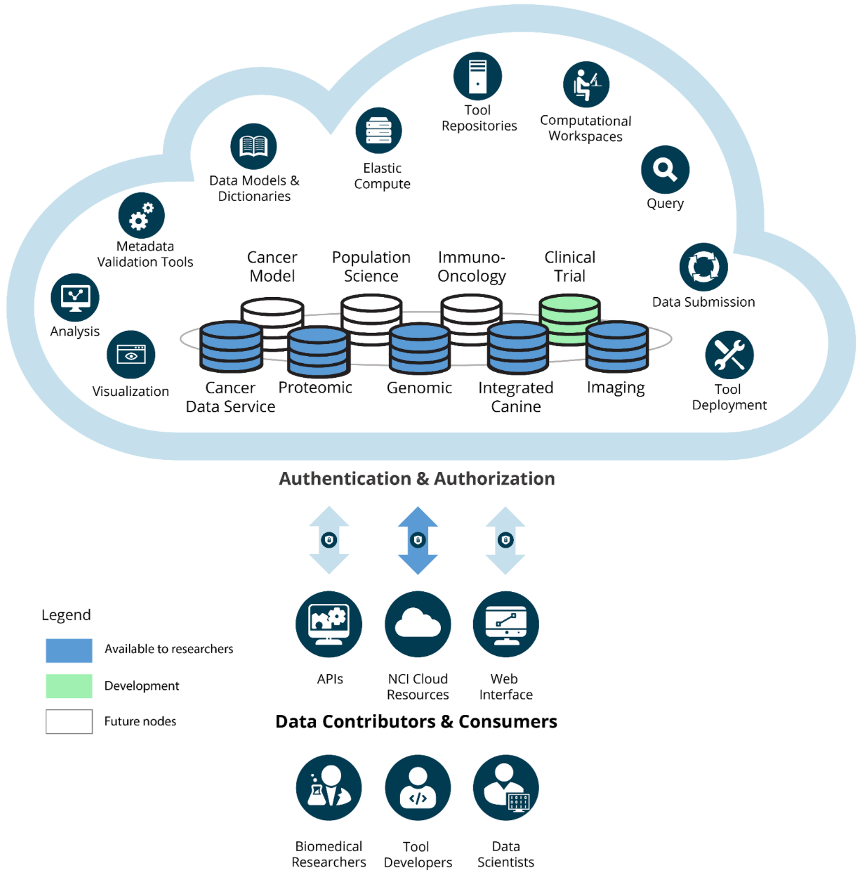
Gen3 Framework Services: Metadata API

Aarti Venkat
Alex VanTol

THE UNIVERSITY OF CHICAGO
CENTER FOR TRANSLATIONAL DATA SCIENCE

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17X053 under Contract No. HHSN261200800001E.

NCI Cancer Research Data Commons (CRDC)



Tool Repositories

Computational Workspaces

Elastic Compute

Data Models & Dictionaries

Query

Metadata Validation Tools

Analysis

Visualization

Cancer Model

Population Science

Immuno-Oncology

Clinical Trial

Data Submission

Cancer Data Service

Proteomic

Genomic

Integrated Canine

Imaging

Tool Deployment

Authentication & Authorization

APIs

NCI Cloud Resources

Web Interface

Data Contributors & Consumers

Biomedical Researchers

Tool Developers

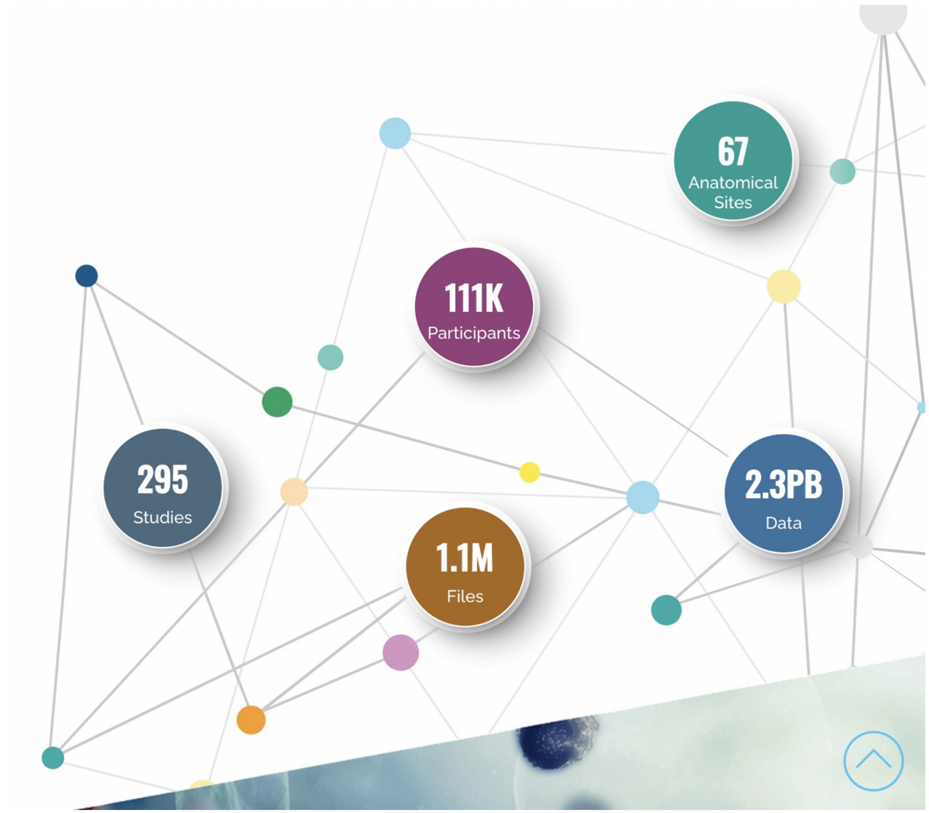
Data Scientists

Legend

🟦 Available to researchers

🟩 Development

□ Future nodes



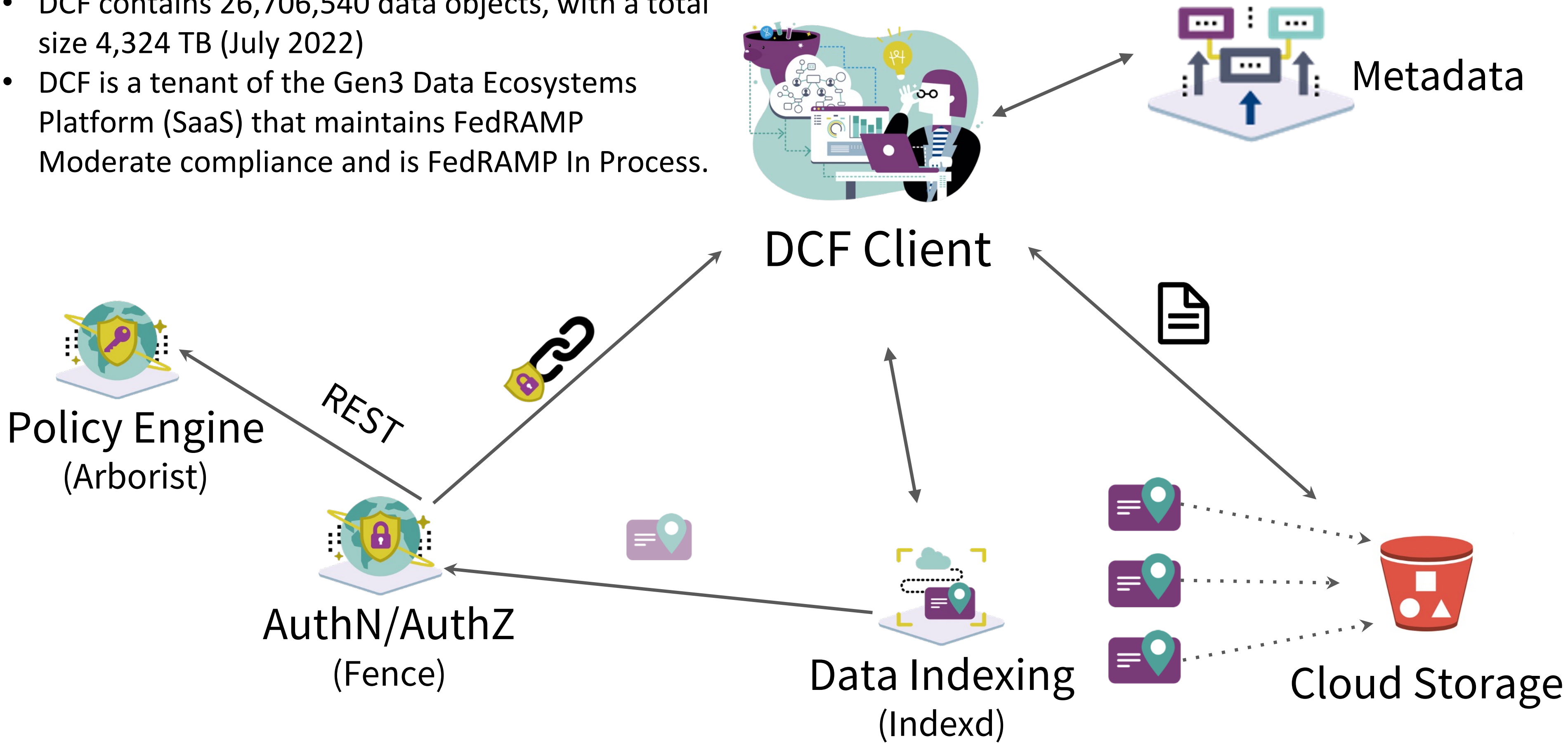
Images from https://datascience.cancer.gov/data-commons

Introduction

- Gen3 is a platform for building data commons, data meshes, and workspaces.
- NCI Data Commons Framework Services (DCFS) runs an instance of Gen3 Framework Services.
- Gen3 is built and maintained by the Center for Translational Data Science at the University of Chicago.

Framework Services Architecture

- DCF contains 26,706,540 data objects, with a total size 4,324 TB (July 2022)
- DCF is a tenant of the Gen3 Data Ecosystems Platform (SaaS) that maintains FedRAMP Moderate compliance and is FedRAMP In Process.



DCF Client

Metadata

Policy Engine
(Arborist)

REST

AuthN/AuthZ
(Fence)

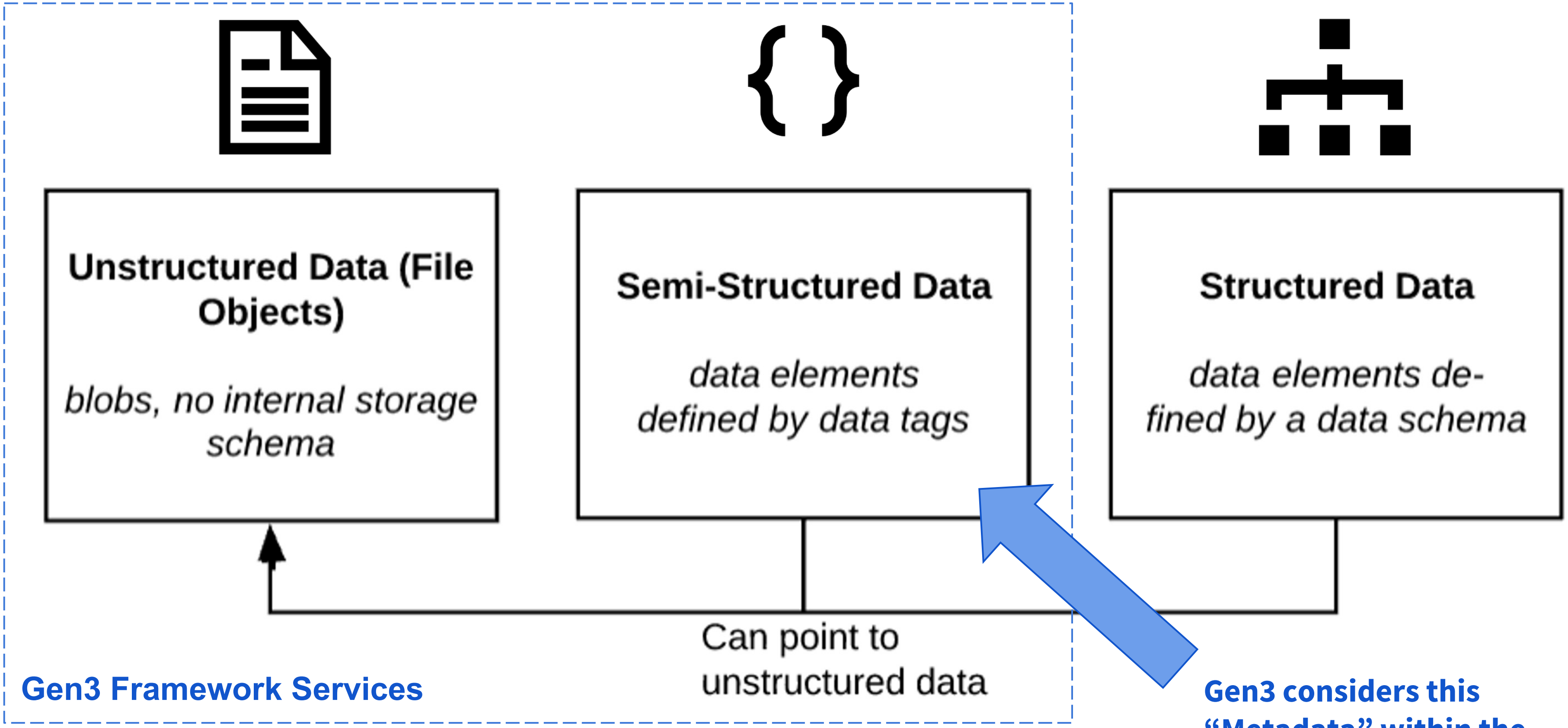
Data Indexing
(Indexd)

Cloud Storage

The Agenda

- Types of Data Gen3 Manages
- Gen3 Metadata Overview
- FAIR Overview
- Use Cases
- Data Commons vs Data Ecosystems
- Data Ingestion into Gen3 Framework Services

Types of Data Gen3 Manages



Unstructured Data (File Objects)

blobs, no internal storage schema

Semi-Structured Data

data elements defined by data tags

Structured Data

data elements defined by a data schema

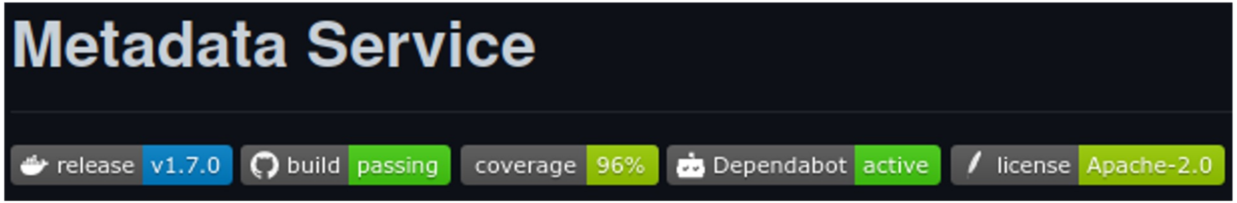
Can point to unstructured data

Gen3 Framework Services

Gen3 considers this “Metadata” within the Framework Services

Gen3 Metadata API

- Part of the Gen3 Framework Services
- Service powers an API to query and retrieve semi-structured data
- Stores data as schema-less JavaScript Object Notation (JSON) blobs attached to globally unique identifiers (GUIDs)
    - GUIDs may be indexed unstructured (file object) GUIDs or non-file-based GUIDs (such as subject or dataset identifiers)



Metadata Service

Gen3 Metadata API

- Gen3 Framework Services provide an open API that allows clients to query and retrieve schema-less JSON blobs associated with GUIDs

Before

- Indexing API (minting persistent identifiers)
  - File size
  - Checksum
  - URLs/locations
  - Data objects are not FAIR

New

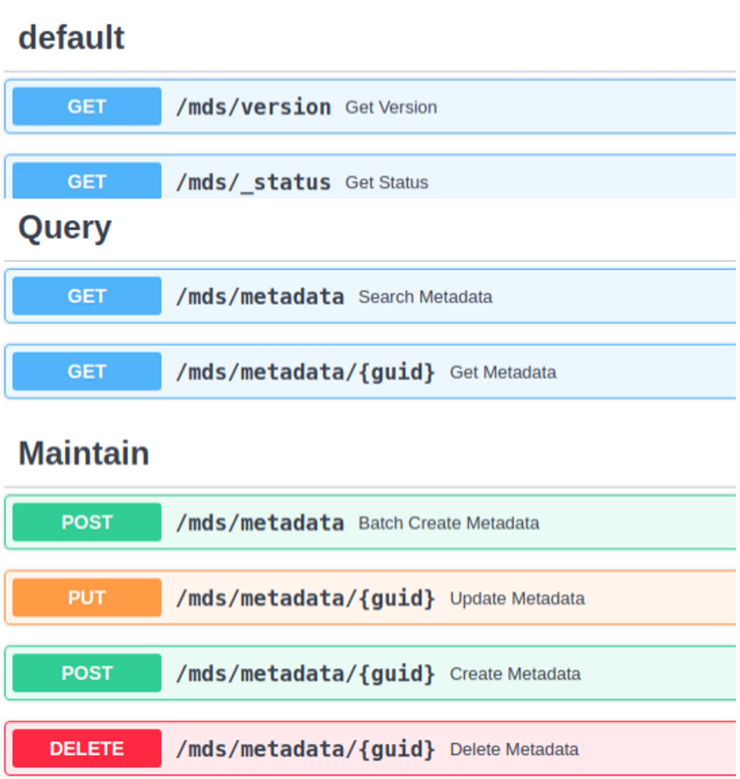
- Metadata API
  - Other arbitrary metadata
  - Schema-less
  - FAIR data objects easily supported
  - **Publicly available** metadata
  - Ideally metadata is available from a stable API or location

Use Cases: Additional Sample-Level Metadata, Study-level Metadata, Subject-level metadata, Subject-level identifier mappings (crosswalks)

Application Programing Interface (API)

GEN3

```
{
 "_guid_type": "indexed_file_object",
 "dbgap": {
  "submitted_sample_id": "93227",
  "consent_code": "1",
  "biosample_id": "SAMN08666480",
  "dbgap_sample_id": "2957086",
  "sra_sample_id": "SRS3389514",
  "submitted_subject_id": "93227",
  "study_subject_id": "phs001554.v1_93227"
  "dbgap_subject_id": "2474022",
  "consent_short_name": "GRU",
  "sex": "female",
  "analyte_type": "DNA",
  "sample_use": ["Seq_DNA_SNP_CNV", "WGS"],
  "repository": "NCI_CRC_Susceptibility",
...
  },
  "study": "phs001554",
  "study_with_consent": "phs001554.c1",
  "study_accession": "phs001554.v1.p1",
  "study_accession_with_consent": "phs001554.v1.p1.c1",
 },
 "{{non dbgap data source}}": {
  "key": "value",
 }
}
```



default

GET `/mds/version` Get Version

GET `/mds/_status` Get Status

Query

GET `/mds/metadata` Search Metadata

GET `/mds/metadata/{guid}` Get Metadata

Maintain

POST `/mds/metadata` Batch Create Metadata

PUT `/mds/metadata/{guid}` Update Metadata

POST `/mds/metadata/{guid}` Create Metadata

DELETE `/mds/metadata/{guid}` Delete Metadata

Findable, Accessible, Interoperable & Reusable (FAIR)

GEN3

F	A	I	R
Findable	Accessible	Interoperable	Reusable

- FAIR F2 requires data to be described with rich metadata
- FAIR F3 requires that metadata clearly and explicitly include the identifier of the data it describes
- FAIR F4 requires metadata are registered in a searchable resources

The Gen3 Framework Services (Metadata + Indexing APIs) satisfies F2-F4 with persistent identifiers and rich metadata.

Source: Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3, no. 1 (2016): 1-9.

Gen3 Metadata API

GEN3

- What use case(s) drove the initial development?
  - Data is not FAIR without metadata
  - There needs to be a common source of truth for metadata in a data mesh with multiple computational resources

F A I R

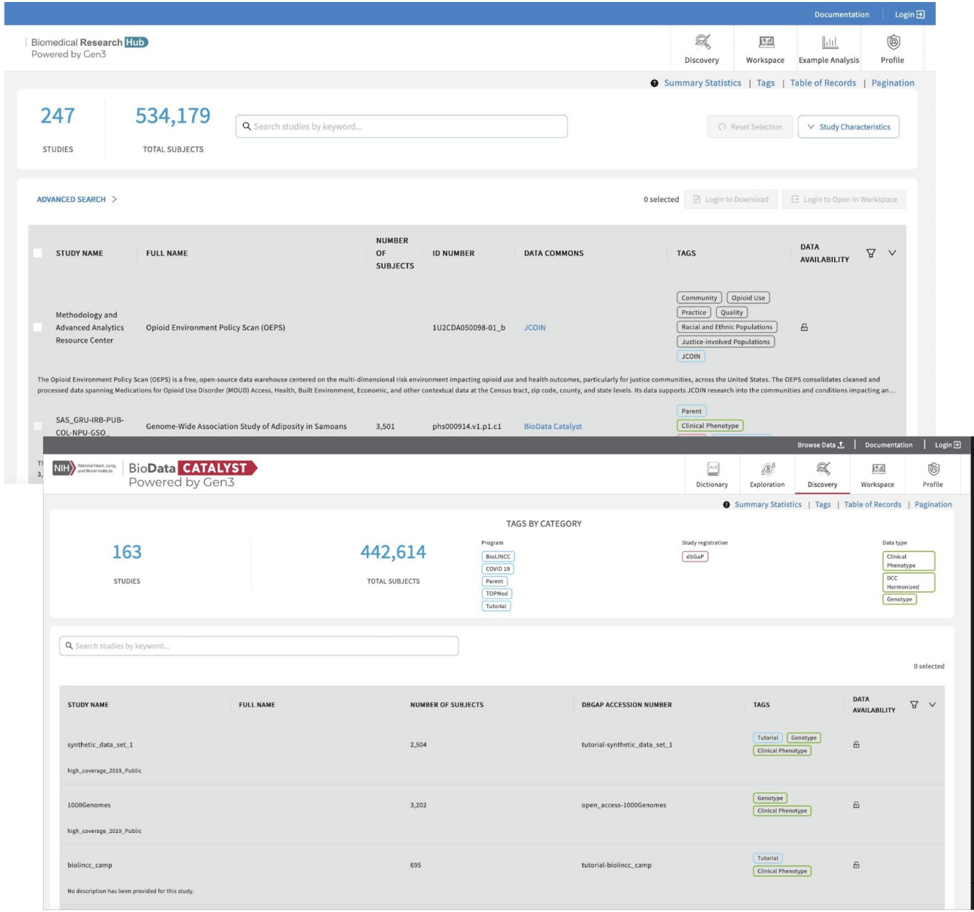
- Who uses the Gen3 Metadata API?
  - BioData Catalyst, MIDRC, Biomedical Research Hub, HEAL Data Platform, …

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

MIDRC MEDICAL IMAGING AND DATA RESOURCE CENTER

Examples of Client Use Cases

- Clients are able to obtain additional metadata for indexed GUIDs from a dynamic, scalable API instead of a static file
- Clients are able to obtain study-level metadata to understand what datasets are available in a commons or mesh
- Clients can obtain metadata about subjects, patients or participants
- Clients can use a crosswalk for privacy preserving record linkage across cloud platforms (in progress)



247 STUDIES

534,179 TOTAL SUBJECTS

	STUDY NAME	FULL NAME	NUMBER OF SUBJECTS	ID NUMBER	DATA COMMONS	TAGS	DATA AVAILABILITY
□	Methodology and Advanced Analytics Resource Center	Opioid Environment Policy Scan (OEPS)		1U2CDA050098-01_b	JCOIN	Community Opioid Use Practice Quality Racial and Ethnic Populations Justice-involved Populations JCOIN	🔓
□	SAS_GRU-IRB-PUB-COL-NPU-GSO_	Genome-Wide Association Study of Adiposity in Samoans	3,501	phs000914.v1.p1.c1	BioData Catalyst	Parent Clinical Phenotype	

The Opioid Environment Policy Scan (OEPS) is a free, open-source data warehouse centered on the multi-dimensional risk environment impacting opioid use and health outcomes, particularly for justice communities, across the United States. The OEPS consolidates cleaned and processed data spanning Medications for Opioid Use Disorder (MOUD) Access, Health, Built Environment, Economic, and other contextual data at the Census tract, zip code, county, and state levels. Its data supports JCOIN research into the communities and conditions impacting an...

163 STUDIES

442,614 TOTAL SUBJECTS

TAGS BY CATEGORY

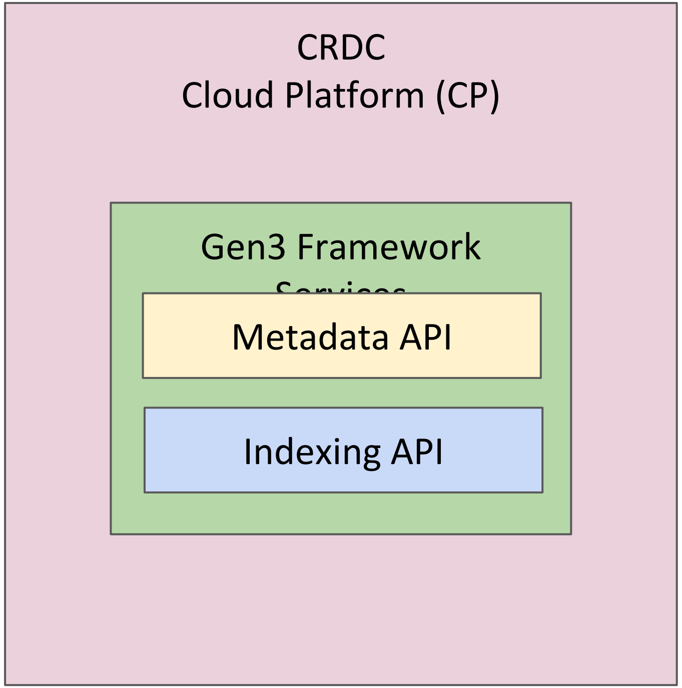
Program: BioLINCC, COVID 19, Parent, TOPMed, Tutorial

Study registration: dbGaP

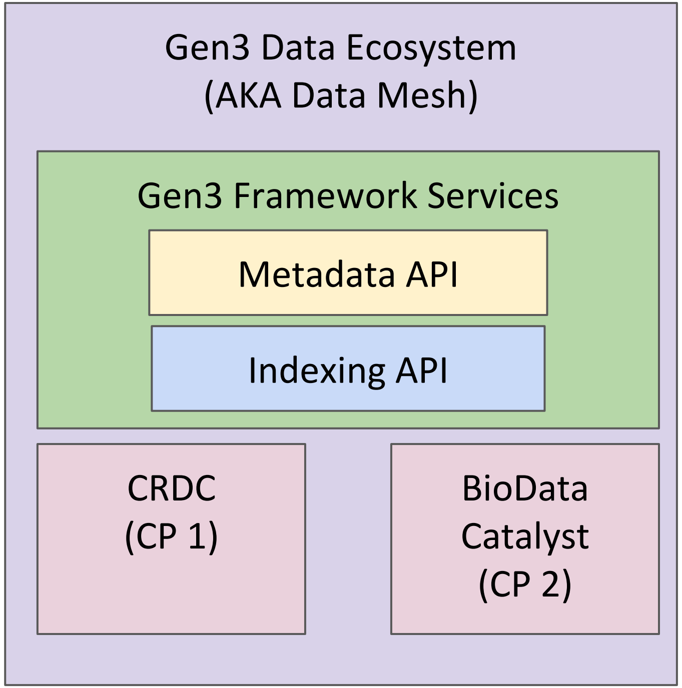
Data type: Clinical Phenotype, DCC Harmonized, Genotype

STUDY NAME	FULL NAME	NUMBER OF SUBJECTS	DBGAP ACCESSION NUMBER	TAGS	DATA AVAILABILITY
synthetic_data_set_1		2,504	tutorial-synthetic_data_set_1	Tutorial Genotype Clinical Phenotype	🔓
high_coverage_2019_Public					
1000Genomes		3,202	open_access-1000Genomes	Genotype Clinical Phenotype	🔓
high_coverage_2019_Public					
biolincc_camp		695	tutorial-biolincc_camp	Tutorial Clinical Phenotype	🔓
No description has been provided for this study.					

Framework Services vs Data Ecosystem

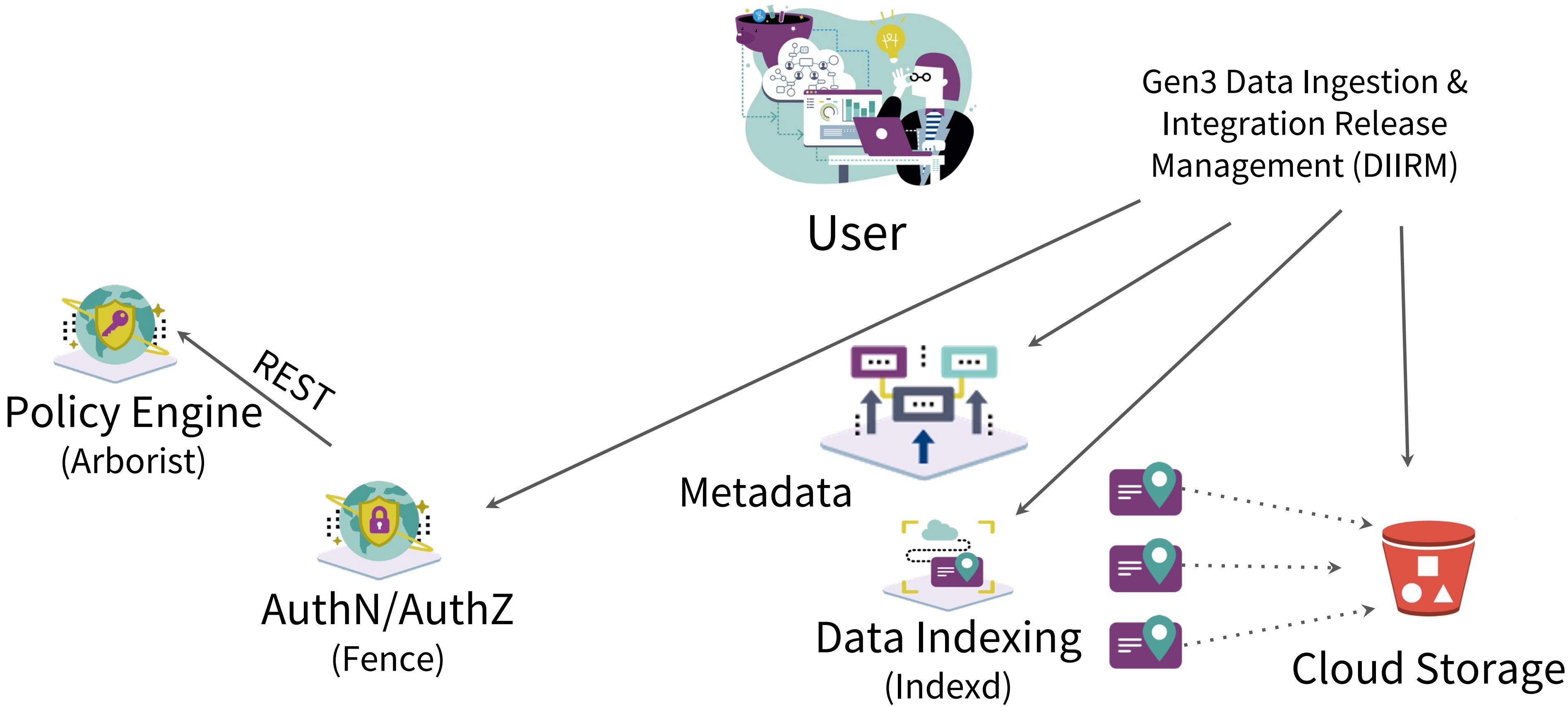


Option A. Set up indexing and metadata for a cloud platform, e.g. CRDC



Option B. Set up indexing and metadata across two or more cloud platforms (CPs)

Framework Services Architecture



User

Gen3 Data Ingestion & Integration Release Management (DIIRM)

Policy Engine
(Arborist)

REST

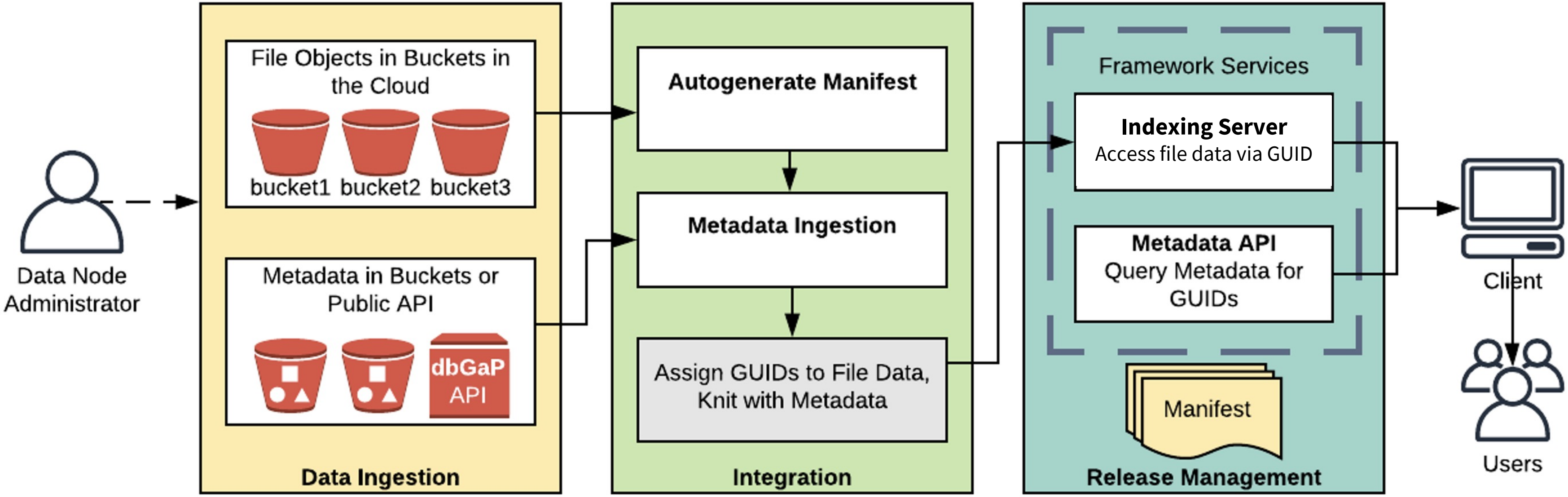
AuthN/AuthZ
(Fence)

Metadata

Data Indexing
(Indexd)

Cloud Storage

Gen3 Data Ingestion & Integration Release Management (DIIRM)



Q&A