Application of iRODS to NIEHS Data Management

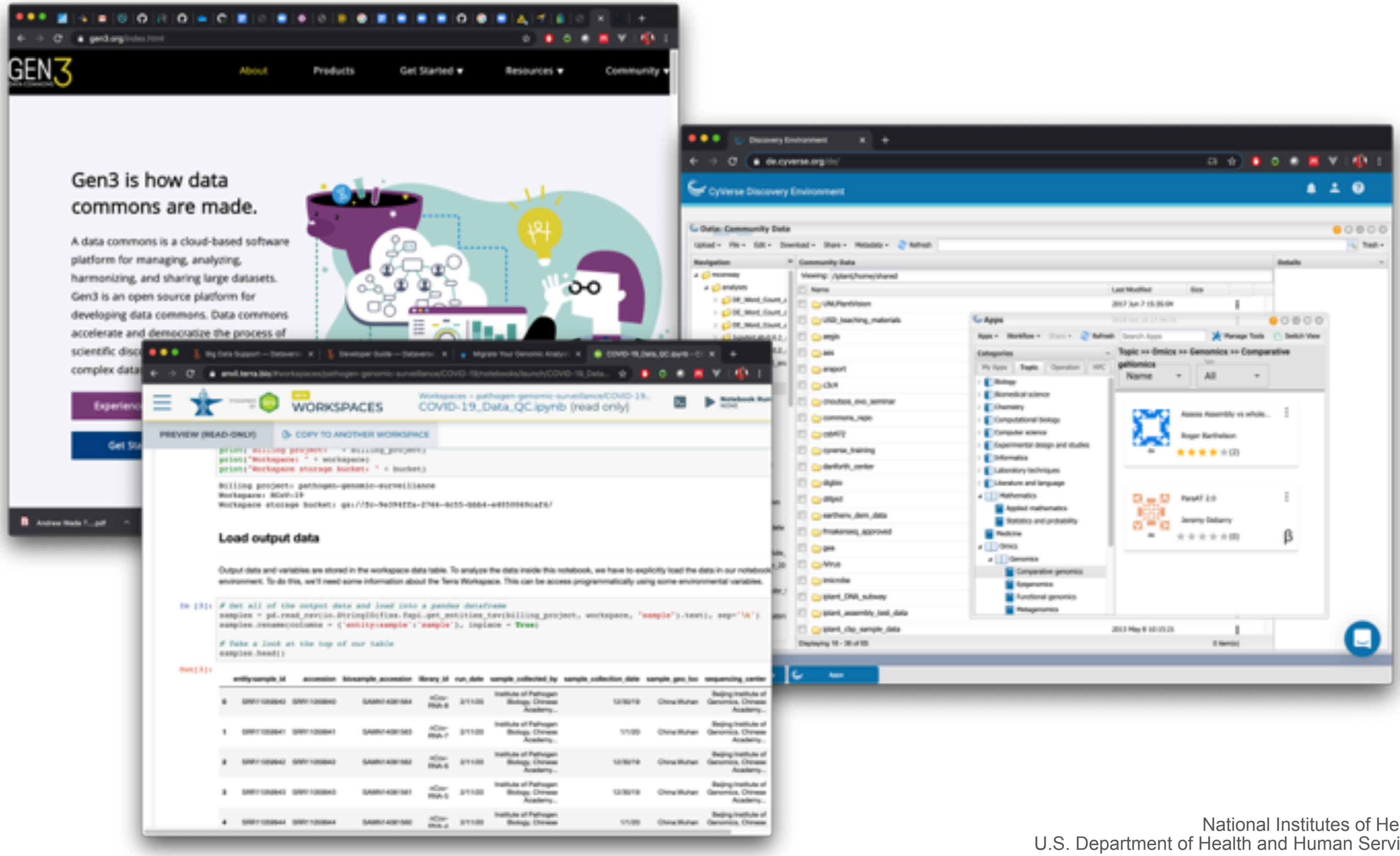
Mike Conway, Deep Patel
Office of Data Science
National Institute of Environmental Health Sciences

What is keeping us awake at night?

Here are two of the things that there's time to mention

https://flic.kr/p/97oo5F

Maintaining Relevance in Key Platforms/Standards



gen3.org/index.html
GEN3
About
Products
Get Started ▾
Resources ▾
Community ▾

Gen3 is how data commons are made.

A data commons is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific disc[illegible] complex data[illegible]

Experienc
Get St
Big Data Support — Dataverse
Developer Guide — Dataverse
Migrate Your Genomic Analysis
COVID-19_Data_QC.ipynb
WORKSPACES
Workspaces › pathogen-genomic-surveillance/COVID-19...
COVID-19_Data_QC.ipynb (read only)
PREVIEW (READ-ONLY)
COPY TO ANOTHER WORKSPACE

```
print("Billing project: " + billing_project)
print("Workspace: " + workspace)
print("Workspace storage bucket: " + bucket)

Billing project: pathogen-genomic-surveillance
Workspace: HCoV-19
Workspace storage bucket: gs://[illegible]/
```

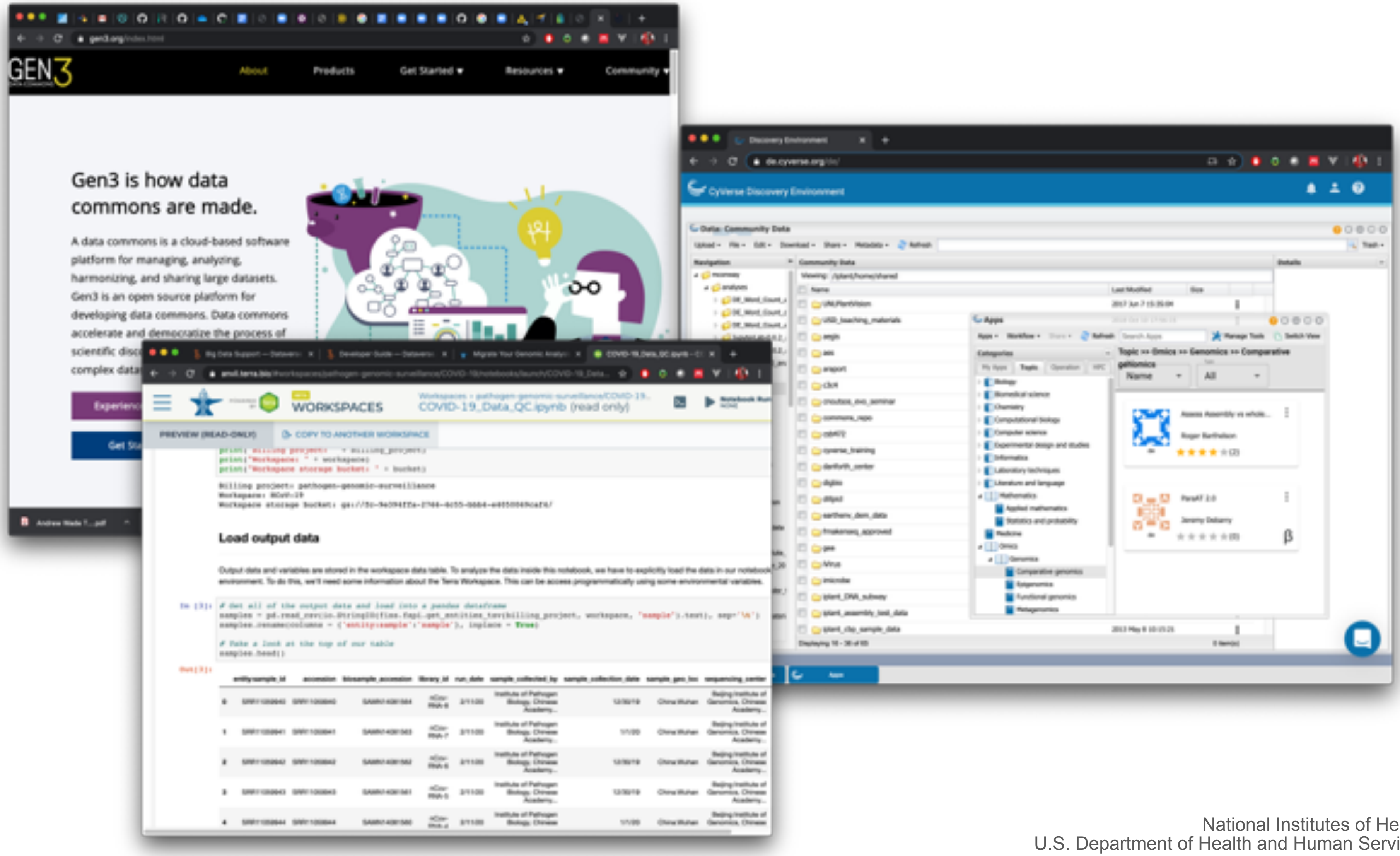
Load output data

Output data and variables are stored in the workspace data table. To analyze the data inside this notebook, we have to explicitly load the data in our notebook environment. To do this, we'll need some information about the Terra Workspace. This can be access programmatically using some environmental variables.

```
In [3]: # Get all of the output data and load into a pandas dataframe
        samples = pd.read_csv(io.StringIO(fiss.fapi.get_entities_tsv(billing_project, workspace, 'sample').text), sep='\t')
        samples.rename(columns = {'entity:sample':'sample'}, inplace = True)

        # Take a look at the top of our table
        samples.head()
Out[3]:
```

	entity:sample_id	accession	biosample_accession	library_id	run_date	sample_collected_by	sample_collection_date	sample_geo_loc	sequencing_center
0	[illegible]	[illegible]	[illegible]	nCov-RNA-[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy	12/30/19	China: Wuhan	Beijing Institute of Genomics, Chinese Academy
1	[illegible]	[illegible]	[illegible]	nCov-RNA-[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy	1/1/20	China: Wuhan	Beijing Institute of Genomics, Chinese Academy
2	[illegible]	[illegible]	[illegible]	nCov-RNA-[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy	12/30/19	China: Wuhan	Beijing Institute of Genomics, Chinese Academy
3	[illegible]	[illegible]	[illegible]	nCov-RNA-[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy	12/30/19	China: Wuhan	Beijing Institute of Genomics, Chinese Academy
4	[illegible]	[illegible]	[illegible]	nCov-RNA-[illegible]	2/11/20	Institute of Pathogen Biology, Chinese	1/1/20	China: Wuhan	Beijing Institute of Genomics, Chinese

Maintaining Relevance in Key Platforms/Standards


gen3.org/index.html
GEN3
About
Products
Get Started ▾
Resources ▾
Community ▾
Experience
Get Started
Andrew Wade T....pdf
Discovery Environment
de.cyverse.org/de/
CyVerse Discovery Environment
Data: Community Data
Upload ▾ File ▾ Edit ▾ Download ▾ Share ▾ Metadata ▾ Refresh
Navigation
Community Data
Details
Viewing: /iplant/home/shared
Name
Last Modified
Size
Apps
Topic >> Omics >> Genomics >> Comparative genomics
Biology
Chemistry
Mathematics
Medicine
Omics
Genomics
Comparative genomics
Apps
Big Data Support — Dataverse
Developer Guide — Dataverse
Migrate Your Genomic Analysis
COVID-19_Data_QC.ipynb
WORKSPACES
Workspaces › pathogen-genomic-surveillance/COVID-19...
COVID-19_Data_QC.ipynb (read only)
PREVIEW (READ-ONLY)
COPY TO ANOTHER WORKSPACE

Gen3 is how data commons are made.

A data commons is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific disc[illegible] complex data[illegible]

```
print("Billing project: " + billing_project)
print("Workspace: " + workspace)
print("Workspace storage bucket: " + bucket)
```

```
Billing project: pathogen-genomic-surveillance
Workspace: HCoV-19
Workspace storage bucket: gs://fc-9e0[illegible]
```

Load output data

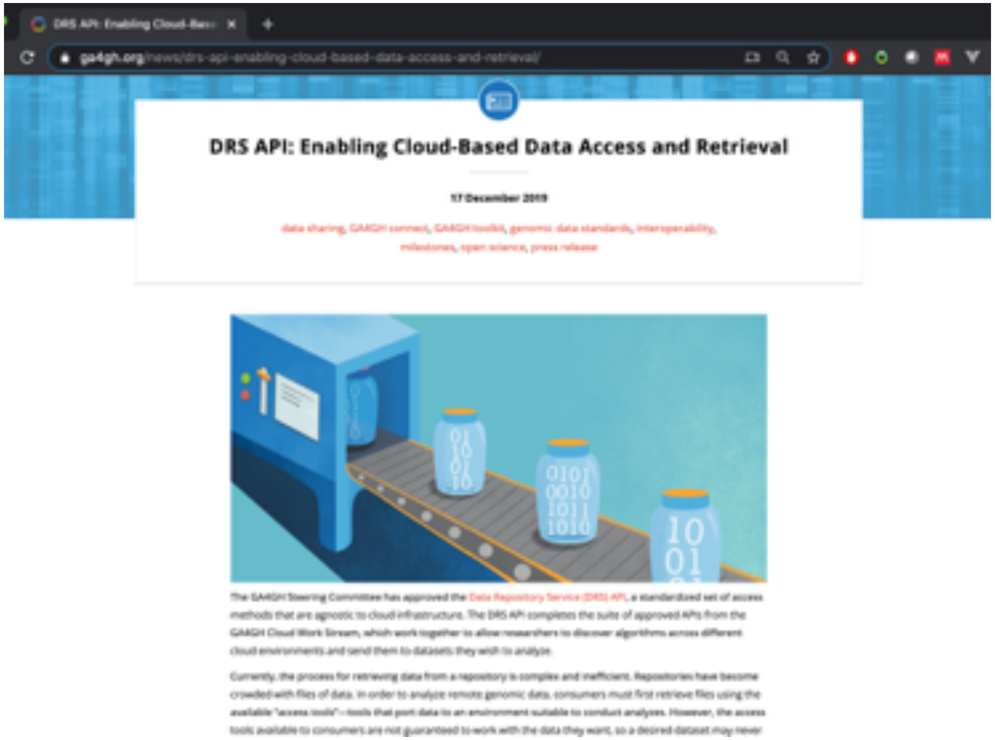
Output data and variables are stored in the workspace data table. To analyze the data inside this notebook, we have to explicitly load the data in our notebook environment. To do this, we'll need some information about the Terra Workspace. This can be access programmatically using some environmental variables.

```
In [3]: # Get all of the output data and load into a pandas dataframe
        samples = pd.read_csv(io.StringIO(fiss.fapi.get_entities_tsv(billing_project, workspace, 'sample').text), sep='\t')
        samples.rename(columns = {'entity:sample':'sample'}, inplace = True)

        # Take a look at the top of our table
        samples.head()
```

Out[3]:

	entity:sample_id	accession	biosample_accession	library_id	run_date	sample_collected_by	sample_collection_date	sample_geo_loc	sequencing_center
0	[illegible]	[illegible]	[illegible]	[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy...	12/30/19	China:Wuhan	Beijing Institute of Genomics, Chinese Academy...
1	[illegible]	[illegible]	[illegible]	[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy...	1/1/20	China:Wuhan	Beijing Institute of Genomics, Chinese Academy...
2	[illegible]	[illegible]	[illegible]	[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy...	12/30/19	China:Wuhan	Beijing Institute of Genomics, Chinese Academy...
3	[illegible]	[illegible]	[illegible]	[illegible]	2/11/20	Institute of Pathogen Biology, Chinese Academy...	12/30/19	China:Wuhan	Beijing Institute of Genomics, Chinese Academy...
4	[illegible]	[illegible]	[illegible]	[illegible]	2/11/20	Institute of Pathogen Biology, Chinese	1/1/20	China:Wuhan	Beijing Institute of Genomics, Chinese

Looking for Pathway to Play Together Nicely


DRS API: Enabling Cloud-Bas...
ga4gh.org/news/drs-api-enabling-cloud-based-data-access-and-retrieval/
10
01

DRS API: Enabling Cloud-Based Data Access and Retrieval

17 December 2019

data sharing, GA4GH connect, GA4GH toolkit, genomic data standards, interoperability, milestones, open science, press release

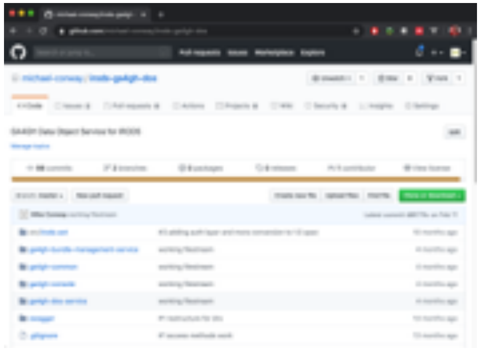
The GA4GH Steering Committee has approved the Data Repository Service (DRS) API, a standardized set of access methods that are agnostic to cloud infrastructure. The DRS API completes the suite of approved APIs from the GA4GH Cloud Work Stream, which work together to allow researchers to discover algorithms across different cloud environments and send them to datasets they wish to analyze.

Currently, the process for retrieving data from a repository is complex and inefficient. Repositories have become crowded with files of data. In order to analyze remote genomic data, consumers must first retrieve files using the available "access tools"—tools that port data to an environment suitable to conduct analyses. However, the access tools available to consumers are not guaranteed to work with the data they want, so a desired dataset may never

- DRS is part of a suite of standards that support distributed execution of tasks, distributed data, and standard workflow execution environments, our “Compute to Data” story
- Gen3 is building DRS support into its platform
- Make iRODS a DRS platform

https://www.ga4gh.org/news/drs-api-enabling-cloud-based-data-access-and-retrieval/

https://github.com/michael-conway/irods-ga4gh-dos



michael-conway / irods-ga4gh-dos

GA4GH Data Object Service for iRODS

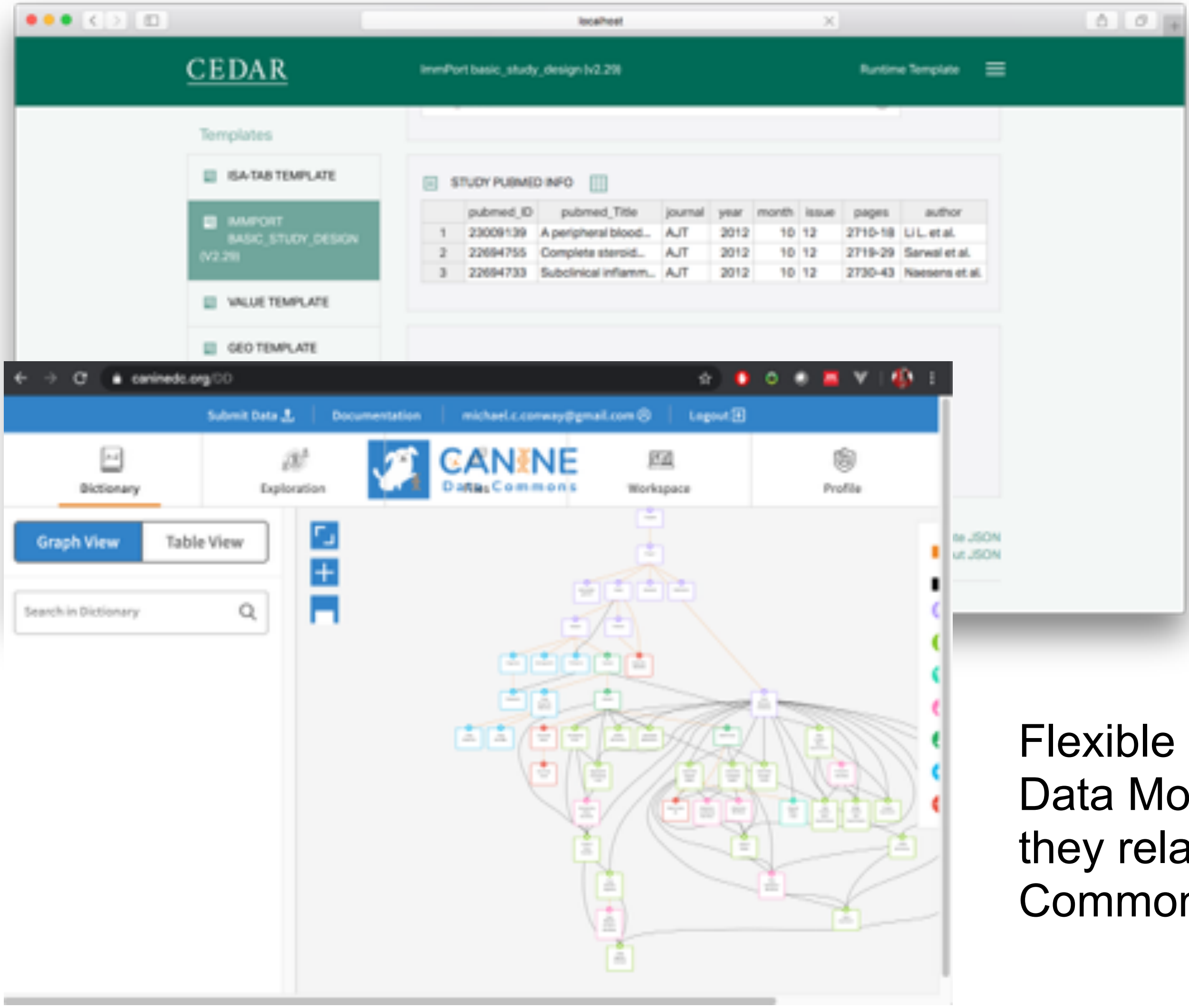
[illegible]

What’s Keeping us Awake at Night

- Handling metadata
  - Curation and getting beyond AVUs
  - Mechanics of ingest of data + metadata
  - Bolting SKOS and Synaptica Graphite to our Commons
  - Indexing (on demand and near real-time)
  - I have an index, how can I search it without polluting community codebases?
  - I can search it, is it useable by relevant communities? How can I micro-target search?

Source: Authors, …………………………………………………. Journal, Vol: pg-pg, year

Structuring Metadata, Metadata Models



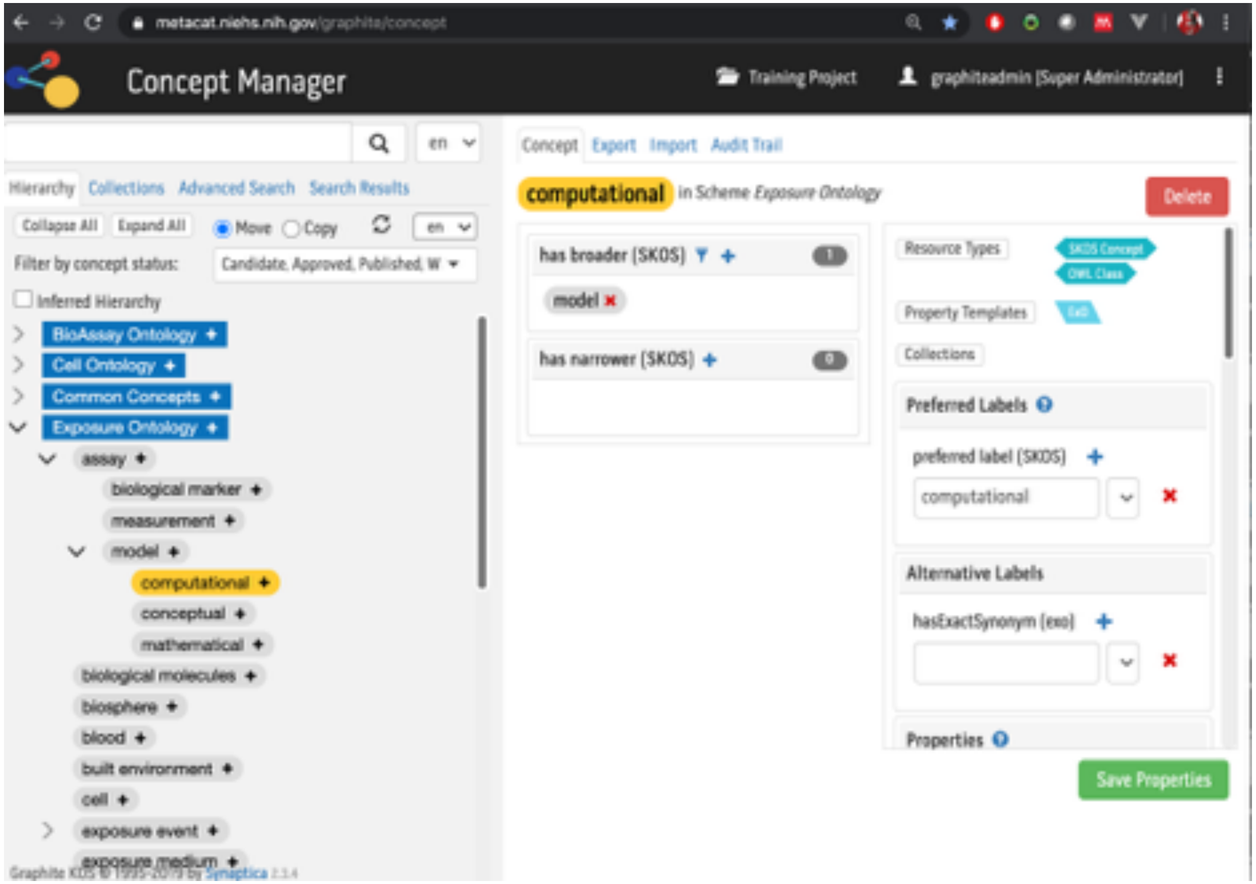
STUDY PUBMED INFO

	pubmed_ID	pubmed_Title	journal	year	month	issue	pages	author
1	23009139	A peripheral blood...	AJT	2012	10	12	2710-18	Li L. et al.
2	22694755	Complete steroid...	AJT	2012	10	12	2719-29	Sarwal et al.
3	22694733	Subclinical inflamm...	AJT	2012	10	12	2730-43	Naesens et al.

Metadata Templates! Working Group making slow but visible progress, this is important!

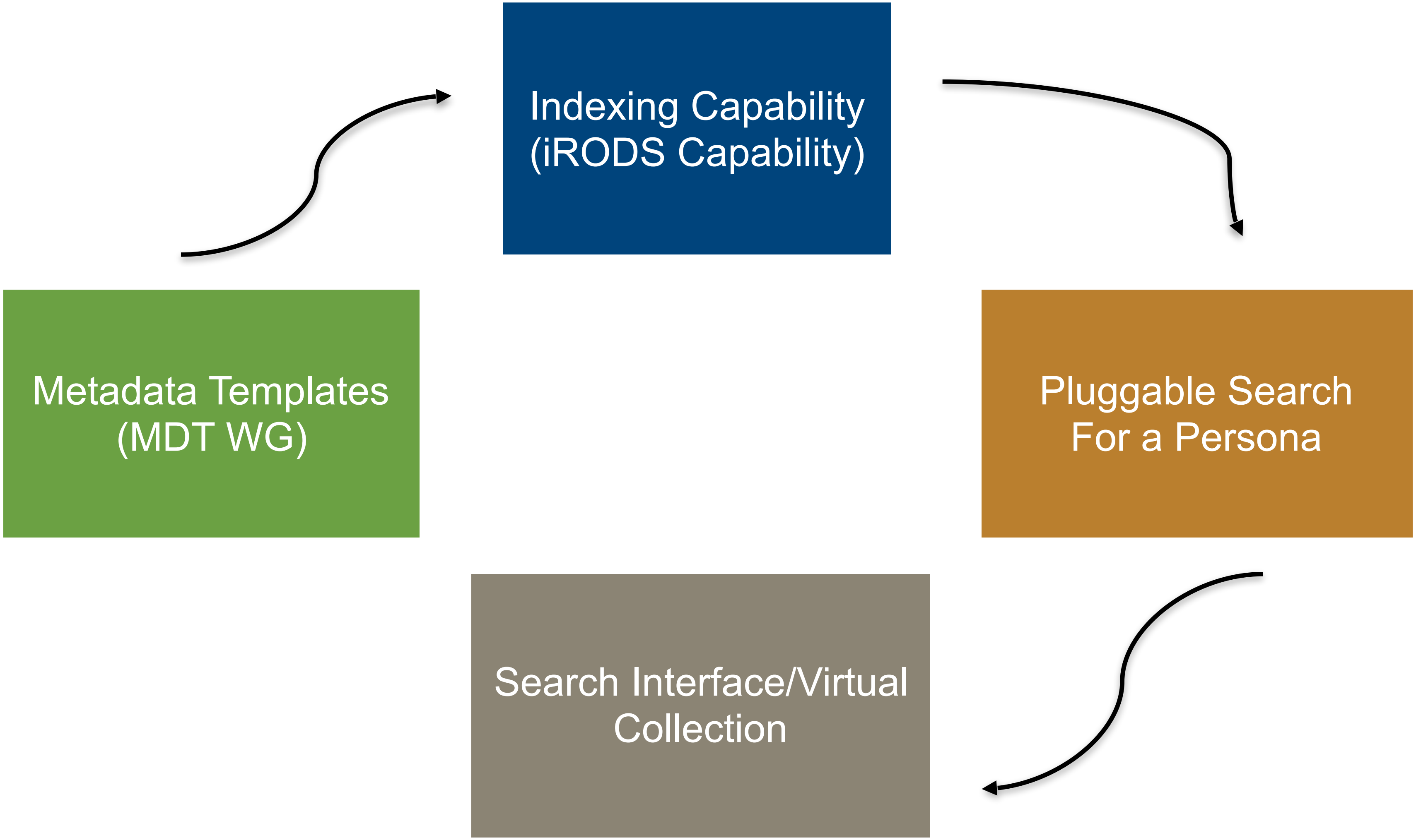
Flexible Semantic Data Models and how they relate to our Commons

Vocabulary and Metadata Management



- How do we incorporate standard terms/labels in templates?
- How can we leverage templates and provide extensible search options and collection formation?

Pluggable Search



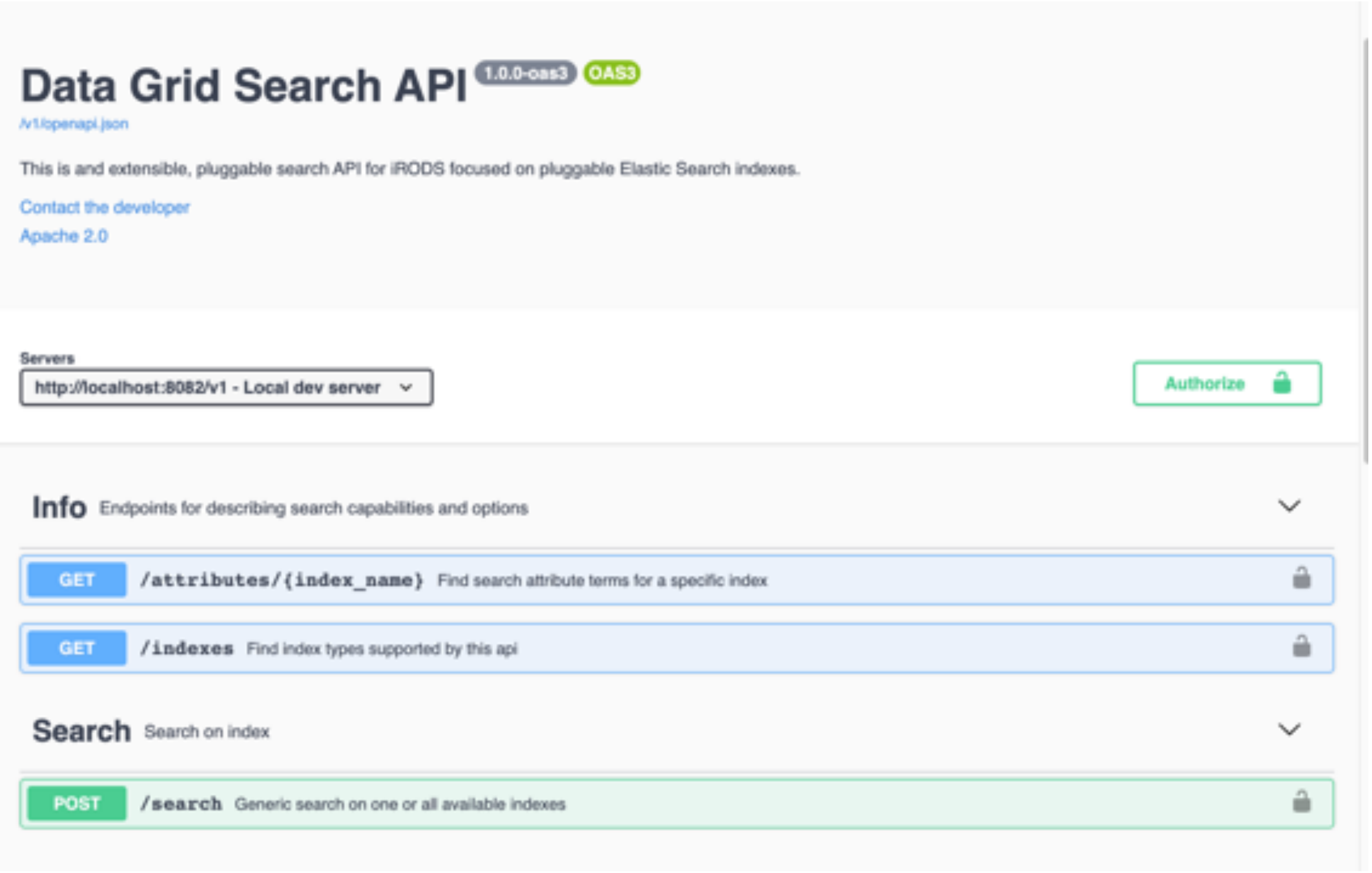
Indexing Capability
(iRODS Capability)

Pluggable Search
For a Persona

Search Interface/Virtual
Collection

Metadata Templates
(MDT WG)

Search Plugins follow simple OpenAPI Spec



Data Grid Search API

/v1/openapi.json

This is and extensible, pluggable search API for iRODS focused on pluggable Elastic Search indexes.

Contact the developer

Apache 2.0

Servers

http://localhost:8082/v1 - Local dev server

Info Endpoints for describing search capabilities and options

GET `/attributes/{index_name}` Find search attribute terms for a specific index

GET `/indexes` Find index types supported by this api

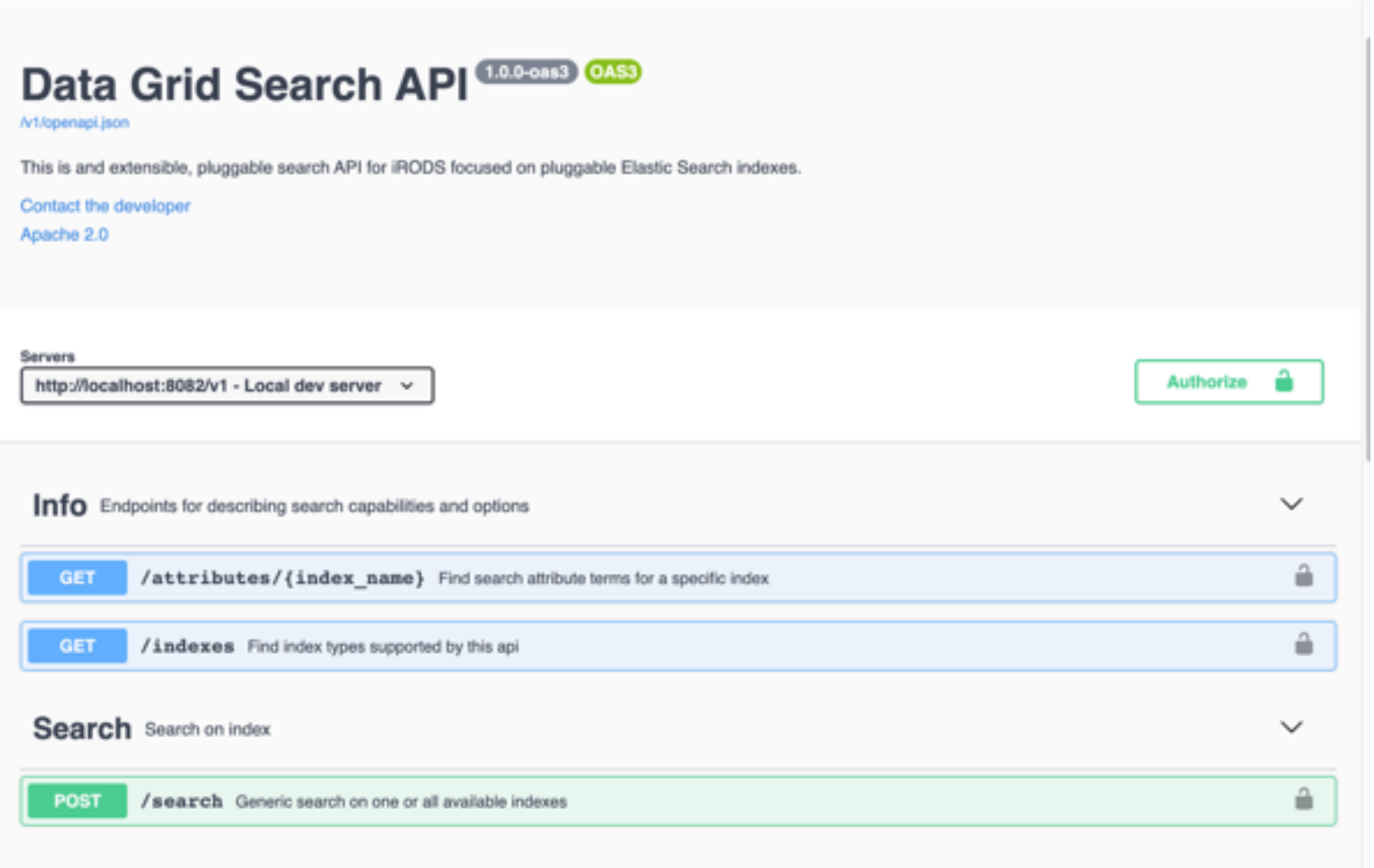
Search Search on index

POST `/search` Generic search on one or all available indexes

Add endpoints in metalnx.properties

```
##################################

# Pluggable search configuration. Turn on and off pluggable search globally, and configure search endpoints.

# N.B. pluggable search also requires provisioning of the jwt.* information above

##################################

# configured endpoints, comma delimited in form https://host.com/v1

pluggablesearch.endpointRegistryList=http://proj_sample_search:8082/v1,http://metadata_search:8082/v1

# enable pluggable search globally and show the search GUI components

pluggablesearch.enabled=true
```

Schema Plugins are interrogated and represented



Data Grid Search API

/v1/openapi.json

This is and extensible, pluggable search API for iRODS focused on pluggable Elastic Search indexes.

Contact the developer

Apache 2.0

Servers

http://localhost:8082/v1 - Local dev server

Info Endpoints for describing search capabilities and options

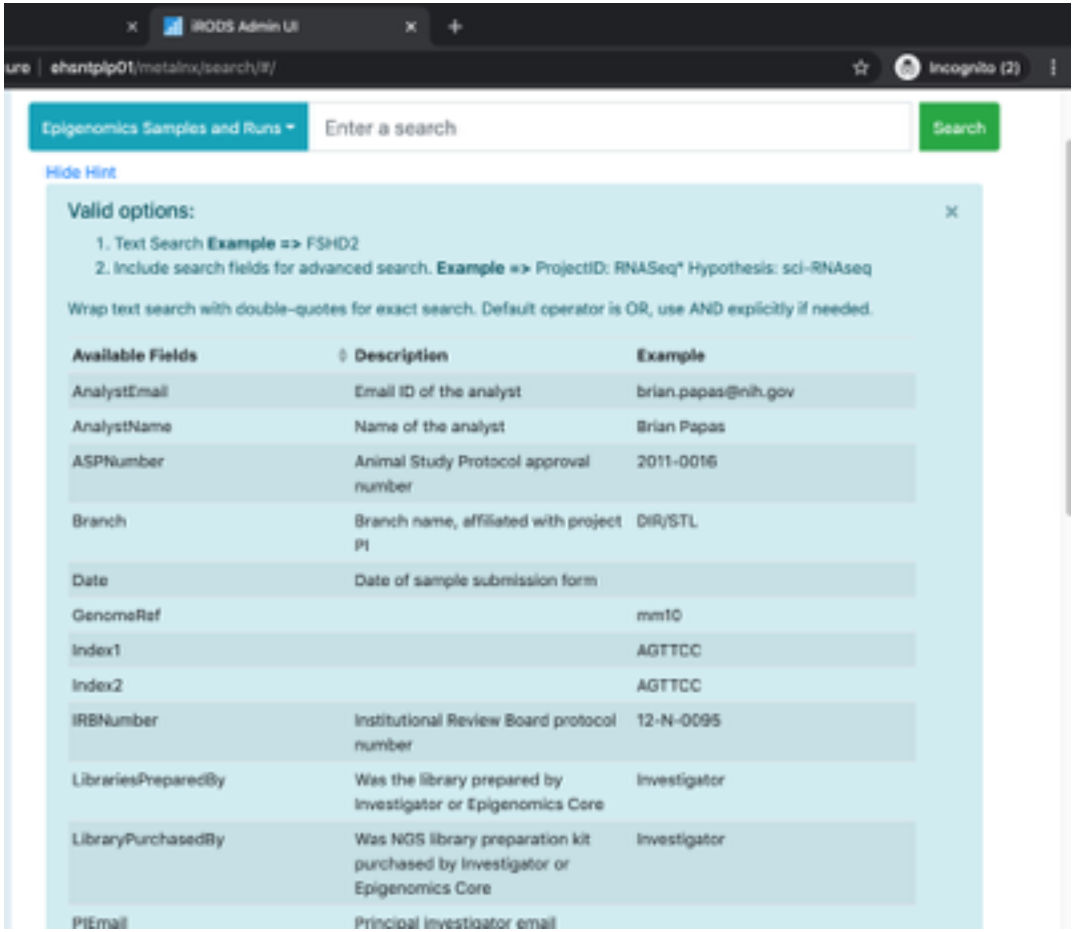
GET `/attributes/{index_name}` Find search attribute terms for a specific index

GET `/indexes` Find index types supported by this api

Search Search on index

POST `/search` Generic search on one or all available indexes

Plugins Advertise Supported Attributes in a Little Language



Valid options:

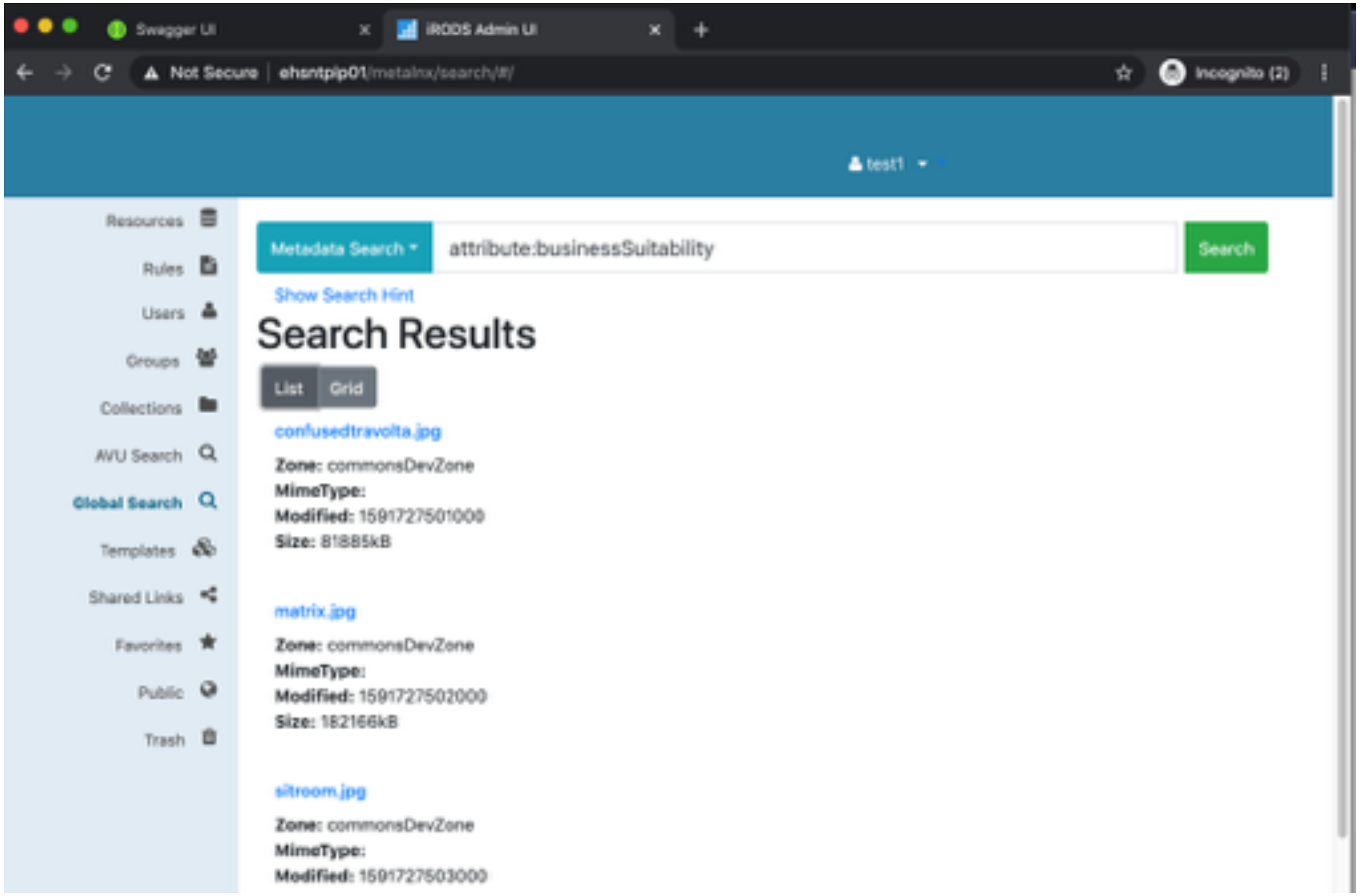
1. Text Search **Example =>** FSHD2
2. Include search fields for advanced search. **Example =>** ProjectID: RNASeq* Hypothesis: sci-RNAseq

Wrap text search with double-quotes for exact search. Default operator is OR, use AND explicitly if needed.

Available Fields	Description	Example
AnalystEmail	Email ID of the analyst	brian.papas@nih.gov
AnalystName	Name of the analyst	Brian Papas
ASPNumber	Animal Study Protocol approval number	2011-0016
Branch	Branch name, affiliated with project PI	DIR/STL
Date	Date of sample submission form	
GenomeRef		mm10
Index1		AGTTCC
Index2		AGTTCC
IRBNumber	Institutional Review Board protocol number	12-N-0095
LibrariesPreparedBy	Was the library prepared by Investigator or Epigenomics Core	Investigator
LibraryPurchasedBy	Was NGS library preparation kit purchased by Investigator or Epigenomics Core	Investigator
PIEmail	Principal investigator email	

- Text entry in familiar ‘advanced query’ form to start
- Builder queries with autocomplete to be supported

Classic Search Result (Plugin can format in interesting ways, including sublinks)



attribute:businessSuitability

Search Results

confusedtravolta.jpg

Zone: commonsDevZone
MimeType:
Modified: 1591727501000
Size: 81885kB

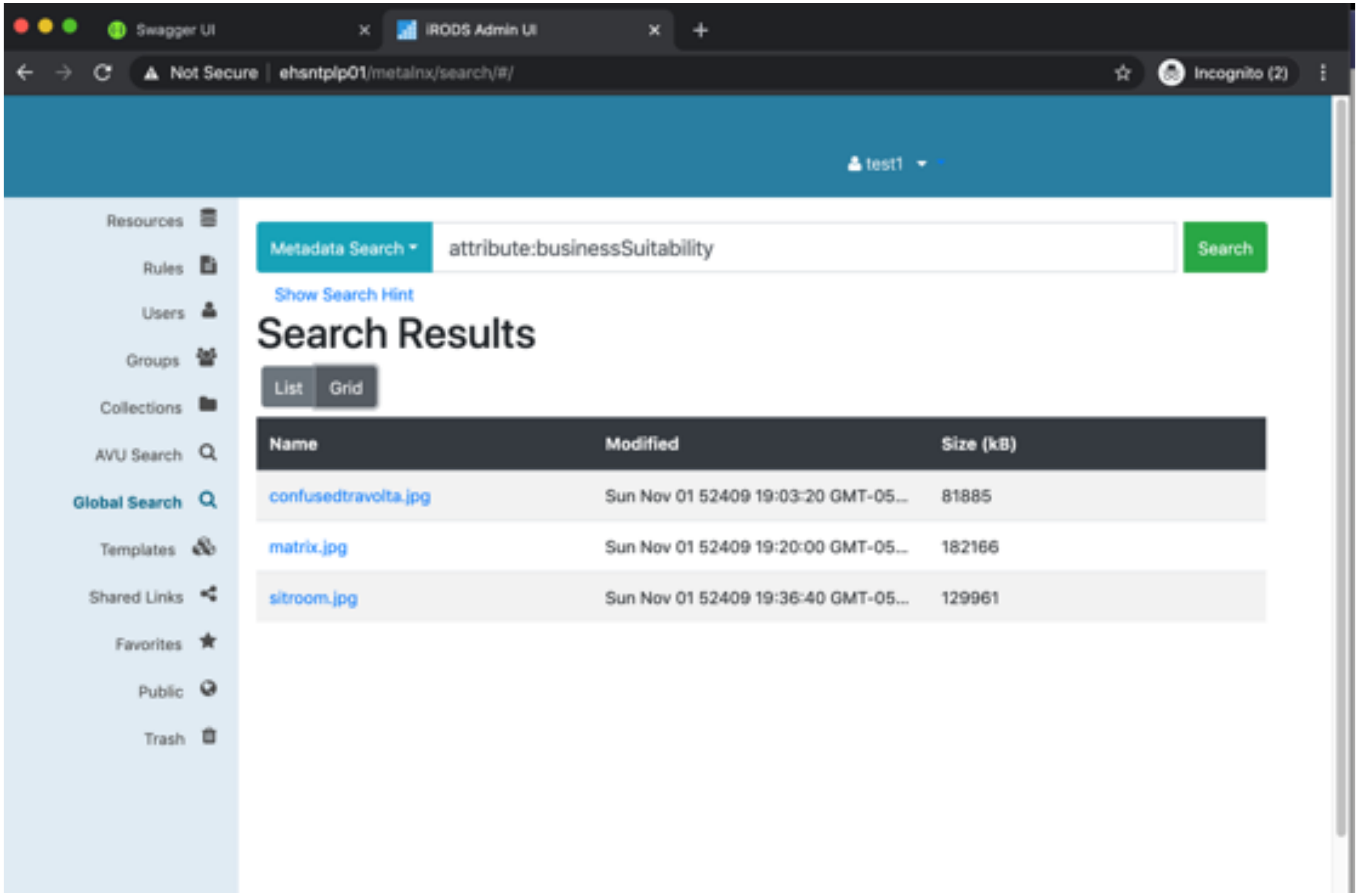
matrix.jpg

Zone: commonsDevZone
MimeType:
Modified: 1591727502000
Size: 182166kB

sitroom.jpg

Zone: commonsDevZone
MimeType:
Modified: 1591727503000

ILS Type File Listing (WIP)



Metadata Search: attribute:businessSuitability

Search Results

Name	Modified	Size (kB)
confusedtravolta.jpg	Sun Nov 01 52409 19:03:20 GMT-05...	81885
matrix.jpg	Sun Nov 01 52409 19:20:00 GMT-05...	182166
sitroom.jpg	Sun Nov 01 52409 19:36:40 GMT-05...	129961