GEN3
DATA COMMONS

Gen3 Development Roadmap

Gen3 Community Forum
24 January 2024

THE UNIVERSITY OF CHICAGO
CENTER FOR TRANSLATIONAL DATA SCIENCE

Australian BioCommons

NeSI
New Zealand eScience Infrastructure

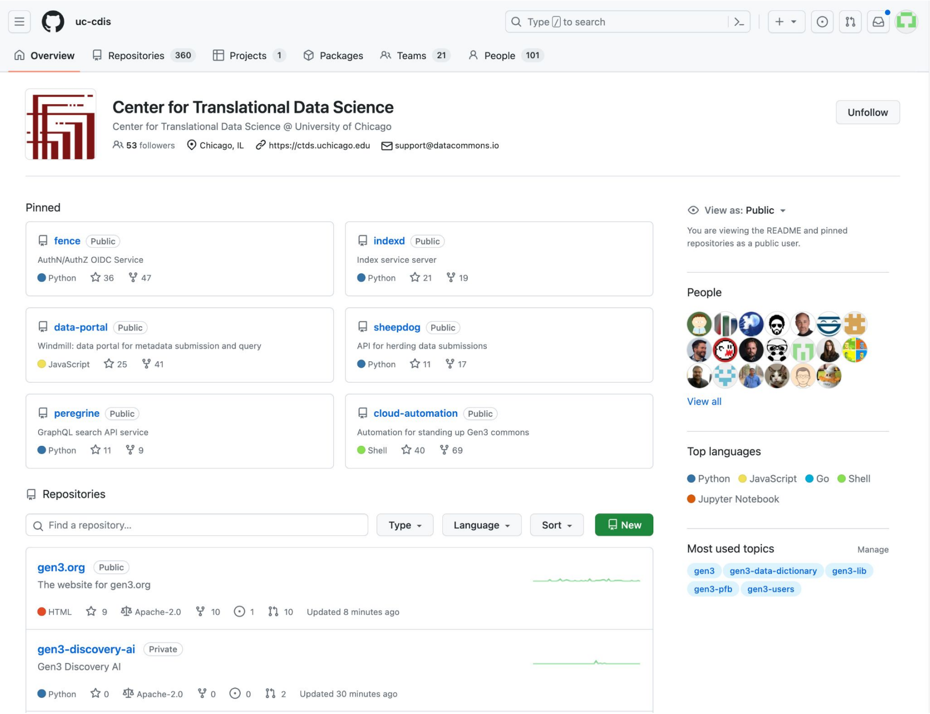
The Agenda

- Introduction
- Gen3 Roadmap
  - Open source support
  - Frontend framework
  - Deployment improvements
  - Workflow execution in workspaces (Nextflow)
  - Large language models for enhanced search
  - Data lakehouse improvements
  - Other improvements
- Steering Committee Discussion
- Q&A

Gen3 Product Roadmap

1. Gen3 Open-Source Support

- Dedicated support for external contributions to Gen3 codebase including timely review of PRs
- Allow for external contributors to test PRs using our integration test suite (to run on your infrastructure)
- Allow issue reporting and tracking
- **Will provide more details in next Gen3 Community Forum**



Center for Translational Data Science

Center for Translational Data Science @ University of Chicago

53 followers · Chicago, IL · https://ctds.uchicago.edu · support@datacommons.io

You are viewing the README and pinned repositories as a public user.

Repository	Visibility	Description	Language	Stars	Forks
fence	Public	AuthN/AuthZ OIDC Service	Python	36	47
indexd	Public	Index service server	Python	21	19
data-portal	Public	Windmill: data portal for metadata submission and query	JavaScript	25	41
sheepdog	Public	API for herding data submissions	Python	11	17
peregrine	Public	GraphQL search API service	Python	11	9
cloud-automation	Public	Automation for standing up Gen3 commons	Shell	40	69

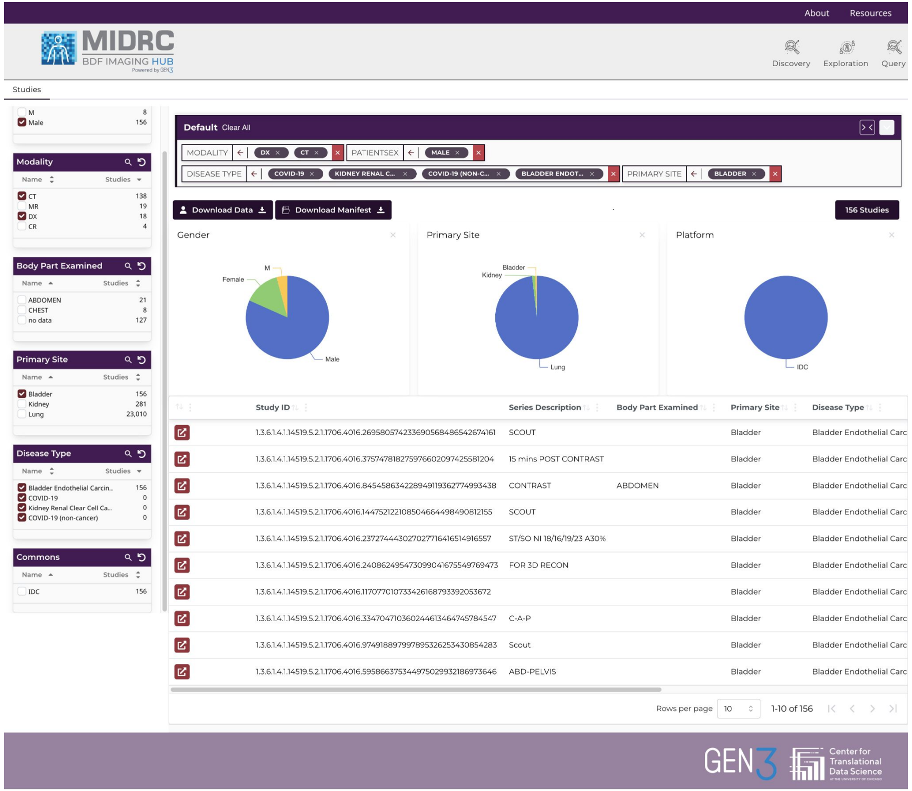
Top languages: Python, JavaScript, Go, Shell, Jupyter Notebook

Most used topics: gen3, gen3-data-dictionary, gen3-lib, gen3-pfb, gen3-users

Repository	Visibility	Description	Language	Stars	License	Forks	Issues	Pull requests	Updated
gen3.org	Public	The website for gen3.org	HTML	9	Apache-2.0	10	1	10	Updated 8 minutes ago
gen3-discovery-ai	Private	Gen3 Discovery AI	Python	0	Apache-2.0	0	0	2	Updated 30 minutes ago

2. Gen3 Frontend Framework & Analysis Tool Framework

- Replacement for Gen3 Data Portal
- Moving to “App Store” like framework for analysis and visualization tools
- Updated technology stack
- Improved development and user experience
- Simplifies project specific content and customizations
- Beta launch at the end of January



	Name	Studies
☐	M	8
☑	Male	156

Modality

	Name	Studies
☑	CT	138
☐	MR	19
☑	DX	18
☐	CR	4

Body Part Examined

	Name	Studies
☐	ABDOMEN	21
☐	CHEST	8
☐	no data	127

Primary Site

	Name	Studies
☑	Bladder	156
☐	Kidney	281
☐	Lung	23,010

Disease Type

	Name	Studies
☑	Bladder Endothelial Carcin...	156
☑	COVID-19	0
☑	Kidney Renal Clear Cell Ca...	0
☑	COVID-19 (non-cancer)	0

Commons

	Name	Studies
☐	IDC	156

Study ID	Series Description	Body Part Examined	Primary Site	Disease Type
1.3.6.1.4.1.14519.5.2.1.1706.4016.269580574233690568486542674161	SCOUT		Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.375747818275976602097425581204	15 mins POST CONTRAST		Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.845458634228949119362774993438	CONTRAST	ABDOMEN	Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.144752122108504664498490812155	SCOUT		Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.237274443027027716416514916557	ST/SO NI 18/16/19/23 A30%		Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.240862495473099041675549769473	FOR 3D RECON		Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.117077010733426168793392053672			Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.334704710360244613464745784547	C-A-P		Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.974918897997895326253430854283	Scout		Bladder	Bladder Endothelial Carc
1.3.6.1.4.1.14519.5.2.1.1706.4016.595866375344975029932186973646	ABD-PELVIS		Bladder	Bladder Endothelial Carc

Rows per page 10 1-10 of 156

2. Frontend Framework

Analysis Tool Framework (ATF)

- Supports the development of custom analysis tools
- Connects to:
    - Gen3 services
    - 3rd-party APIs
    - Other data sources
- Uses context to filter tools that can be applied to current data selection
- New Gen3 Frontend:
    - Redesigned standard pages: Explorer, Discovery…
    - Analysis tool center for accessing applications
    - Commons specific functionality/components (for use in tables, charts)
    - More integration/data sharing between applications

3. Deployment Improvements

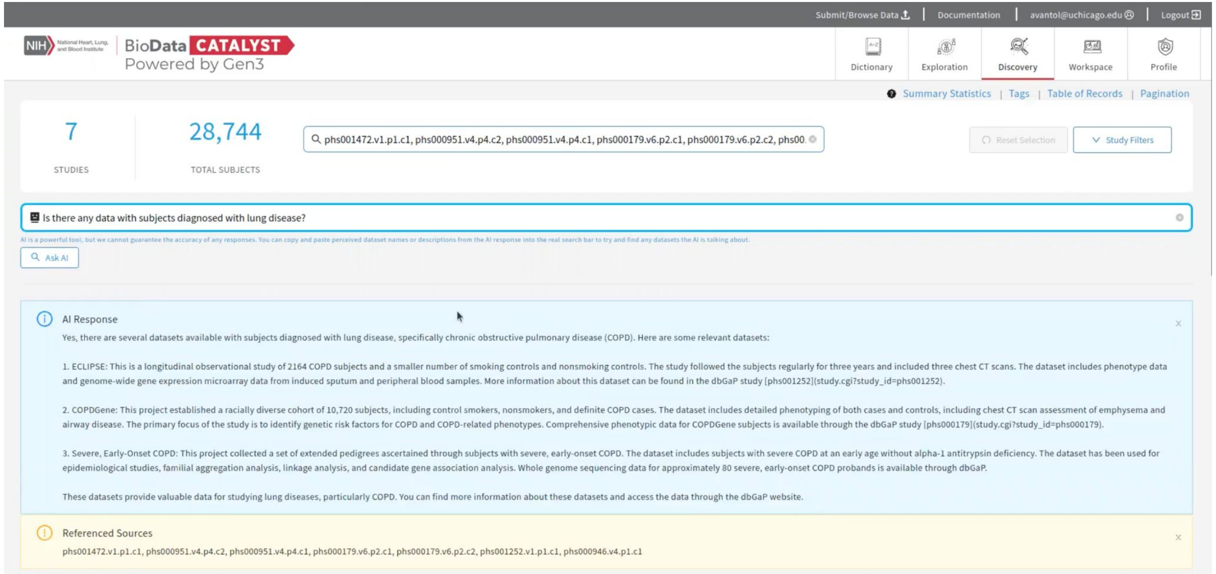
- In 2023, Helm was rolled out and can now be used to deploy Gen3
- In 2024, CTDS will begin to use Helm for production deployments. Also planned for 2024:
    - Full test suite w/ helm - Getting a full testing suite running against Helm deployments, which will be available to external users
    - Observability - Incorporating observability tools in the Gen3 Helm charts – better monitoring and logging, and general visibility into the health of the deployments.
- Gen3 Lite - Run a lightweight Gen3 on a single instance. Cheaper and great for non-production environments.

4. Workflow execution in workspaces (Nextflow)

- A general purpose workflow execution system in Gen3 that researchers can use to run *containers* on the cloud for various applications in a secure and isolated manner
- Phase 1
  - Develop and test infrastructure on containers developed by users and run by Gen3 operator - Complete
- Phase 2
  - Develop and test infrastructure with CLI push credentials for containers developed by users and run by users - Currently being tested.
- Phase 3
  - Develop a friendly portal for users to submit containers, track jobs

5. Large Language Models for Enhanced Search

- Natural language queries and responses of data on the Discovery page
- Will use a RAG approach initially (Retrieval Augmented Generation) and in the future may involve training a new model or fine-tuning an existing model
- Planning to include an API for vectorbases to support RAG



phs001472.v1.p1.c1, phs000951.v4.p4.c2, phs000951.v4.p4.c1, phs000179.v6.p2.c1, phs000179.v6.p2.c2, phs00

Is there any data with subjects diagnosed with lung disease?

AI is a powerful tool, but we cannot guarantee the accuracy of any responses. You can copy and paste perceived dataset names or descriptions from the AI response into the real search bar to try and find any datasets the AI is talking about.

AI Response

Yes, there are several datasets available with subjects diagnosed with lung disease, specifically chronic obstructive pulmonary disease (COPD). Here are some relevant datasets:

1. ECLIPSE: This is a longitudinal observational study of 2164 COPD subjects and a smaller number of smoking controls and nonsmoking controls. The study followed the subjects regularly for three years and included three chest CT scans. The dataset includes phenotype data and genome-wide gene expression microarray data from induced sputum and peripheral blood samples. More information about this dataset can be found in the dbGaP study [phs001252](study.cgi?study_id=phs001252).

2. COPDGene: This project established a racially diverse cohort of 10,720 subjects, including control smokers, nonsmokers, and definite COPD cases. The dataset includes detailed phenotyping of both cases and controls, including chest CT scan assessment of emphysema and airway disease. The primary focus of the study is to identify genetic risk factors for COPD and COPD-related phenotypes. Comprehensive phenotypic data for COPDGene subjects is available through the dbGaP study [phs000179](study.cgi?study_id=phs000179).

3. Severe, Early-Onset COPD: This project collected a set of extended pedigrees ascertained through subjects with severe, early-onset COPD. The dataset includes subjects with severe COPD at an early age without alpha-1 antitrypsin deficiency. The dataset has been used for epidemiological studies, familial aggregation analysis, linkage analysis, and candidate gene association analysis. Whole genome sequencing data for approximately 80 severe, early-onset COPD probands is available through dbGaP.

These datasets provide valuable data for studying lung diseases, particularly COPD. You can find more information about these datasets and access the data through the dbGaP website.

Referenced Sources

phs001472.v1.p1.c1, phs000951.v4.p4.c2, phs000951.v4.p4.c1, phs000179.v6.p2.c1, phs000179.v6.p2.c2, phs001252.v1.p1.c1, phs000946.v4.p1.c1

6. Gen3 Data Lakehouse Improvements

- Allow data files to be shared through Gen3 either without or before creation of a data dictionary and population of the data model
- Can associate searchable metadata with files or groups of files and projects (via Gen3 Metadata API)
- Adding a per-user data library and updating pages in Gen3 to improve ability to access files before metadata harmonization
- Groups of files and/or a data dictionary could also be distributed in the data lake
  - Will support a packaged file format like Portable Format for Bioinformatics (PFB)
    - PFB includes the structured data (i.e. graph model) and pointers to data files all in one
  - Will also support a .zip or other combination of file formats

7. Other improvements

- Frictionless
  - Export Gen3 data into Frictionless format
  - Import metadata from frictionless data packages
  - Adding tools in workspaces for working with frictionless data packages
- We plan to release an annual Gen3 product roadmap

Gen3 Panel Discussion

Gen3 Steering Committee

- Robert Grossman - University of Chicago
- Claire Rye - New Zealand eScience Infrastructure
- Steven Manos - Australian BioCommons
- Plamen Martinov - Open Commons Consortium
- Kyle Ellrott - Oregon Health and Science University

Questions