Gen3 Webinar

NCI DCFS

Interoperability with Framework Services

Tuesday, June 9, 2020
2:00 PM - 3:00 PM (CST)

Interoperability with Framework Services

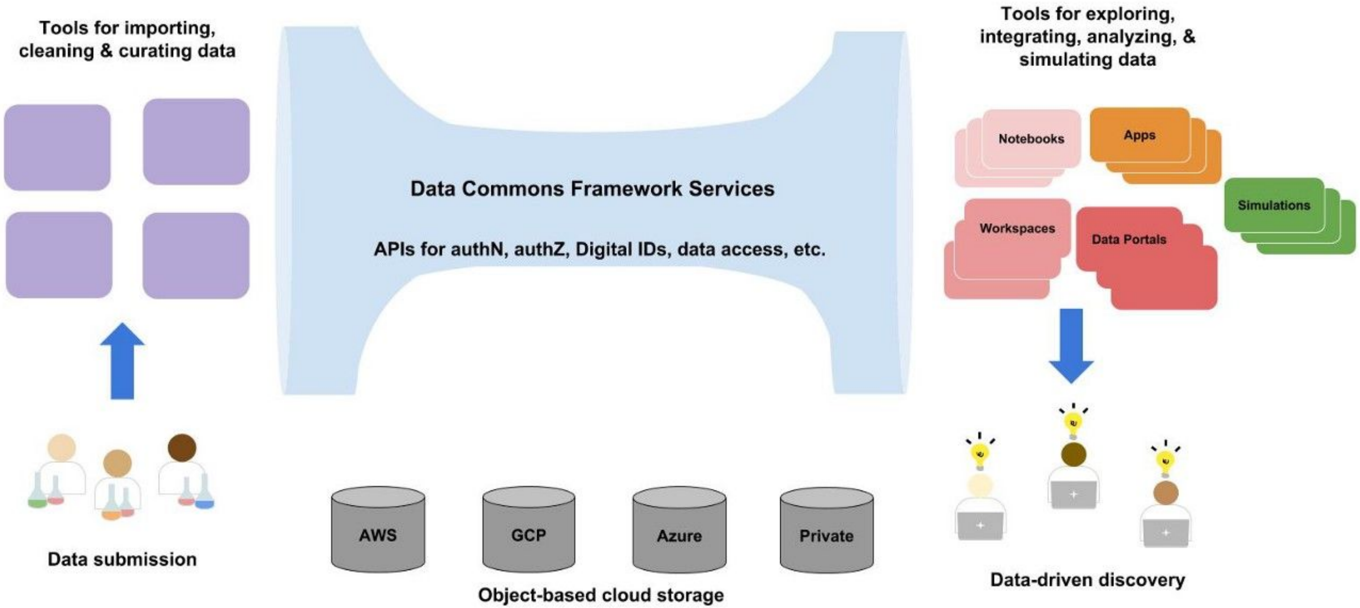
Chris Meyer
Jiaqi Liu
Alexander VanTol

Center for Translational Data Science,
University of Chicago

GEN3 Data Commons

*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. Source of image: The CDIS, GDC, & OCC data commons infrastructure at the University of Chicago Kenwood Data Center.


Tools for importing, cleaning & curating data
Data Commons Framework Services
APIs for authN, authZ, Digital IDs, data access, etc.
Tools for exploring, integrating, analyzing, & simulating data
Notebooks
Apps
Simulations
Workspaces
Data Portals
Data submission
AWS
GCP
Azure
Private
Object-based cloud storage
Data-driven discovery

*The Framework Services is a set of **interoperable** software services with public APIs that enable data commons and compute environments to receive, manage and share structured clinical data and object data in a secure and scalable way.*

Source of image: Saltzer, Jerome H., David P. Reed, and David D. Clark. “End-to-end arguments in system design,” ACM Transactions on Computer Systems (TOCS) 2, no. 4 (1984): 277–288.

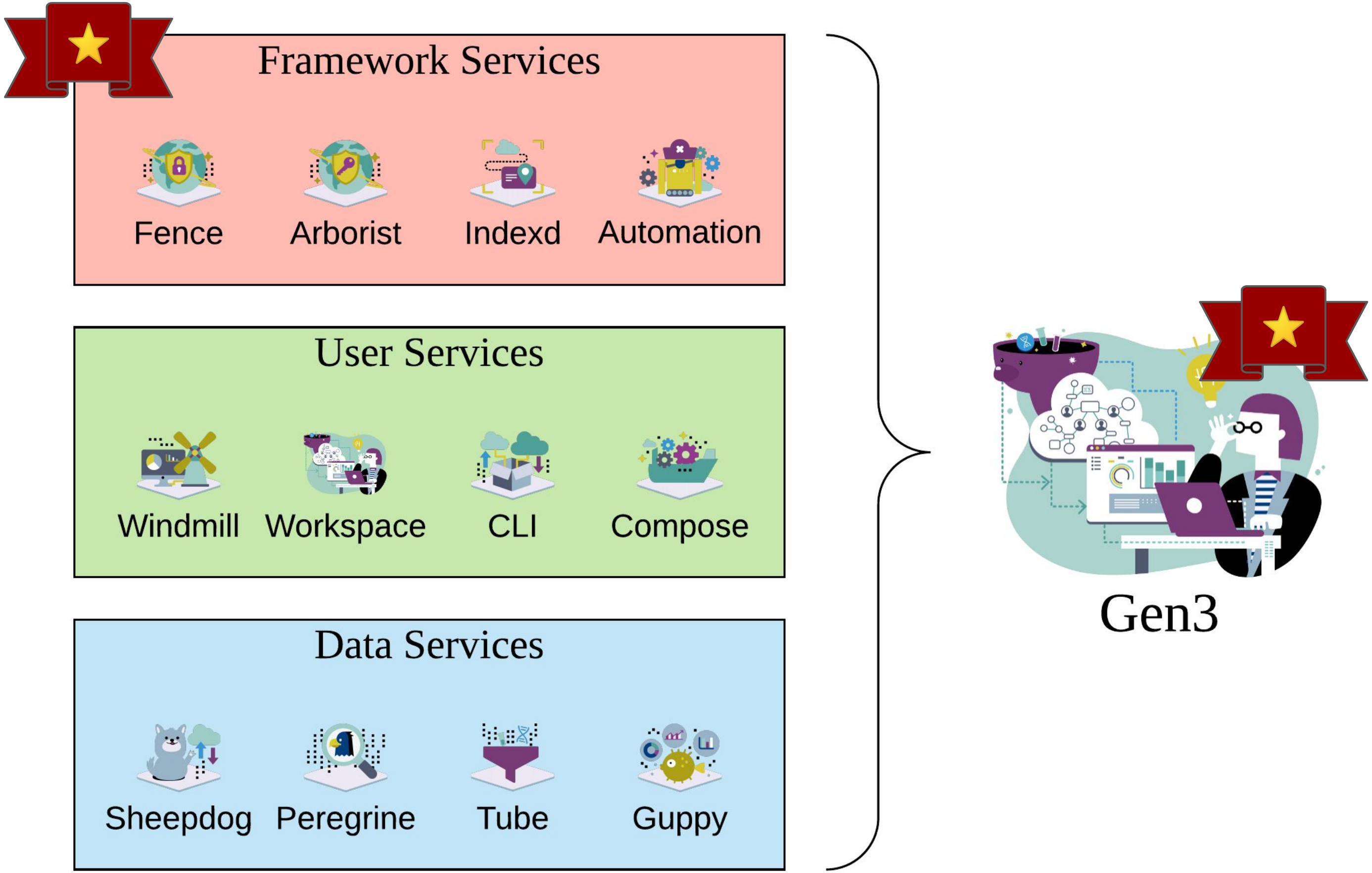
Technical Principles for Interoperability

1. Identify Data through persistent Digital IDs that remain unchanged regardless of the physical location of your data
2. Expose data through an API
3. Expose the data model through an API
4. Interoperate with third party authN and authZ services from trusted platforms
5. Interoperate with other trusted resources with similar security and compliance.

Gen3 Implementation of Framework Services

Services & Core Products

GEN3
DATA COMMONS



Framework Services

Fence Arborist Indexd Automation

User Services

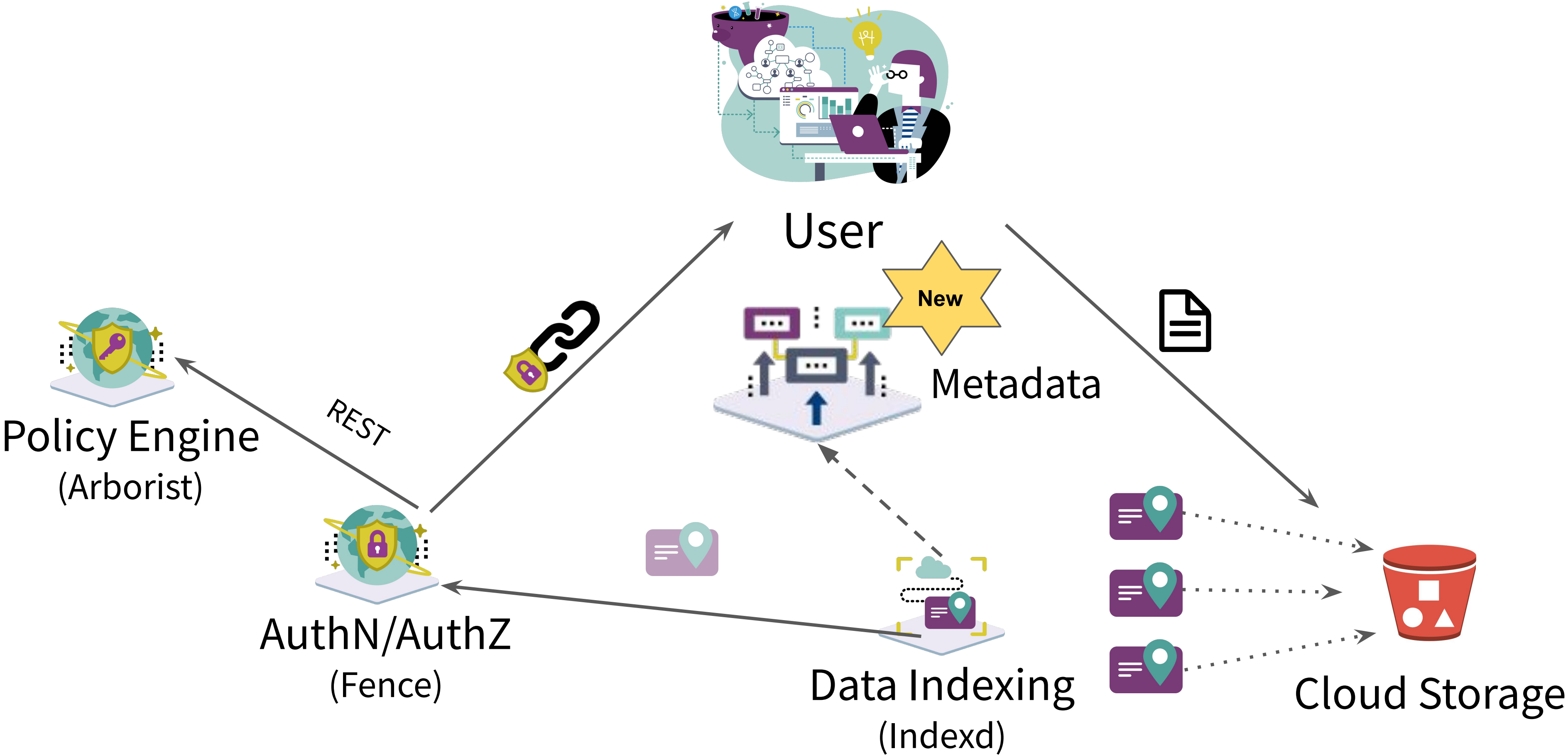
Windmill Workspace CLI Compose

Data Services

Sheepdog Peregrine Tube Guppy

Gen3

Framework Services Architecture



Standards Used by NCI DCFS

Where do Standards Come From

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

OpenID

OAUTH 2 OAUTH

GA4GH Data Repository Service (DRS)

Indexd

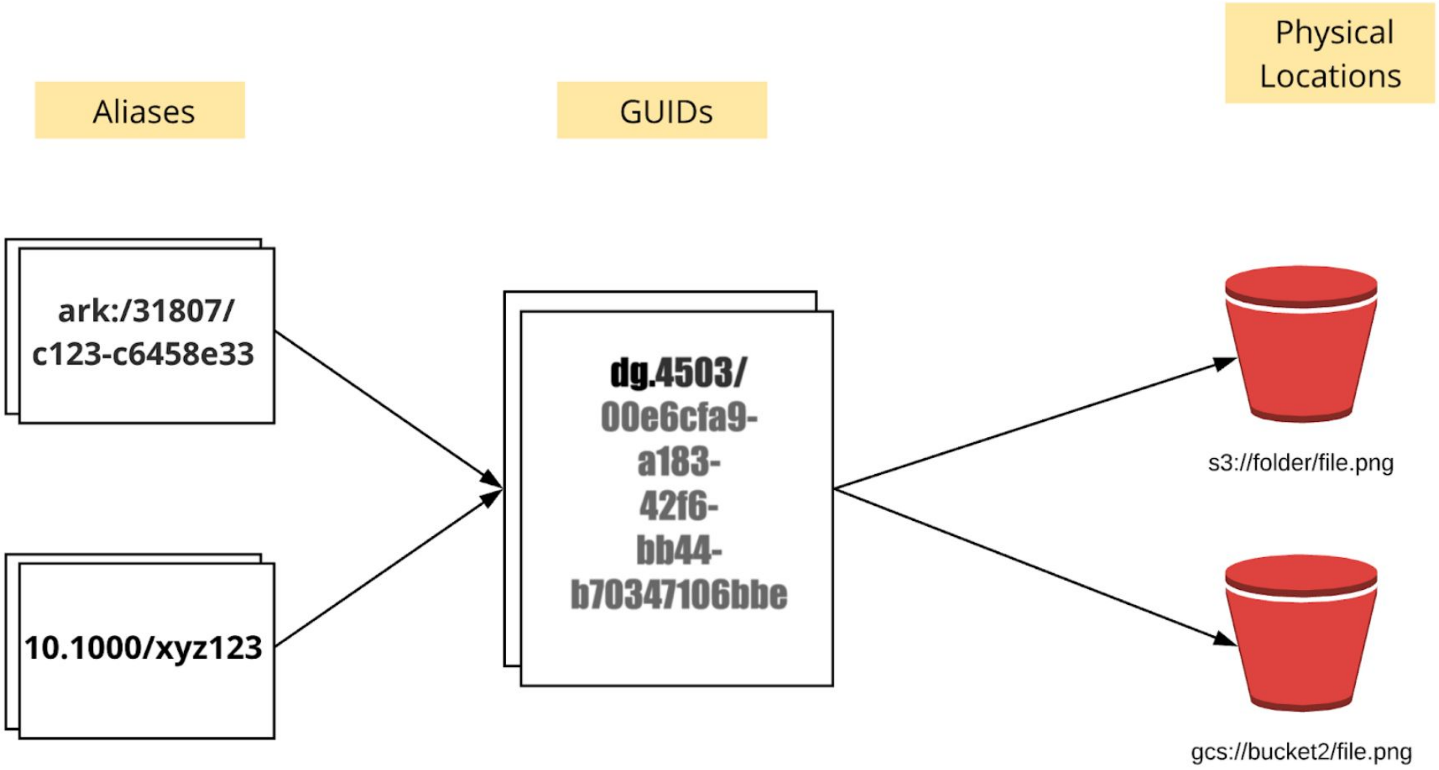
Gen3 data indexing service

Indexd

Gen3 data **indexing** service

indexing: locate data with easily used identifiers

Gen3 Implementation in Indexd



Aliases

GUIDs

Physical Locations

ark:/31807/
c123-c6458e33

10.1000/xyz123

dg.4503/
00e6cfa9-
a183-
42f6-
bb44-
b70347106bbe

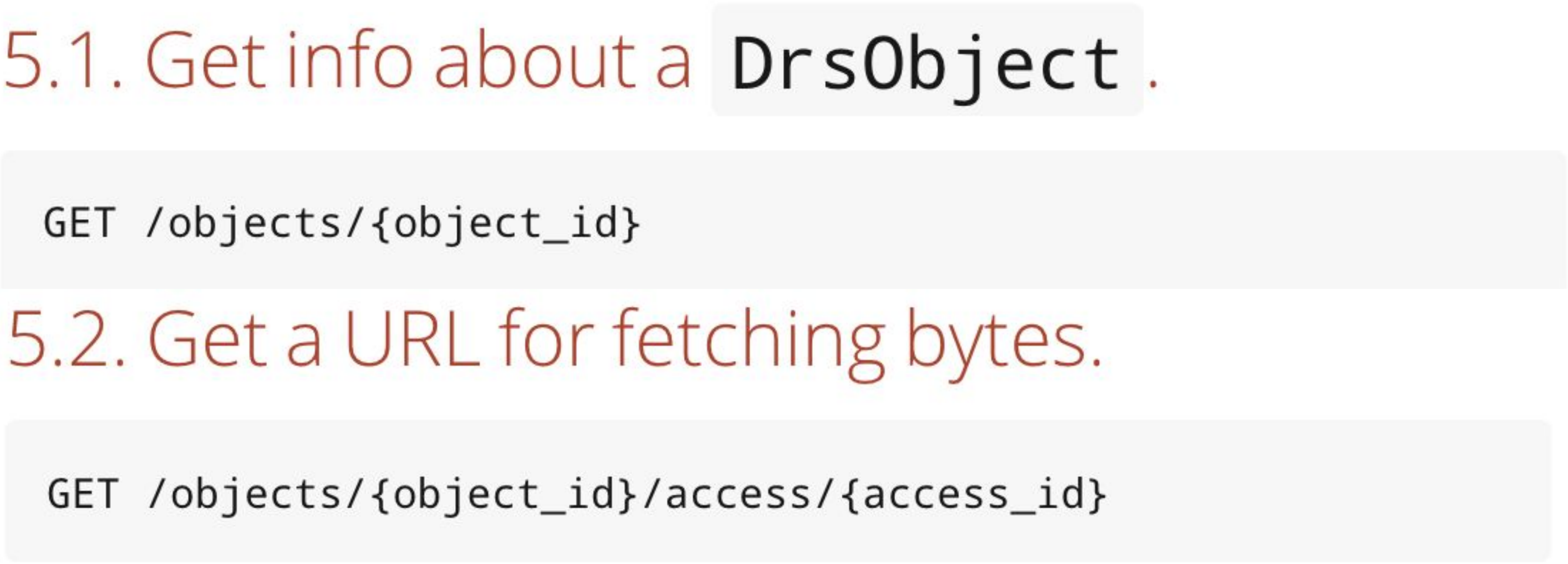
s3://folder/file.png

gcs://bucket2/file.png

Support New DRS endpoints in Indexd

BasePath : `/ga4gh/drs/v1`

Schemes : HTTPS



5.1. Get info about a `DrsObject`.

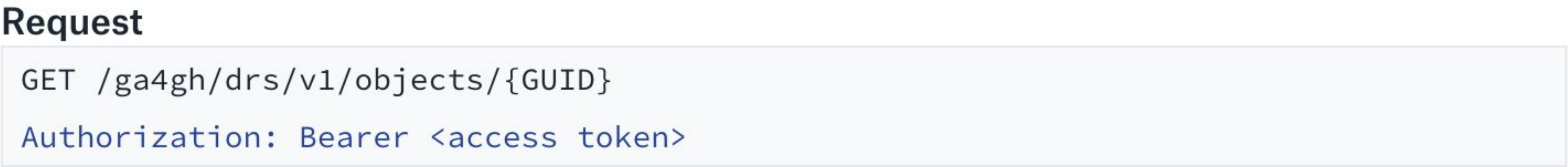
```
GET /objects/{object_id}
```

5.2. Get a URL for fetching bytes.

```
GET /objects/{object_id}/access/{access_id}
```

Fetching Data using Signed URLs

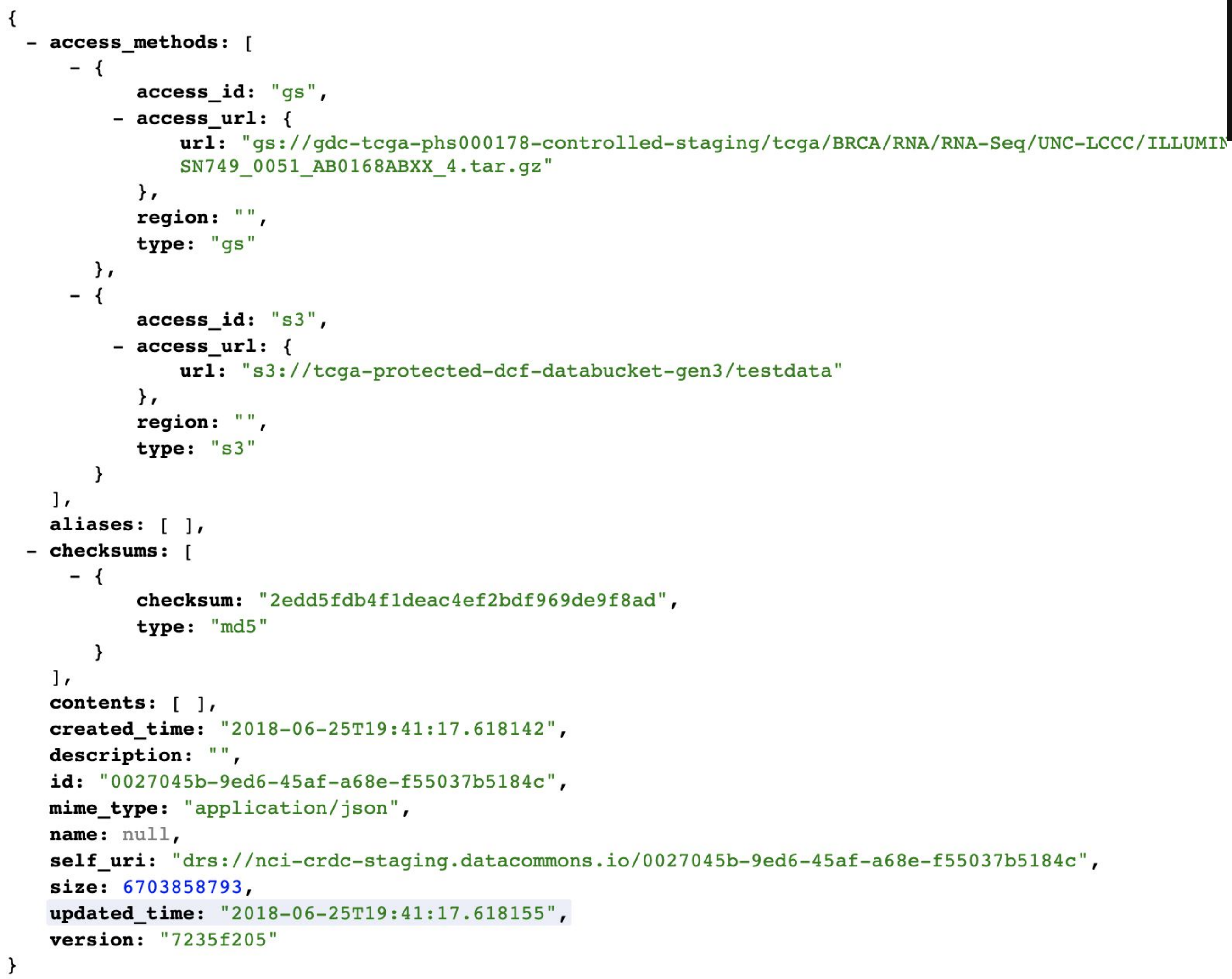
- Indexd will location of file object with additional file metadata in the `/objects/{object_id}` endpoint (open access)
- For signed URLs:
  - Users will get an OAuth2.0 access token from Fence
  - Users will auth with an OAuth2.0 access token in the header
  - Indexd will return a signed URL in `/object/{object_id}/access/{access_id}` with proper authorization
  - If user is not authorized to access data, Indexd will return access denied



Request

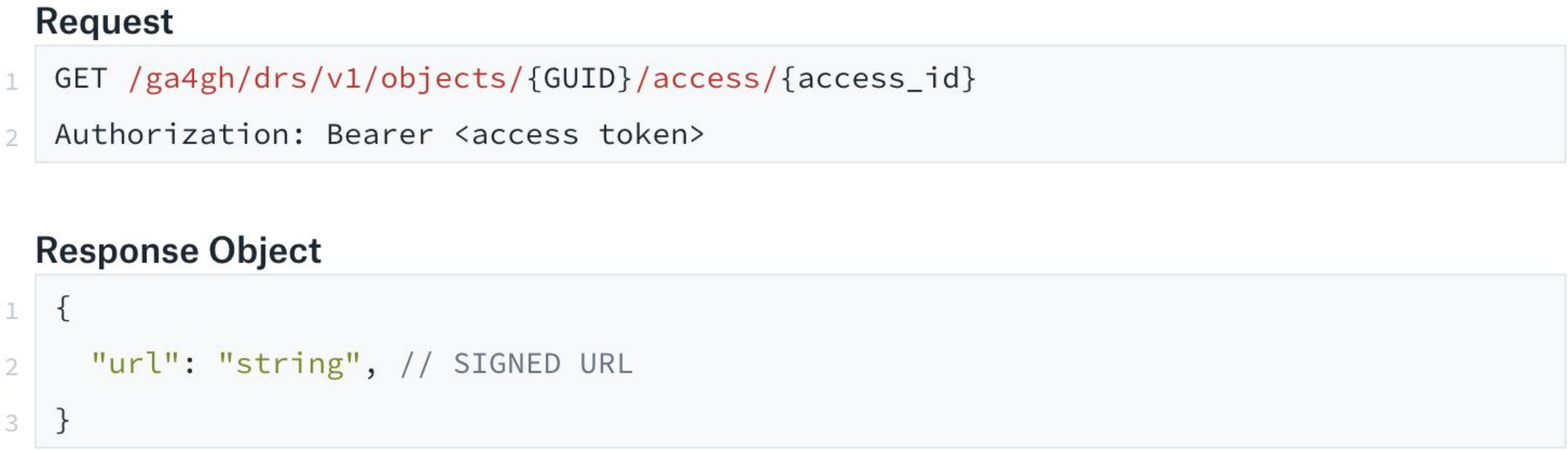
```
GET /ga4gh/drs/v1/objects/{GUID}
Authorization: Bearer <access token>
```

Example DRS Response for Single File Object (DRSObject)



```
{
  - access_methods: [
      - {
             access_id: "gs",
           - access_url: {
                 url: "gs://gdc-tcga-phs000178-controlled-staging/tcga/BRCA/RNA/RNA-Seq/UNC-LCCC/ILLUMIN
                 SN749_0051_AB0168ABXX_4.tar.gz"
             },
             region: "",
             type: "gs"
        },
      - {
             access_id: "s3",
           - access_url: {
                 url: "s3://tcga-protected-dcf-databucket-gen3/testdata"
             },
             region: "",
             type: "s3"
        }
    ],
    aliases: [ ],
  - checksums: [
      - {
             checksum: "2edd5fdb4f1deac4ef2bdf969de9f8ad",
             type: "md5"
        }
    ],
    contents: [ ],
    created_time: "2018-06-25T19:41:17.618142",
    description: "",
    id: "0027045b-9ed6-45af-a68e-f55037b5184c",
    mime_type: "application/json",
    name: null,
    self_uri: "drs://nci-crdc-staging.datacommons.io/0027045b-9ed6-45af-a68e-f55037b5184c",
    size: 6703858793,
    updated_time: "2018-06-25T19:41:17.618155",
    version: "7235f205"
}
```

Example Requests: Access Endpoint (Signed URLs)



Request

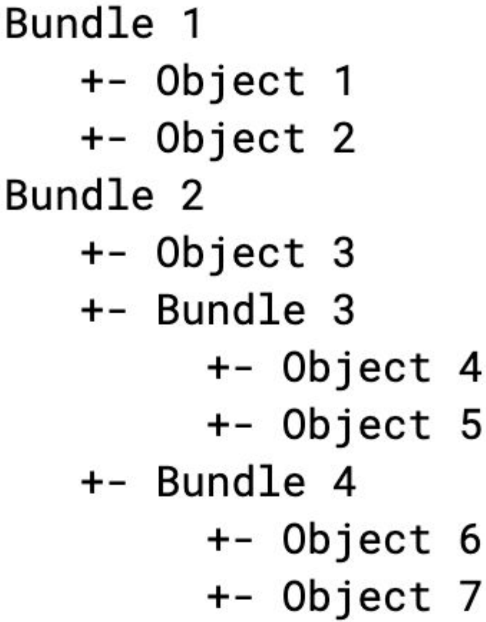
```
GET /ga4gh/drs/v1/objects/{GUID}/access/{access_id}
Authorization: Bearer <access token>
```

Response Object

```
{
  "url": "string", // SIGNED URL
}
```

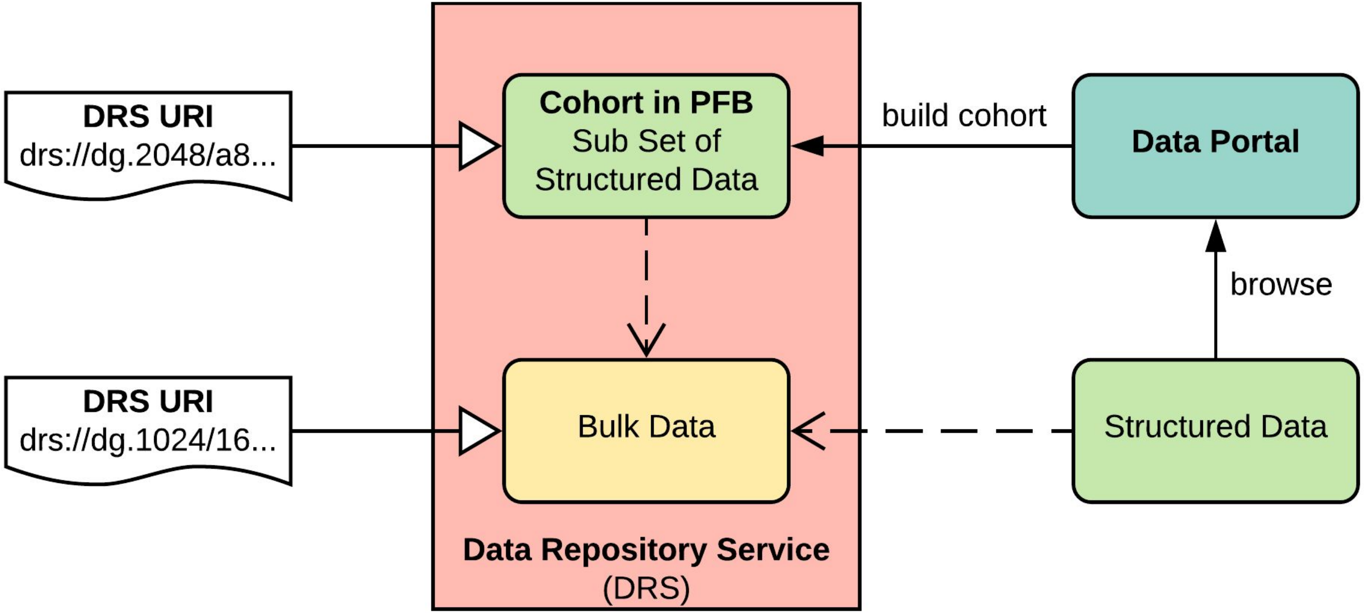
Coming Soon: Bundles

- A Data Bundle is like a folder - contains a collection of data objects (can also contain other bundles)
- Support Bundles as new object type in Indexd
- Support expansion of Bundles in ContentObjects array per DRS spec



```
Bundle 1
    +- Object 1
    +- Object 2
Bundle 2
    +- Object 3
    +- Bundle 3
           +- Object 4
           +- Object 5
    +- Bundle 4
           +- Object 6
           +- Object 7
```

Future DRS usage: Interoperating with Clinical Data



GA4GH Passports & Visas

Passports & Visas

What is a Passport?

- An identity that travels with the researcher across data platforms
- A collection of visas

What is a Visa?

- An assertion signed by a visa issuer
- Designed for machine interpretation only

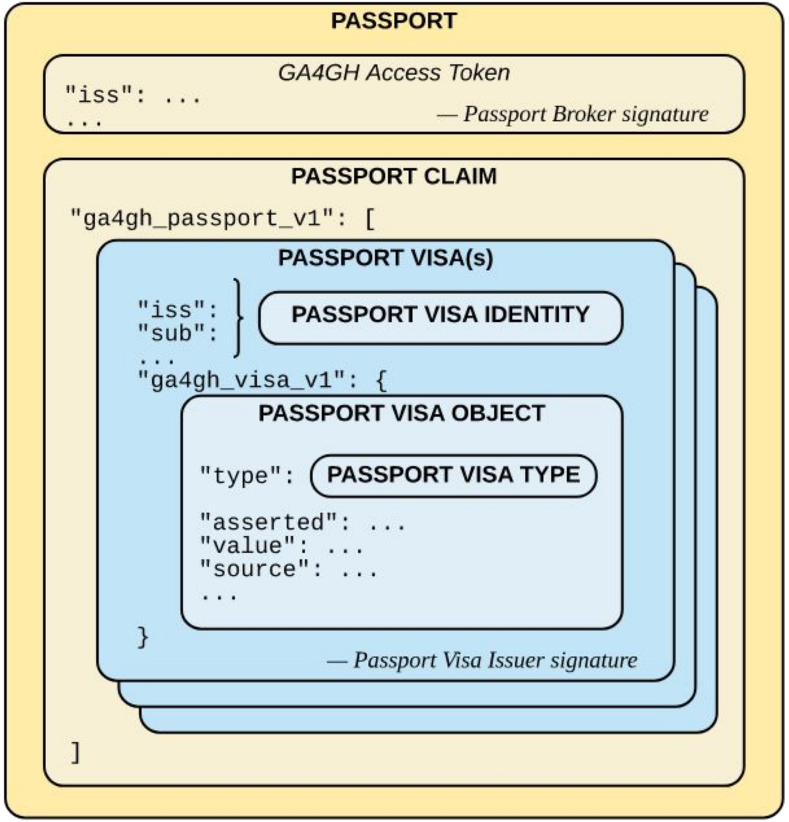
Behind the Curtain: JWTs

GEN3 DATA COMMONS

- Cryptographically signed by fence
  - Use tokens for authentication
  - Any service can verify that a token was issued by the fence instance it expects
- Contains user information
  - User tokens for authorization
- Open source libraries for working with JWTs
  - jwt.io for list of all libraries
  - We use:
    - github.com/mpdavis/python-jose
    - github.com/jpadilla/pyjwt

```
{
  "sub": "7",
  "azp": "test-client",
  "pur": "access",
  "aud": ["openid", "user"],
  "context": {
    "user": {
      "is_admin": false,
      "name": "test",
      "projects": {
        "test": ["read", "create", "upload"]
      }
  },
  "iss": "https://portal.occ-data.org/",
  "jti": "2e6ade06-5afb-4ce7-9ab5-e206225ce291",
  "exp": 1516983302,
  "iat": 1516982102
}
```

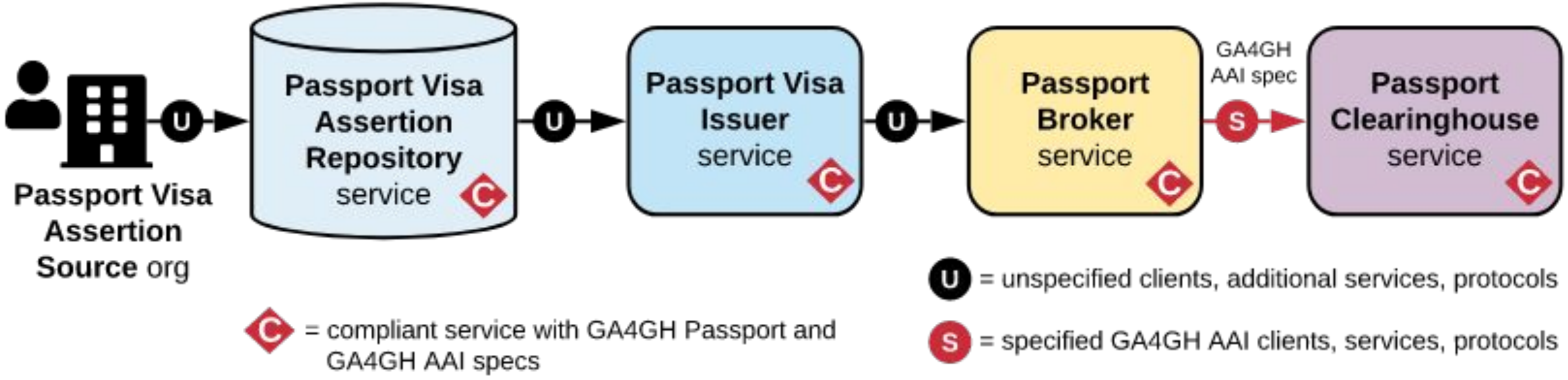
GA4GH Passport



Source of Image: GA4GH DURI Passport overview

Passports / Visas & Fence

GEN3 DATA COMMONS


Passport Visa Assertion Source org
U
Passport Visa Assertion Repository service
C
U
Passport Visa Issuer service
C
U
Passport Broker service
C
GA4GH AAI spec
S
Passport Clearinghouse service
C
U = unspecified clients, additional services, protocols
C = compliant service with GA4GH Passport and GA4GH AAI specs
S = specified GA4GH AAI clients, services, protocols

When Interoperating with Visa issuers to compile information about a user's access, Fence will be a Passport **Broker**

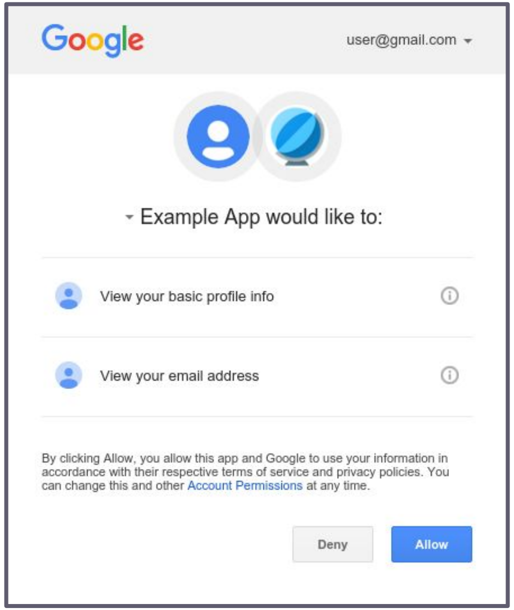
By interpreting and enforcing the authz information in Visas, Fence will act as a Passport **Clearinghouse**

Source of Image: GA4GH DURI Passport overview

OIDC & OAuth 2.0

What is OAuth2?

OAuth2 is a protocol allowing an application to securely access a resource on behalf of a user

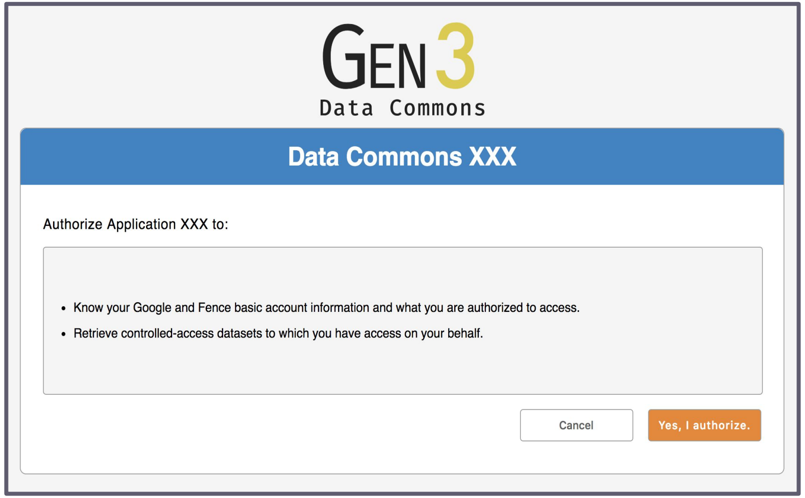


user@gmail.com

Example App would like to:

- View your basic profile info
- View your email address

By clicking Allow, you allow this app and Google to use your information in accordance with their respective terms of service and privacy policies. You can change this and other Account Permissions at any time.



Data Commons XXX

Authorize Application XXX to:

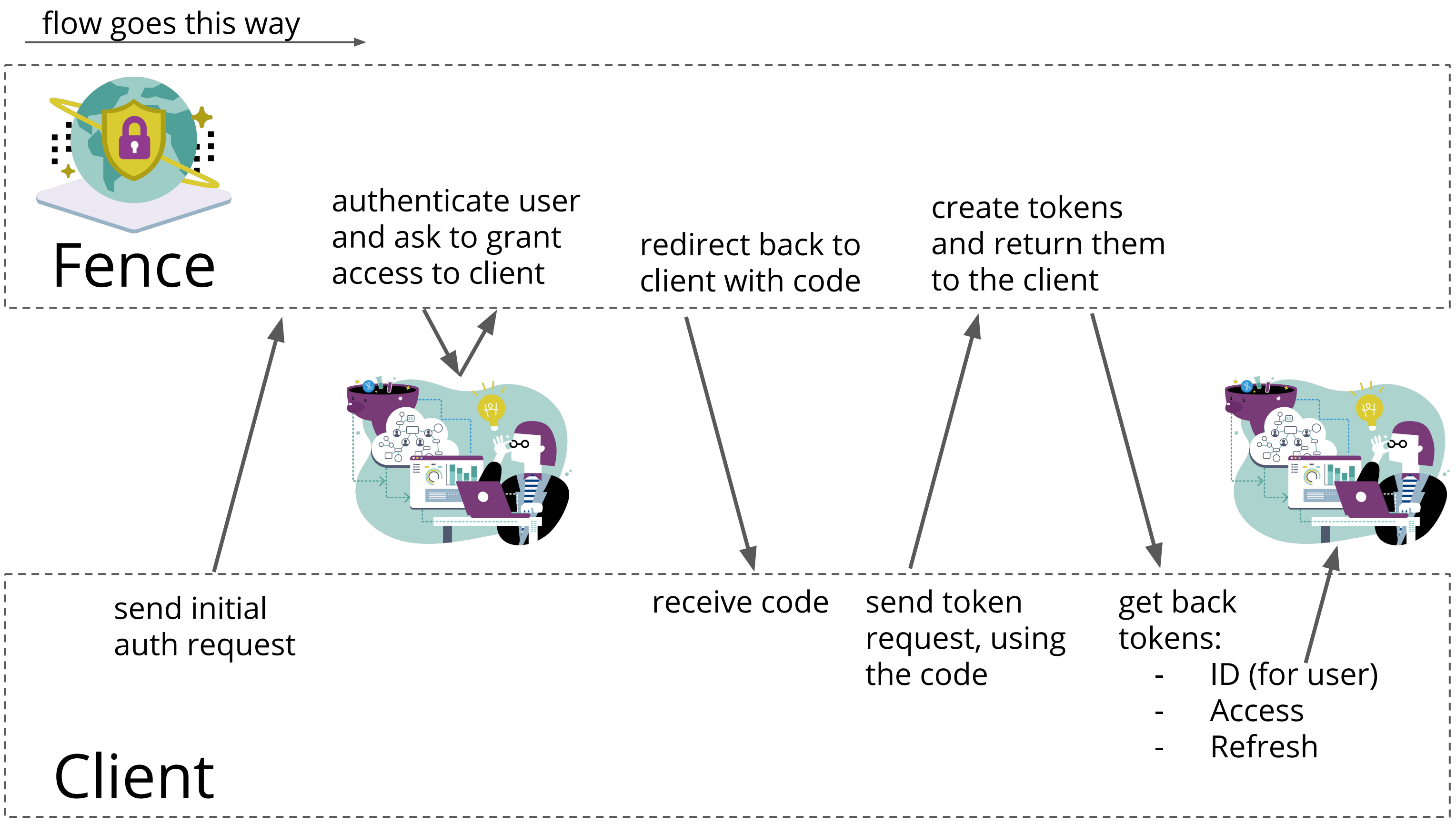
- Know your Google and Fence basic account information and what you are authorized to access.
- Retrieve controlled-access datasets to which you have access on your behalf.

What is OpenID Connect (OIDC)?

(Identity + Authentication) + OAuth2.0 = Open ID Connect

- Authentication Layer on top of OAuth2.0
- Enables secure interoperability across systems

Overview of OAuth2 & OpenID Connect



flow goes this way

Fence

authenticate user and ask to grant access to client

redirect back to client with code

create tokens and return them to the client

send initial auth request

receive code

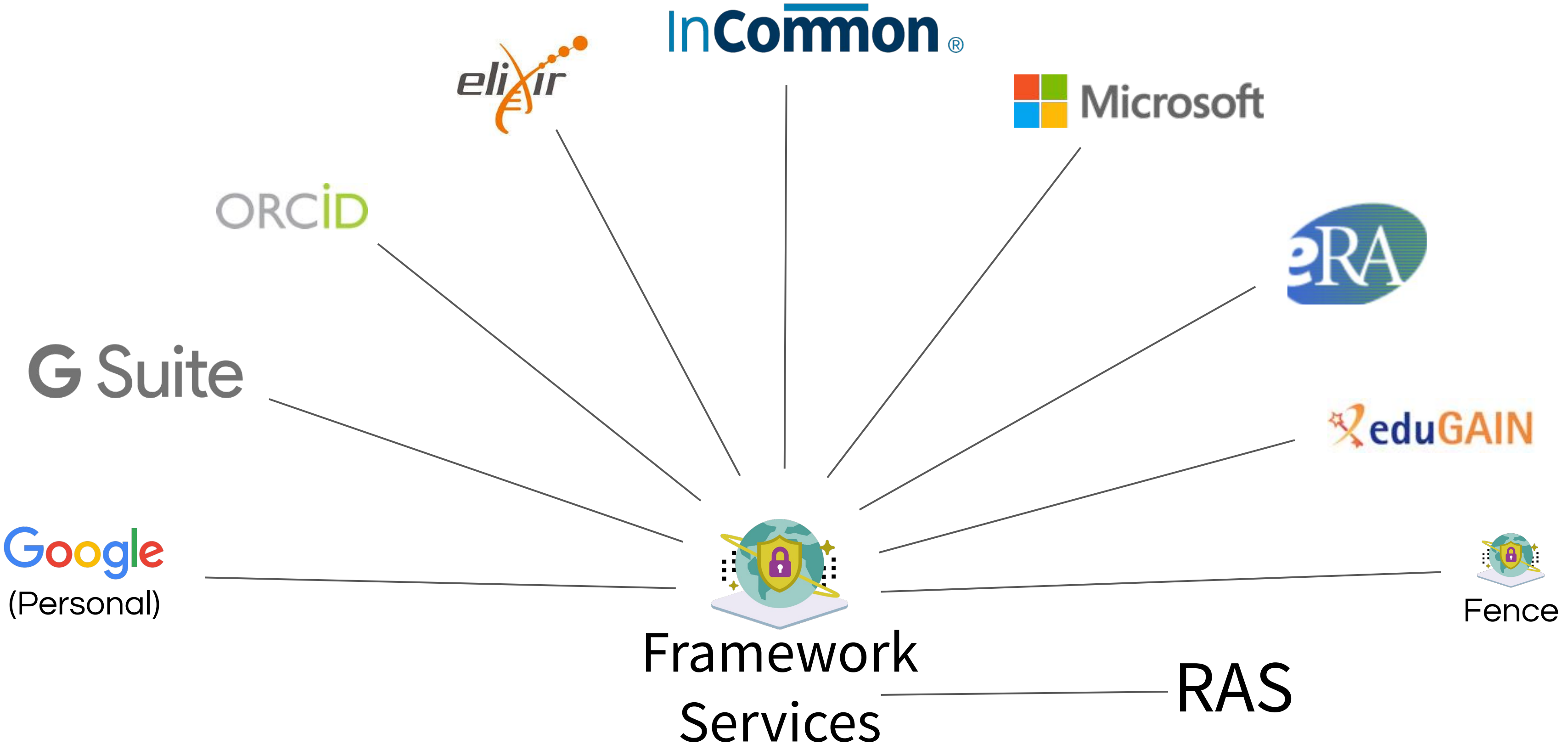
send token request, using the code

get back tokens:

- ID (for user)
- Access
- Refresh

Client

AuthN Provider



Google (Personal)

G Suite

ORCID

elixir

InCommon®

Microsoft

eRA

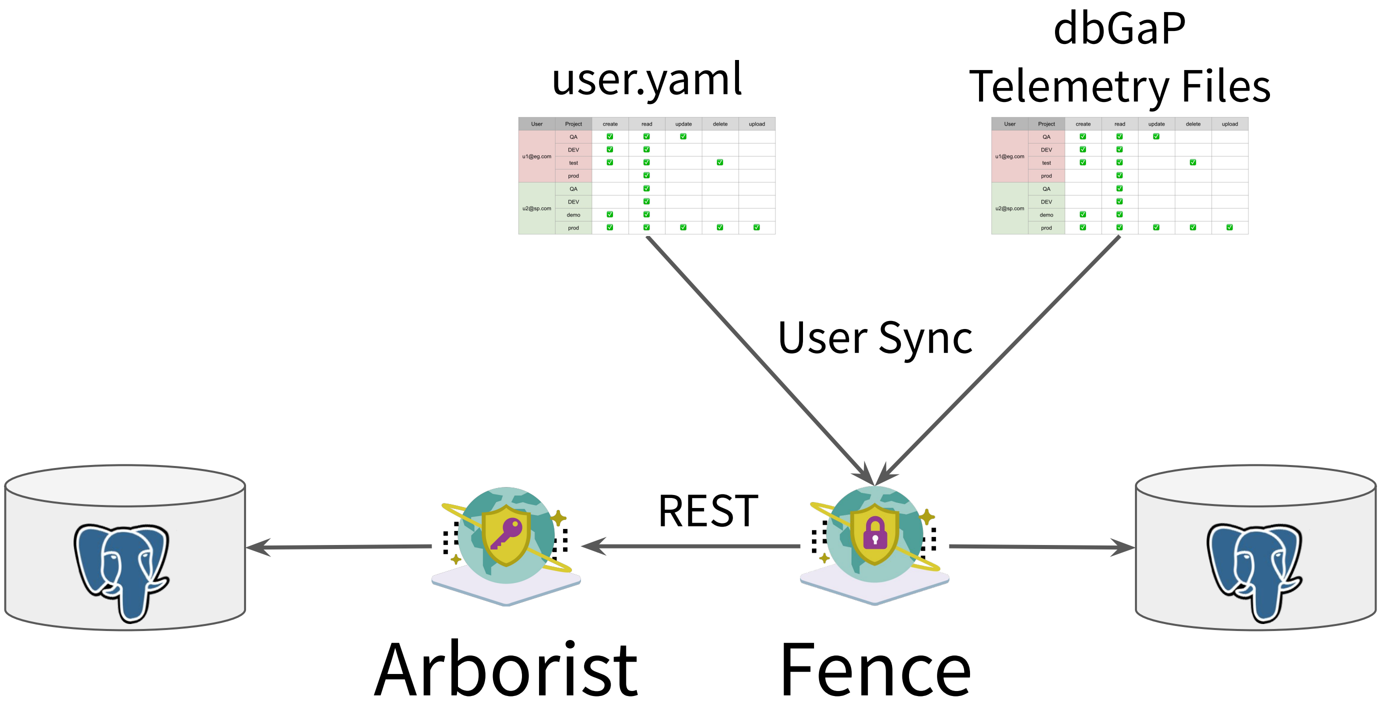
eduGAIN

Fence

Framework Services

RAS

AuthZ Provider



user.yaml

User	Project	create	read	update	delete	upload
u1@eg.com	QA	✅	✅	✅		
	DEV	✅	✅			
	test	✅	✅		✅	
	prod		✅			
u2@sp.com	QA		✅			
	DEV		✅			
	demo	✅	✅			
	prod	✅	✅	✅	✅	✅

dbGaP
Telemetry Files

User	Project	create	read	update	delete	upload
u1@eg.com	QA	✅	✅	✅		
	DEV	✅	✅			
	test	✅	✅		✅	
	prod		✅			
u2@sp.com	QA		✅			
	DEV		✅			
	demo	✅	✅			
	prod	✅	✅	✅	✅	✅

User Sync

REST

Arborist

Fence

Metadata API

A Framework Services API that allows clients to query and retrieve schema-less JSON blobs for GUIDs

Metadata

Current

- Indexd (persistent identifier service)
  - File name
  - File size
  - Checksum
  - URLs/locations

New

- Metadata API
  - Other arbitrary metadata

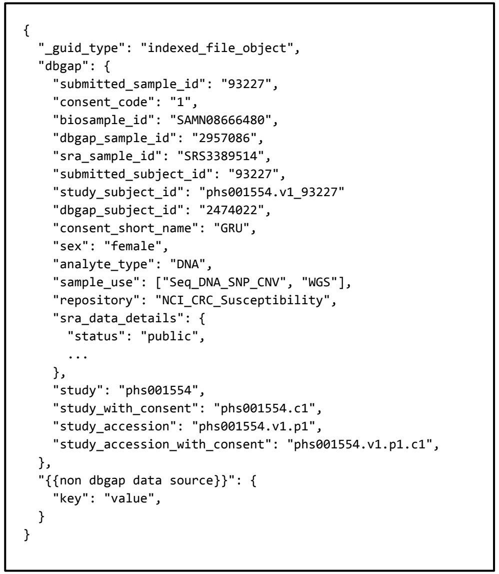
→

Requirements for metadata:

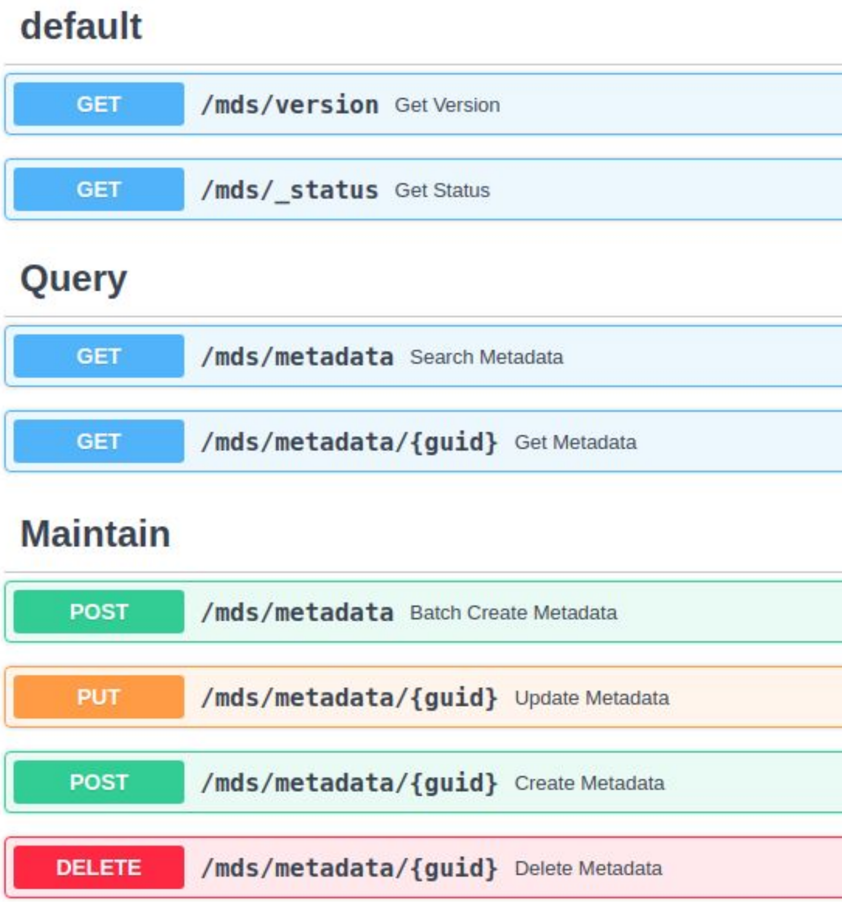
- Publically available data
- Available fully programmatically from a stable API
  - Not manually curated
- Schema-less
  - Cannot enforce restrictions on format

Metadata API

- API for retrieving schema-less JSON metadata blob for GUIDs



```
{
  "_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",
    "sra_data_details": {
      "status": "public",
      ...
    },
    "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

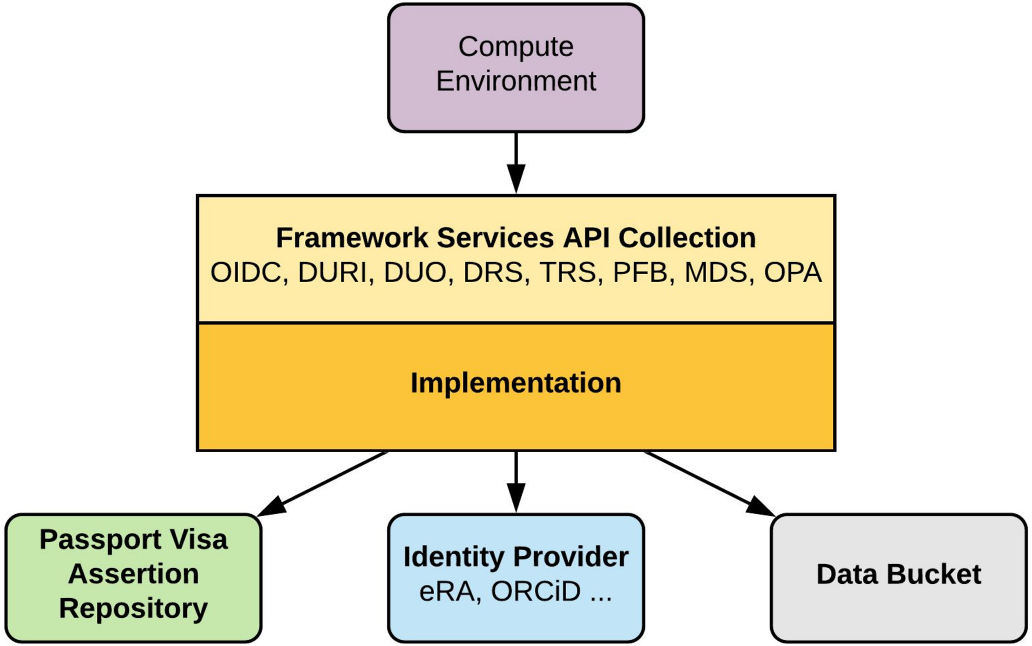
POST /mds/metadata Batch Create Metadata

PUT /mds/metadata/{guid} Update Metadata

POST /mds/metadata/{guid} Create Metadata

DELETE /mds/metadata/{guid} Delete Metadata

Framework Services API Collection


Compute Environment
Framework Services API Collection
OIDC, DURI, DUO, DRS, TRS, PFB, MDS, OPA
Implementation
Passport Visa Assertion Repository
Identity Provider
eRA, ORCiD ...
Data Bucket

The Framework Services mean any collection of services that implements the APIs in the Framework Services API Collection.

Learn More

- github.com/uc-cdis
- Gen3.org
- Gen3 Community on Slack
- dcf-support@datacommons.io
- ctds.uchicago.edu

Selected Data Commons Using Gen3

GEN3
DATA COMMONS

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM

NIH National Heart, Lung, and Blood Institute | BioData CATALYST
Powered by Gen3

NIAID DATA HUB PILOT

NIH National Heart, Lung, and Blood Institute

AnVIL

BRAINCOMMONS

NIH National Human Genome Research Institute

Q&A

GEN3 Data Commons