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)
Gen3 Framework Services: Metadata API

Aarti Venkat
Alex VanTol

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.
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.
• 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.
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

Gen3 Framework Services

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

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)
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)

```
{
    "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",
    ...
}
```
Findable, Accessible, Interoperable & Reusable (FAIR) (GEN3)

- 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.

Gen3 Metadata API

● 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

● Who uses the Gen3 Metadata API?
  ○ BioData Catalyst, MIDRC, Biomedical Research Hub, HEAL Data Platform, …
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).
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

- **Policy Engine (Arborist)**
- **AuthN/AuthZ (Fence)**
- **Metadata**
- **Data Indexing (Indexd)**
- **Cloud Storage**
- **User**
- **Gen3 Data Ingestion & Integration Release Management (DIIRM)**

Diagram showing the integration of services for data management and security.
Gen3 Data Ingestion & Integration Release Management (DIIRM)

Data Ingestion
- File Objects in Buckets in the Cloud
- Metadata in Buckets or Public API

Integration
- Autogenerate Manifest
- Metadata Ingestion
- Assign GUIDs to File Data, Knit with Metadata

Release Management
- Framework Services
  - Indexing Server
    - Access file data via GUID
  - Metadata API
    - Query Metadata for GUIDs

Client
Users