Gen3 2026 Development Roadmap

Gen3 Community Forum
January 21st, 2026

Meeting Agenda

- Welcome & Announcements
- Community Collaboration Request - Luca Graglia
- 2025 Product Roadmap Review
- Preview: 2026 Product Roadmap
- Steering Committee Discussion
- Open Q&A Session

Community Collaboration Request - Limitation of Loading Data in GEN3 with Sheepdog

- 46k subjects with a total of 1.85 millions lines/rows across 384 files occupying 214MB. This takes about 3 to 4 days to load so about 8 to 10 records per seconds.
- Time per record is hard to report because it depends on the width of the record

Main problems:

- The number of sheepdog instances doesn't scale up under load in the deployment (need to change it manually).
- The current transaction system is slow.

Contact info: Luca (lgraglia@uchicago.edu)

2025 Product Roadmap Review

Summary of 2025 Roadmap

Topic	Status	Description
Front End Framework (FEF) roll out	Mostly complete	In production and continue to add new features
Task Execution Service (TES) Support	Mostly complete	Q1 target for internal production use, Q2-3 target for general use
Improved Documentation	Complete	Migrated to docs.gen3.org
Planned Frontend Improvements	Mostly complete	Full status provided in slide 11
Mesh and Node Cards	Partially complete	Full status provided in slide 15
Commons Services Operations Center (CSOC) Support	Prototyped	Full status provided in slide 17
Integrated AI capabilities	Partially complete	Full status provided in slides 19-21

Progress Updates on Last Year's Improvement Proposals

Topic	Status	Description
Airlock or Data Enclave feature set	In progress	Adding additional controls to keep data from leaving a Gen3 workspace
Native Graph Database	No resources to start	There may be some advantage to working with a native graph database like neo4j
Gen3 in a Box	No resources to start	Completely self-contained Gen3 platform tied to a computer. May be critical for indigenous peoples or other sensitive data that must not be transferred
Stand Alone Schema Validation	Complete (by community)	Ability to review dictionary schemas without the need to set up an entire Gen3 system

Outcomes Beyond the Initial Plan

Topic	Status	Specific Tools
Analysis Tooling (part of Front End Framework, FEF)	Complete	GDC V2's Cohort Builder
		Clinical Data Visualization
		Cohort Comparison
		Set Operations
		Mutation Frequency
		Repository with cart support

Preview: 2026 Product Roadmap

Preview: 2026 Product Roadmap

1. DRS 1.5 Update
2. Enhancing PFB Capabilities
3. Gen 3.2 Front End Framework (FEF) Improvements
4. Workspace Improvements
5. Backend support for Cell Level Data Access
6. Mesh Improvements
7. CSOC Dashboard
8. FHIR Server
9. AI Capabilities (embedding service; model repo service; inference engine)
10. Data Access Committee Integration in the Mesh
11. Areas Under Evaluation for Further Investment

Data Repository Service (DRS) 1.5 Update

- Updating DRS 1.2 → 1.5. Implementation of the GA4GH DRS 1.5 specification by adding the required new endpoints in Gen3 and updating existing endpoints in Gen3.
- The upgrade contains:
  - Bulk retrieval; Request and resolve multiple DrsObjects at once.
  - Authorization info; Know which tokens/issuers are needed for access.
  - cloud /region awareness; See where data is stored for optimal compute placement.
  - Availability status; Check if data is online, in cold storage, or unavailable.

Enhancing PFB Capabilities

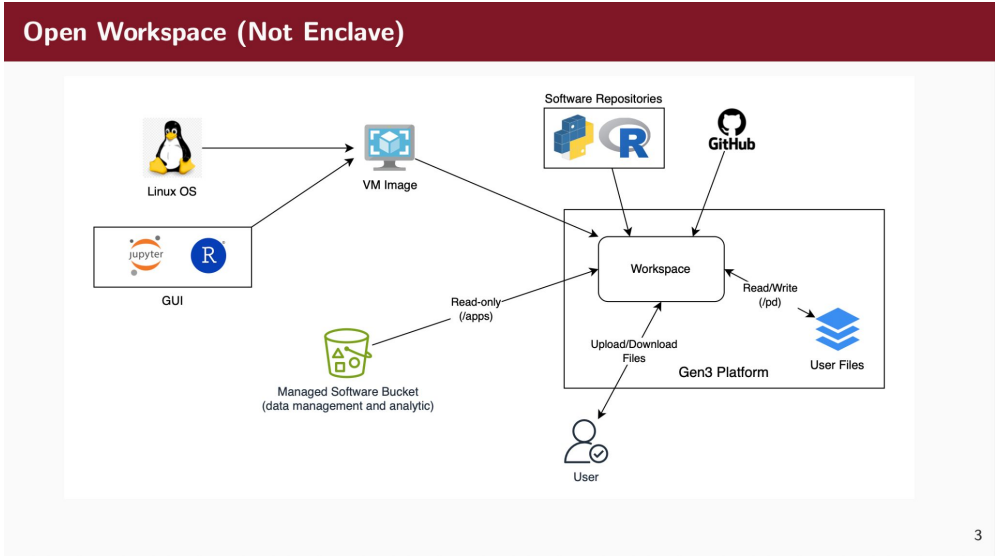
- Implement Portable Format for Bioinformatics (PFB) Search Functionality
  - Serialized file format used to transfer both dictionary schema and graph data content
  - Heavily used by the BioData Catalyst (BDC) ecosystem
  - Will develop new capabilities to search variables and aggregations within PFBs
  - Variables and aggregations for PFBs available for export will be searchable from Gen3 Discovery
  - Supports multiple Gen3 Data Models simultaneously
- To enhance capabilities, Gen3 will also focus on resolving and handling nested PFBs, streamlining access to complex hierarchical datasets.
- Enhance PFB schema flexibility to support integration with FHIR standards.

Gen 3.2 Front End Framework (FEF) Improvements

- Gen3.2 Front End Framework (FEF)
  - 2025 brought new features and improved user experiences, 2026 priorities include
    - Update deployment documentation
    - Enable migration of all projects to FEF
    - Improving developer and operator experience
      - Faster hot reloads & config validation
      - Storybook for components & accessibility
    - New workflow/pages for data submission
    - FHIR support (See slide 18)
    - Expanded Gen3.2 applications catalog
    - Jupyter-lite notebooks & UI for AI models/embeddings

Workspace Improvements

- Workspace software repository
  - Data management and analytic software stored in a centralized repository, using a similar model to HPC environments
  - Benefits
    - Large library of software makes workspaces more capable
    - Provides multiple versions
    - Persistent preferences and settings
    - Smaller image may help to provide faster start up



Workspace Improvements

- Trusted execution environment
  - Adding additional isolation to existing workspaces to prevent data from leaving the closed analysis environment
  - Controlled environment workspaces will be created that will prohibit users from:
    - Downloading and/or uploading files to/from a local computer
    - Accessing the internet from within a workspace (with limited exceptions as necessary)
    - Installing software (limited to administrators)
- Support for TES and WES
  - Finalize TES implementation using Funnel
  - Provide a GUI to manage workflows
  - Integrate cost tracking
- Simplify file download into workspaces

Cell-Level Data Access (Backend Support)

- Cell level access is a granular permission system that allows users to request and obtain authorization for a specific subset of files within a larger dataset. This capability ensures targeted and secure data sharing with external repositories/ partners.
- Key Features
  - Selective File Permission
    - Users can request access to specific files, rather than the entire dataset, or rows in the dataset.
    - Permissions maintained at the file/subset level, streamlining collaboration.
  - User-Friendly Request Process
    - Users submit access requests for chosen files and is informed throughout the approval process.
  - Efficient Authorization Handling
    - Administrative review and management of requests.
    - Notification sent upon approval.
  - Secure, Time-Limited Access
    - Access is granted via tokens, ensuring authorization is temporary and controlled.

Mesh Improvements

Mesh and Node Cards are machine readable configuration files to enable automatic joining of nodes to a hub in a mesh. Prototype created in 2025. In 2026 we will work towards a GA4GH standard and a production example.

Mesh Card Example

```
{
    "card_type": "mesh",
    "meshcard_version": "1.0.0",
    "description": "This is a data mesh for imaging files",
    "data_type": "imaging",
    "usage_endpoint": "/usage",
    "data_governance": "data can leave boundaries",
    "metadata_endpoint": "/mds/metadata",
    "metadata_valid_apis": ["data_connect", "gen3_mds"],
    "auth_valid_apis": ["passports", "gen3_fence"],
    "data_valid_apis": ["DRS", "gen3_indexd"],
    "nodes": [
        "example.com/node-card",
        "example2.com/node-card"
    ]
}
```

Node Card Example

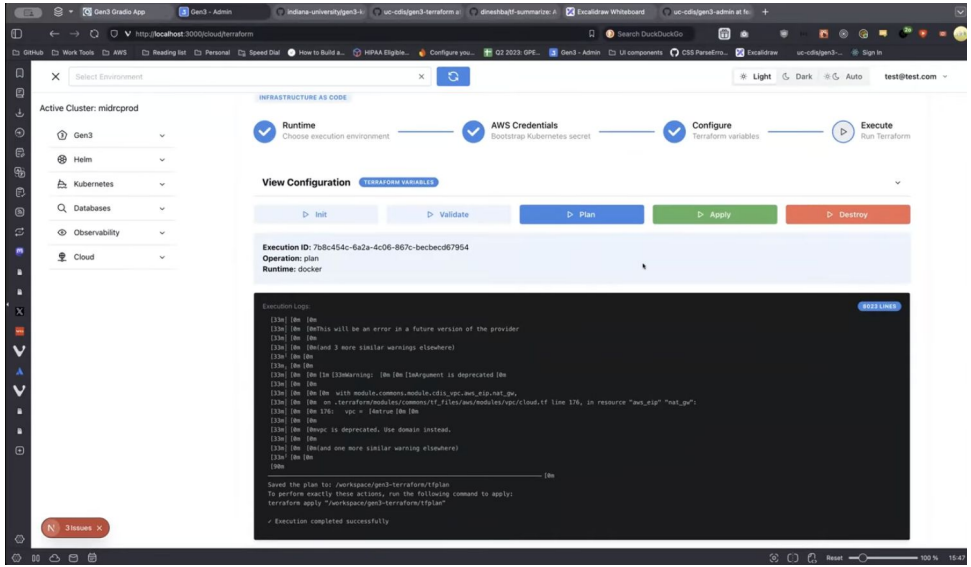
```
{
    "card_type": "node",
    "id": "example-node",
    "meshcard_version": "1.0.0",
    "description": "This would be the description of a node in the mesh",
    "node_endpooint": "example.com",
    "metadata_api": {
        "standard": "data_connect",
        "version": "1.0.0",
        "endpoint": "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/"
    }
}
```

Mesh Improvements

- We plan to deliver a new scalable solution for Data Nodes to send updated mesh data to a Data Hub
- Currently testing a **pub/sub** (publish/subscribe) solution

CSOC Dashboard

- Significant progress in 2025 as part of the CSOC WG
- In 2025
  - Prototype dashboard for setting up infrastructure (using terraform), deploying Gen3 (using Helm), and monitoring deployment
- In 2026
  - Productionalize this dashboard and begin using it ourselves
  - Collaborative plugin development with community will continue



FHIR Server

- Start productionalizing the FHIR Proof of Concept (PoC)
  - Refactor and harden the initial PoC to meet community wide standards.
  - Enhance scalability, reliability, and security for production use.
  - Support Gen3 Authorization applied over FHIR resources.
  - Design and execute integration test suites covering cross-system interactions.
  - Validate data flow and API compatibility between Gen3 and FHIR endpoints.
- Start Implementing Scalable FHIR Search Functionality
  - Support FHIR-compliant search operations.
  - Optimize query performance for large datasets.
  - Build an intuitive UI for browsing and querying FHIR resources.
- Start Implementing FHIR Data Submission Tooling
  - Provide tools for submission of FHIR-compatible data.
  - Document submission procedures and troubleshooting guides.

** Due to the scale of this initiative, we are still scoping the deliverables for 2026

AI Capabilities

- We are working to roll out integrated AI capabilities in Gen3
- These will be added to the M3 AI Commons as they become available
- See the following slides to get a sense of expected features in 2026
- Note that these are projected timelines, that we will be able to update during the year

AI Capabilities

Items/Features	Anticipated 2026 Q1	Anticipated 2026 Q2	Anticipated 2026 Q3	Anticipated 2026 Q4
Refine embeddings in batch through API	Arborist and fence are currently being re-engineered to handle batch embeddings. New version of batch embeddings service has been deployed to AI Commons, which is ~2x faster for testing and continue refinement.	– Continue refinement of batch embeddings to increase speed – Align with diverse embeddings formats and data modalities, particularly proteomics and radiology images	– Develop and test an ingestion service for users to share their own embeddings – Develop security parameters for controlled-access embeddings (may need to be addressed earlier in Q2)	– Test how embeddings may be pulled across Gen3 platforms for model sharing and federated training/inferencing – Potentially interoperate with containerized workflow services (TES/Nextflow)
AI commons applications	Address bugs and feature issues within the AI Commons applications - QAG, Cohort Copilot, and Data Model Generator		– Explore how QAG and LaB-RAG may be leveraged by others – Assess if any additional AI applications should be hosted in AI Commons (particularly AI curation) – Need to publish QAG and LaB-RAG publications	
Embedding visualization and search	Support embedding visualizations FEF	User flow, UI/UX search and discovery use-case definitions	Design embeddings explorer page in FEF	
Federated learning and data visitation services	Define and develop federated learning use-case leveraging Gen3 data commons/mesh data sources, embeddings and model deployed in AI Commons, and notebook shared in AI Commons workspace.	Define data visitation use-cases feasible within AI Commons	Align with user-provided embeddings service and MESH cards	
GPU on prem	Support deployment of on-prem version of Gen3			

AI Capabilities

Items/Features	Anticipated Q1 2026	Anticipated Q2 2026	Anticipated Q3 2026	Anticipated Q4 2026
Model Inference Services	Pilot version of LaB-RAG inference workflow is deployed in Gen3 AI Commons notebooks. Continue refinement and deploy existing Gen3 AI Inferencing service in AI Commons https://github.com/uc-cdis/gen3-discovery-ai/			– Scale inferencing workflow to align with emerging standards and technologies (TES, Nextflow) to embed small to midscale model inferencing while balancing rapid response.
Model Training Services	Need close collaboration with GPU teams Will focus on model training for CTDS developed models initially			Leverage on-prem GPUs and small to midscale multi-modal models in a federated data mesh empowering federated training approaches leveraging with 1) Gen3 cloud-based computational workspaces; 2) integration with workflow execution services (Nextflow, TES)
Model Repository Services	Leveraging Huggingface SafeTensors opensource format for model repository services. Need to define what Gen3 approaches are needed to justify move away from Huggingface. For instance, FedRAMP certification and additional security parameters.	Test and deploy Gen3 Model Repository Services	Continue to develop, test, and deploy model repository service through Q3 and Q4	Extend to diverse embedding formats and security parameters as needed
Agentic Interfaces	Explore alignment with MCP and use-cases	Scaling Query Augmented Generation (QAG) & Cohort Co-Pilot apps for diverse disease categories and use-cases –Expand underlying disease ontologies and data types for multi-modal use-cases	Refine UX/UI user journeys for multiple Gen3 Platforms	

Data Access Committee Integration in the Mesh

Goal:

- Enable secure and auditable review of data access requests for partner-held datasets.

Solution:

- Build a centralized interface for Data Access Committee (DAC) review.
- Users submit requests; DAC reviews and approves/denies.
- Integrate into Gen3 authorization flow.

Technology:

- Evaluating REMS (Resource Entitlement Management System) and other alternatives.

Benefits:

- Transparent, compliant approvals.
- Streamlined collaboration across partners.

Areas Under Evaluation for Further Investment

- Re-evaluating our Elasticsearch-based search backend and assessing StarRocks as a potential replacement to enhance search performance and scalability
- Enhanced tools and features for seamless integration and analysis of OMOP data
- Keycloak as an Authentication/Authorization tool
- Increasing Community Access to Gen3 with OpenShift
  - Enterprise-grade Kubernetes platform enables secure, scalable, and user-friendly on-prem Gen3 deployments
  - Streamlines onboarding, management, and collaboration. Enables easier launch of new commons by community members and provides a path to share data commons easily, with strong security and governance
- Advanced networking with Cillium
  - Integrating Cilium with Gen3 on Kubernetes has the potential to significantly enhance security, network visibility, and real-time monitoring, while also optimizing pod networking

Steering Committee Discussion

Open Q&A Session