

Gen3 2025 Development Roadmap

Gen3 Community Forum
26 February 2025

- Announcements
- Review of 2024 Roadmap
- 2025 Roadmap
- Steering Committee Discussion
- Q&A

- New documentation site released! <https://docs.gen3.org>
- Next community forum will be on May 7, but the topic has not been finalized

2024 Gen3 Product Roadmap

Summary of 2024 roadmap

1. Gen3 Open-source Support **Complete**
2. Frontend Framework **Mostly Complete**
3. Helm for Deployment **Mostly Complete**
4. Integration Test Suite **Mostly Complete**
5. Observability **Complete**
6. Gen3 Lite **Mostly Complete**
7. Nextflow Integration **Phase 2 Complete**
8. RAG Interface for dataset search **Complete**
9. Data Lakehouse **Partially Complete**
10. Annual product roadmap **Complete**

Other requests provided at last year's roadmap meeting

1. Better visibility into roadmap progress **Partially complete**
2. More fine-grained access control **Not started**
3. Trusted research environment like capabilities **Partially complete**
4. Better data access request management **Partially complete**
5. Implementation of a native graph database like neo4j or neptune **Not started**
6. Better documentation for each microservice to help with extending Gen3 (API definitions and service interactions) **Partially complete**
7. Self-assembled working groups of interest **Not started**

2025 Gen3 Product Roadmap (Draft)

- Commons Services Operations Center (CSOC) Support
- Task Execution Service (TES) Support
- Integrated AI capabilities
- Improved Documentation
- Frontend Improvements
- Mesh and Node Cards
- TBD
- TBD

- A commons services operations center (CSOC) is used by organizations that run more than one Gen3 system.
- A CSOC allows a team of engineering and security staff to set up, configure, secure, operate, and monitor two or more data commons or data meshes.
- The CSOC Working Group is developing a resource to automate both infrastructure set up and Gen3 deployment
- This is a key step so that organizations can easily run 2 or more Gen3 commons

One Portal for End-to-End Gen3 Lifecycle Management

- Seamlessly deploy, configure, and manage **multiple** Gen3 instances.
- Enable **zero to production** Gen3 deployments effortlessly.

Multi-Cloud Infrastructure Provisioning

- Support **Kubernetes** deployments across multiple cloud providers.
- Simplify cloud infrastructure setup and maintenance.

Community-Driven

- Incorporate **laC contributions** from AU Biocommons, Krumware, OCC and other community partners.
- Collaborate on design and other features

Comprehensive Monitoring & Management

- Unified visibility into **Gen3, Kubernetes, and cloud resources** for proactive issue resolution.

- Support container execution by implementing the GA4GH TES (Task Execution Service) standard, through an integration of the distributed task execution tool Funnel
- Support workflow execution by implementing support for Nextflow workflows through the Gen3 TES API
- The Gen3 TES API improves our ability to isolate tasks and workflows and allows for greater flexibility in running jobs outside of a workflow.
- This will allow us to uncouple Nextflow from workspaces and simplify the addition of other workflow specification languages in the future

- Integrate open-source vector store into Gen3
- Add APIs for embeddings
- Add Gen3 services to create embeddings
- Add services to fine tune open source AI models over data in commons
- Add services for creating synthetic data from data into commons
- Explore architectures for building small to midscale LLM/GenAI over data in Gen3 commons
- Explore architectures for supporting federated / distributed AI in Gen3 meshes
- This will be a multi-year project different from the other updates included in the roadmap

Gen3 Documentation

[Home](#)

[About Gen3](#)

[Gen3 Community](#)

[User Guide](#)

[Operator Guide - Deploy Gen3](#)

[Developer Guide - Extend Gen3](#)

[Blog](#)

[Frequently Asked Questions](#)

[Glossary](#)

[Demo @ gen3.datacommons.io](#)

Gen3 Documentation

This is your home for all technical documentation related to the design, deployment, use, and maintenance of a Gen3 data commons or mesh.

Please visit [Gen3.org](#) if you would like a high-level overview of Gen3 as well as details about the Gen3 philosophy, community events, and governance.

Gen3 documentation is organized by the category of person interacting with Gen3:

- **Gen3 User** - This is a data scientist, researcher, or analyst who needs to explore, download, or analyze data found within an existing instance of Gen3.
- **Gen3 Developer** - This is a software engineer who wants to extend Gen3 either by contributing to the source code or by integrating Gen3 services into a larger system. This section will cover the Gen3 architecture including the individual microservices and how they interact with each other.
- **Gen3 Operator** - This is for those organizations who operate their own Gen3 instances. It will include content on how to Deploy or Spin up a Gen3 instance, configure a data dictionary and upload data, and customize the frontend.

Next
[About Gen3](#) →

- Migration of all previous technical content on gen3.org to docs.gen3.org. Revision and addition of many new pages.
- Overall structure more clearly organized into separate sections based on use cases: Users, Developers, and Operators.
- Integration of helm documentation previously found at docs.gen3.org. Old versions now found at old.docs.gen3.org.
- New site is now generated using MkDocs and nearly all content is now exclusively in markdown format.
- Contributions by community to docs are better supported due to simpler content and contributor instructions.
- Content will now be versioned to align with Gen3 software.
- **More content to come in 2025!**

Frontend Improvements

Gen3.2

- Roll out remaining features (workspaces and dictionary view)
- Migrate existing data commons and meshes to Gen3.2
- Improve documentation
- Additional features in the form of apps

The screenshot displays the MIDRC BDF IMAGING HUB interface. On the left, there are four filter panels: Gender, Modality, Body Part Examined, and Primary Site. The Gender panel shows counts for Female (342,568), Male (280,394), no data (17,545), and Not Reported (667,645). The Modality panel lists CR (137,381), CT (406,384), DX (104,049), FUSION (18), KO (40), and MSD (2,328). The Body Part Examined panel lists ABD (2), ABD_PEL (9), ABD_PELV (60), ABDOMEN (4), and ABDOMEN (17,823). The Primary Site panel lists Abdomen (1,019), Abdomen, Arm, Bladder, C... (570), Abdomen, Mediastinum (352), Abdomen, Pelvis (230), Adrenal (177), and Adrenal Glands (708). Below these is a Disease Type panel with a list of conditions and counts.

The main content area features a Filters bar with download and upload icons. Below it is a 'Download Table' button and a badge indicating '1,298,152 Imaging Series'. Three pie charts are displayed: Platform (Stanford AIMI, MIDRC, TCIA), Primary Site (Chest, No Data, Abdomen, Arm, Bladder, Chest, Head..., Pelvis, Prostate, Anus, Abdomen, Mediastinum, Whole Body, Thymus, Abdomen Cervix, Breast), and Disease Type (Lung Cancer, No Data, Pancreatic Ductal Adenocarcinoma, Lung Carcinoma, Squamous Cell Carcinoma, Uterine Carcinosarcoma, COVID-19).

A table of data is shown below the charts, with columns for Study ID, Study Description, Body Part Examined, Primary Site, Disease Type, and Platform. The table contains 10 rows of data, including study IDs like 1.3.6.1.4.1.14519.5.2.1.7009.9004.947472417000223653848235685369 and descriptions like NLSL-ACRIN and CT CHEST PULMONARY EMB.

At the bottom, there is a pagination bar showing 'Maximum of 10,000 Total Records, Records per Page: 10' and a page number '1' out of a total of 1000 pages.

Data Library

User managed list of data sets/cohorts for:

- Workspaces
- Exporting
- Sharing
- Applications

The screenshot displays the BioData CATALYST Data Library interface. At the top, there is a navigation bar with links for 'Submit Data', 'Documentation', and 'Login'. Below this is the NIH logo and the 'BioData CATALYST Powered by Gen3' branding. The main content area shows a list of datasets with columns for 'NAME', 'ID', '# FILES', and 'ADDITIONAL DATA SOURCES'. A 'Dataset 1' is expanded to show a table of files with columns for 'NAME', 'DESCRIPTION', 'TYPE', and 'SIZE'. The footer contains various policy and contact links.

NAME	ID	# FILES	ADDITIONAL DATA SOURCES
open_access-1000Genomes		1	False

NAME	DESCRIPTION	TYPE	SIZE
Serialized PFB created with test data from data-simulator		GA4GH_DRS	

Privacy Policy | Data Sharing Policy | HHS Vulnerability Disclosure | Freedom of Information Act (FOIA) | Accessibility | U.S. Department of Health & Human Services | National Institutes of Health | National Heart, Lung, and Blood Institute | USA.gov

Frontend Improvements

Greater data visualization capabilities

- Will model many improvements based on the GDC to allow interactive gene and mutation visualizations

Filters

Collapse All

Mutated Gene

Upload Genes

Somatic Mutations

Upload Somatic Mutations

Biotype

Name = Genes: 2

lncRNA

antisense_coding

transcribed_processed_transcript

In Cancer Gene Census

Genes: 714

VEP Impact

Mutations: 2

high

low

moderate

modifier

SIFT Impact

Mutations: 2

deleterious

deleterious_low_confidence

tolerated

tolerated_low_confidence

Polyphe Impact

Mutations: 2

benign

probably_damaging

damaging

unknown

Consequence Type

Mutations: 2

Genes Mutations

Distribution of Most Frequently Mutated Genes

Overall Survival Plot

S₁ (N = 94) TP53 Not Mutated Cases
 S₂ (N = 248) TP53 Mutated (GISTIC2) Cases
 Log Rank Test P-value = 1.2E-2
 Use the Survival History on the tabs below to change the survival plot.

TOTAL OF 714 Genes

Cohort	Survival	Symbol	Name	# SSM Affected Cases in Cohort	# SSM Affected Cases Across the GDC	# CNV Gain	# CNV Loss	# Mutations	Annotations
		TP53	tumor protein p53	542 / 1,577 (34.39%)	4,364 / 16,508 (26.44%)	48 / 1,393 (3.45%)	498 / 1,393 (35.75%)	348	
		ERBB2	epidermal growth factor receptor 2, erbB2	421 / 1,577 (26.70%)	1,885 / 16,508 (11.42%)	302 / 1,393 (21.66%)	32 / 1,393 (2.29%)	82	
		KRAS	KRAS proto-oncogene, GTPase	281 / 1,577 (17.83%)	1,763 / 16,508 (10.68%)	289 / 1,393 (20.74%)	199 / 1,393 (14.29%)	14	
		BRCA1	breast cancer 1, early onset	188 / 1,577 (11.98%)	2,959 / 16,508 (17.93%)	199 / 1,393 (14.29%)	576 / 1,393 (41.36%)	387	
		CDH1	cadherin 1	188 / 1,577 (11.98%)	490 / 16,508 (2.97%)	116 / 1,393 (8.33%)	498 / 1,393 (35.75%)	188	
		CDKN2A	Cyclin D1	154 / 1,577 (9.77%)	388 / 16,508 (2.35%)	294 / 1,393 (21.17%)	142 / 1,393 (10.19%)	94	
		MGMT	O6-methylguanine-DNA methyltransferase 2	121 / 1,577 (7.67%)	1,963 / 16,508 (11.89%)	256 / 1,393 (18.38%)	181 / 1,393 (12.99%)	148	
		MDM2	cellular myelocytomatosis 2	102 / 1,577 (6.47%)	385 / 16,508 (2.33%)	163 / 1,393 (11.69%)	253 / 1,393 (18.16%)	148	

GEN3
Browser Data & Documentation | Login

Discovery
Dictionary
Exploration
Query
Analysis
Workspace
Profile

Summary Charts

Age at Index

Sex

Race

Ethnicity

Age at index Bar Radar Statistics

Sex Bar Radar Statistics

MEAN ABSOLUTE ERROR (MAE) 0.026
 ROOT MEAN SQUARE ERROR (RMSE) 0.036
 PEARSON CORRELATION COEFFICIENT 10000
 BOMBIER CHI-SQUARE STATISTIC 0.0308
 CHI-SQUARE CONTRIBUTIONS BY GROUP

Group	Contribution
Female	0.0005
Male	0.0007
Not reported	0.0297

Race Bar Radar Statistics

Ethnicity Bar Radar Statistics

Center for Translational Data Science

- Machine readable document provided by a data commons to a data mesh (and vice versa) in order to participate in the mesh.
- Example mesh card:

```
{
  "card_type": "mesh",
  "meshcard_version": "1.0.0",
  "description": "description of data mesh",
  "data_type": "imaging",
  "usage_endpoint": "/usage",
  "metadata_endpoint": "/mds/metadata",
  "metadata_valid_apis": ["data_connect", "gen3_mds"],
  "auth_valid_apis": ["passports", "gen3_fence"],
  "data_valid_apis": ["DRS", "gen3_indexd"],
  "nodes": [
    {
      "card_type": "node",
      "id": "example-node",
      "meshcard_version": "1.0.0",
      "description": "description of a node in the mesh",
      "node_endpoint": "example.com",
      "metadata_api": {
        "standard": "data_connect",
        "endpoint": "example.com/mds/metadata/"
      },
      "auth_api": {
        "standard": "passports",
        "endpoint": "example.com/ga4gh/passports/"
      },
      "data_api": {
        "standard": "DRS",
        "endpoint": "example.com/ga4gh/drs/"
      }
    }
  ]
}
```

- Example Node Card:

```
{
  "card_type": "node",
  "id": "example-node",
  "meshcard_version": "1.0.0",
  "description": "This would be the description of a node in the mesh",
  "node_endpoint": "example.com",
  "metadata_api": {
    "standard": "data_connect",
    "version": "1.0.0",
    "endpoint": "example.com/mds/metadata/"
  },
  "auth_api": {
    "standard": "passports",
    "version": "1.0",
    "endpoint": "example.com/ga4gh/passports/"
  },
  "data_api": {
    "standard": "DRS",
    "version": "1.5",
    "endpoint": "example.com/ga4gh/drs/"
  }
}
```

Other topics from the community

Gen3 Panel Discussion

Questions