EPHI National - Genomics Data Engineering Collaborative Research Project.

TOR:

Prepared By:

Mihret Kitaw

Frew Legese

November 5, 2025

Addis Ababa,
Ethiopia

Acronyms:

FAIR: Findable, Accessible, Interoperable, Reusable

NDMC: National Data Management Center

RTDS: Research Tracking Database System

DMP: Data Management Plan

DOI: Document Object Identifier

IRB: Institutional Review Board

PDF: Portable Document Format

PhD: Doctor of Philosophy

MSc: Master of Science

HTTPS: Hypertext Transfer Protocol Secure

SFTP: Secure File Transfer Protocol

API: Application Programming Interface

PII: Personally Identifiable Information

FASTQ: A common plain‑text format for storing raw sequencing reads and their quality scores.

BAM: Binary Alignment/Map

VCF: Variant Call Format

RBAC: Role-Based Access Control

IAM: Identity and Access Management

ETL: Extract, Transform, Load

DB: Database

BI: Business Intelligence

CI/CD: Continuous Integration / Continuous Delivery

DR: Disaster Recovery

Table of contents

1. Introduction

1.1. Overview of the Genomics and Bioinformatics Core Facility of EPHI

The Genomics and Bioinformatics Core Facility at the Ethiopian Public Health Institute (EPHI) delivers advanced genomic sequencing and bioinformatics analysis services to a diverse user base, including external collaborators, grant providers, universities, PhD students, hospitals, and EPHI researchers for public health research and surveillance. The facility generates substantial data volumes, ranging from tens of gigabytes (Gb) to several terabytes (Tb) per sequencing run, with cumulative outputs reaching petabyte-scale annually due to its high-throughput sequencing capabilities.

The appropriate and digitalized data management provides a comprehensive framework for managing data across its lifecycle, ensuring compliance with national and international standards, including EPHI's National Health Data Access and Sharing Guideline 2021. Furthermore, it promotes Findable, Accessible, Interoperable, Reusable (FAIR) data principles, ensures reproducibility, and fosters equitable access for all users. It also aims to ensure that all genomic and bioinformatics data are handled with integrity, accuracy, and transparency while safeguarding ethical, legal, and security standards. Ultimately, this supports EPHI's mission to advance data-driven public health research, promote collaboration among national and international partners, and strengthen Ethiopia's capacity in genomics and bioinformatics for long-term sustainability and global health impact.

1.2. Project overview

This project will establish a centralized, secure, and intuitive website application to serve as the digital backbone for the EPHI Genomics and Bioinformatics Core Facility. The platform is designed to transform the data lifecycle by integrating key project initiation documents (DMP and Project Information Form) and ethical clearance verifications directly into the digital workflow. It will enable service utilizers to seamlessly register projects, create standardized metadata, and manage secure, tiered access to their genomic data and analysis results. By automating data ingestion and ensuring seamless integration with the National Data Management Center (NDMC) Research Tracking Database System (RTDS), the project guarantees data sovereignty, promotes reproducibility, and creates a sustainable, FAIR-compliant national resource for advanced genomic big data analysis.

1.3. Purpose:

The purpose of this project will be delivering a fully functional and fully fledged system to automate the office Genomics department operation and data management, and utilization practices. This section of the document will help understand what is the main purpose of the development and implementation of the GenDE platform for EPHI. Some, but not limited to the following are the purpose of implementation of the GenDE platform.

- To establish a centralized, website application digital platform that Automates the archival, secure storage, and management of genomic data and its associated metadata.
- Streamlines project onboarding by digitally capturing the Data Management Plan (DMP) and Project Information Form.
- Empowers users to create and manage standardized metadata for their sample data.
- Implements a controlled access model for raw data and analyzes results based on user roles and data sensitivity.
- Ensures data sovereignty and long-term sustainability of the data and metadata by linking the system with EPHI NDMC's RTDS for safe storage in standardized repositories of EPHI-NDMC.

- Integrates the web application system seamlessly with RTDS of EPHI for secure data sharing according to EPHI’s National Health Data Access and Sharing Guideline (2021) to enable advanced secondary analysis.

1.4. Project Scope:

- Development of a web portal for user registration, authentication, and project initiation.
- Digitalization of the Data Management Plan (DMP) and Project Information Form as core components of project registration.
- API-based integration with the EPHI IRB system for real-time ethical clearance validation.
- Dynamic metadata creation interfaces for projects and samples.
- A data access control system along with a secure data request and approval workflow.
- Automated pipelines for collecting data from sequencers and bioinformatics tools, plus secure archiving to the NDMC RTDS.
- Dashboards for data discovery (with limited public metadata), project tracking, and administrative oversight.

1.5. Benefits of the project

- Streamlined Workflow: Complete digitalization from project proposal to data archiving.
- Improved Compliance: Automated enforcement of DMP and ethical review requirements.
- Strong Data Governance: Ensures data lineage, origin tracking, and adherence to FAIR principles.
- Faster Research: Supports secure data sharing and discovery for advanced secondary analysis.
- National Data Sovereignty: Strengthens EPHI's role as the central hub for genomic data through seamless NDMC RTDS integration.

2. Existing work setting (AS-IS)

As-Is part of this document will describe how the current operation is performed by the office, as well as the tools, settings, and standard operation followed by the stakeholders to achieve the Genomics data management.

2.1. Existing system

The existing system is under the Genomic department, has been undertaken by manual and tiresome practices, which led the office operation to the data silos, and well-structured governance for the genomic data. Some of the practices AS-IS are:

- ❖ Project Initiation: Reliance on static, paper-based, or PDF versions of the DMP and Project Information Form leads to scattered information and manual data entry.
- ❖ Ethical Compliance: Manual, offline verification of IRB certificates causes administrative delays and increases the chance of errors.
- ❖ Metadata & Data Management: Fragmented use of Excel spreadsheets and manual file transfers results in inconsistent metadata, data silos, and difficulties in tracking data lineage, as well as a lack of transparency.
- ❖ Data Access & Sharing: The absence of a standard process for requesting and granting data access hinders collaboration and secondary use.
- ❖ NDMC Integration: No automated link between the core facility's bioinformatics data outputs and the national RTDS for archiving and cataloguing.

2.2. Problem statement.

The current process for managing genomic data at EPHI is fragmented and manual, leading to significant inefficiencies and risks of data loss, causing bottlenecks and potential for error.

- ★ Disjointed Project Initiation: Reliance on offline, paper-based static forms (Data Management Plan, Project Information) creates data silos and manual entry burdens.
- ★ Poor Data & Metadata Management: Metadata is captured in inconsistent formats (e.g., Excel spreadsheets), decoupled from raw data, and hinders findability and reuse.
- ★ No Automated Archival: Lack of integration with the NDMC's Research Tracing Data System (RTDS) prevents systematic data preservation, cataloguing, and secondary use.

- ★ Insecure Data Access for both the service user and provider: The data-sharing methods lack audit trails and controlled, tiered access, compromising security and governance.

2.3. Focus areas

This project aims to address the following areas, but is not limited to:

- ❖ Workflow Digitalization: Creating a seamless digital workflow from project registration (including DMP and project information) and IRB validation to metadata submission and data archival, and sharing.
- ❖ System Integration: Ensuring seamless interoperability with key systems, specifically the EPHI IRB for ethical clearance and the NDMC RTDS for long-term data preservation and use.
- ❖ Data Governance & Security: Implementing a robust, tiered data access model and a formal request/approval process to ensure secure and compliant data sharing.
- ❖ FAIR Data Implementation: Enforcing standardized metadata creation and management to make data Findable, Accessible, Interoperable, and Reusable (FAIR).

3. Methodology

3.1. Overview:

The proposed system is a centralized, secure, web-based application platform that acts as the digital backbone for the Genomics Core Facilities. It will transform the data lifecycle by integrating project initiation, ethical review, and metadata creation into a single, streamlined workflow. The system automates the Data Management Plan (DMP), Project Information Form, metadata development, and data archiving while providing controlled, tiered access to data for authorized users, ensuring compliance with EPHI guidelines and promoting advanced secondary genomic and bioinformatics data analysis.

3.2. System Architecture

The proposed system will have the following core components and processes with four layers. Core layers of the system are an ingest layer, a validation and de-identification layer, a storage layer, a processing and catalog layer, and an access and presentation layer, all governed by consent, RBAC, encryption, and audit logging. Below is a detailed brief of each layer describing responsibilities, key components, and operational expectations.

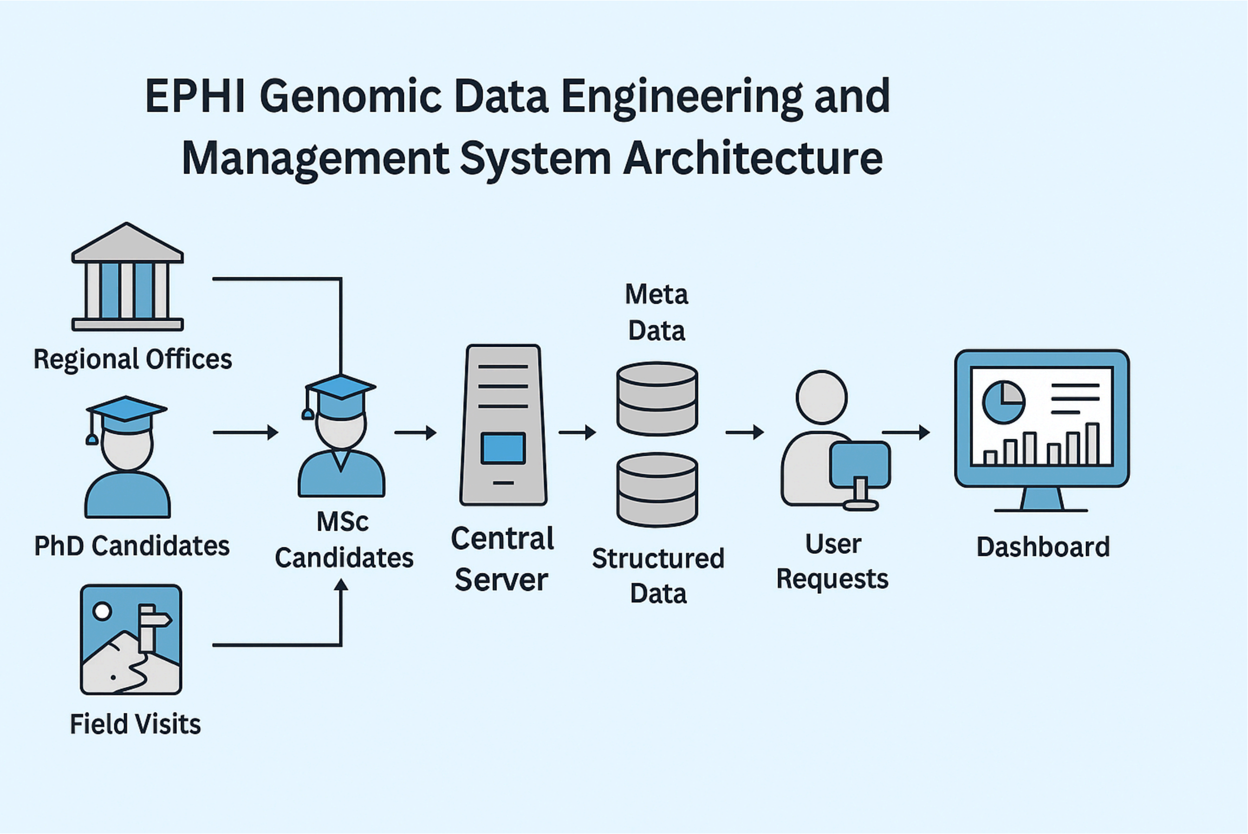

Fig.1: Architecture of the system

3.2.1. Ingest Layer

The ingest layer is the system's front door and must securely accept data from regional offices, PhD and MSc students, and field visits. It provides multiple secure transfer methods such as HTTPS uploads, authenticated APIs, and SFTP endpoints, and captures standardized metadata at submission so every file arrives with provenance, submitter identity, collection context, and consent flags. This layer also implements immediate syntactic validation to reject corrupt or incomplete files and returns structured ingestion receipts that feed the downstream pipeline. For operational resilience, it supports resumable uploads, rate limiting, and automated quarantine of suspicious submissions.

3.2.2. Validation and De‑identification Layer

After the ingest, data flows into a validation and de‑identification service that enforces consent rules and privacy policies before any raw genomic data is persisted. This layer runs schema checks, checksum verification, and automated PII detection; it applies deterministic or probabilistic de‑identification transforms and records the exact transformation steps as immutable provenance metadata. Consent enforcement is policy-driven so that datasets retain flags indicating permitted uses, retention windows, and sharing restrictions. All actions here are logged for audit and legal review to ensure traceability of every change to the original submission.

3.2.3. Storage Layer

The storage layer separates concerns between raw object storage for large genomic files, a structured relational metadata database for sample and consent records, and a processed data lake for analysis-ready artifacts. Raw files are stored in an encrypted object store with immutable snapshots and lifecycle policies, while the metadata database stores normalized sample, project, and provenance tables to enable fast discovery and policy evaluation. The processed data lake holds curated VCFs, aggregated tables, and derived features optimized for query engines and analytics. This separation supports cost optimization, scalable compute, and reproducible reprocessing workflows

3.2.4. Processing and Catalog Layer

Processing is implemented as containerized ETL pipelines and serverless functions that transform raw files into standardized, analysis-ready formats while recording lineage. The data catalog provides searchable discovery, dataset versioning, and lineage visualization so researchers can understand dataset provenance and consent constraints before requesting access. Automated quality metrics and validation reports are attached to catalog entries to accelerate reuse and reproducibility. Where appropriate, notarization or tamper-evident checksums are used to strengthen integrity guarantees

3.3. Deliverables.

For the successful completion of this project, it is expected for the following deliverables are implemented, but not limited to the following.

- ➔ A fully operational genomic data management web application portal.
- ➔ Digitally integrated Data Management Plan (DMP) and Project Information Form modules.
- ➔ Technical design documentation and API specifications.
- ➔ User Training Materials and Administrator Guides.
- ➔ Trained administrative and user teams from EPHI and partner institutions.

3.4. System Functionalities

The system functionality will be varied according to the changes in the requirements and needs which would be arized from the user until the project is deployed. The minimal system functionalities identified at this stage are listed below.

- ➢ Digital project onboarding: Dynamic forms for DMP and Project Information, integrated with IRB validation.
- ➢ Metadata management: User-friendly interfaces for creating and managing sample and project metadata.
- ➢ Data ingestion pipeline: Automated registration and transfer of sequencing data from core facility instruments to users.
- ➢ Tiered access & data request workflow: Configurable access levels with an online request system that has audit trails.
- ➢ NDMC RTDS Connector: Automated service to package and transfer finalized datasets and metadata to the EPHI central server/repository after achieving its primary objective.
- ➢ Data Discovery Catalog: Public-facing search portal for project and sample metadata (where approved).
- ➢ Dashboard & Reporting: Real-time views of project status, storage metrics, and system usage for users and admins.

3.5. Non-Functional Requirements

- ➢ Security: Compliance with national data laws, end-to-end encryption, and strong authentication and authorization.
- ➢ Interoperability: Adherence to standard metadata schemas (e.g., MIxS) and APIs for seamless integration (IRB, NDMC RTDS).
- ➢ Scalability & Performance: Architecture designed to manage large-scale data growth and high user demand.
- ➢ Usability: Intuitive interface for users with varying levels of bioinformatics expertise.
- ➢ Reliability: High availability to support essential research activities.

3.6. Use case

Actor: EPHI and external researchers

- ➔ Logs into the portal and starts a new project.
- ➔ Completes the digital Project Information Form (PIF) and Data Management Plan (DMP), specifying data types, volume, and sharing intentions.
- ➔ Enters the IRB approval code; the system automatically verifies its validity and status with the IRB database.
- ➔ Verification and validation by the genomic core facility center experts.
- ➔ Once verified, the researcher registers 50 samples using the standardized metadata creation tool.
- ➔ After sequencing, the core facility's automated pipeline collects the FASTQ files, linking them to the registered samples.
- ➔ The researcher accesses preliminary analysis results via the portal, with access adjusted to their role.
- ➔ Upon project completion, the system automatically packages the final dataset and metadata and transfers them to the NDMC RTDS for long-term archival and national cataloging.

3.7. Third-party integration requirement

The proposed system will be interacting with some of the following (but not limited to) internal and external systems.

3.7.1. EPHI IRB System: For ethical clearance validation.

3.7.2. EPHI NDMC: As the primary endpoint for data archival, cataloging, and persistent identification (DOI minting).

3.7.3. DataCite API (via NDMC): For registering datasets and obtaining DOIs.

3.8. Resource requirement

The minimal resource requirement for this project to come to life, the following, but not limited to are listed below:

Human: Project Manager, Bioinformatician, Full-Stack Developers, DevOps/System Administrator.

Technical: Development & Staging environments, Production Server/Cloud Infrastructure, Secure Object Storage.

Collaborative: Dedicated liaisons from the IRB and NDMC for API specification and testing.

3.9. Stakeholders' engagement

The engagement of the actors in the successful implementation of this project can be listed as follows:

Primary Lead: EPHI Genomics and Bioinformatics Core Facility.

Key Integrators: EPHI NDMC (for RTDS integration) and EPHI IRB (for ethical clearance integration).

End-Users: Representatives from collaborating universities and hospitals.

Advisory Board: Senior scientists from EPHI, Jimma University, Mekelle University, and Hawassa PHI.

3.10. Expectation or outcome:

At the end of the project implementation and deployment, the successful establishment of a national, integrated genomic data platform that significantly improves research efficiency, data quality, and collaboration will be expected. The platform will be the trusted entry point for all projects using the core facility, ensuring that all genomic data generated is FAIR, ethically compliant, and sustainably archived within the national research infrastructure (NDMC RTDS).

3.11. Implementation strategy

Phase 1:

Foundation (1-4 months): Develop core platform, user management, and digital DMP/project info forms.

Phase 2:

integration (months 5-8): implement IRB and metadata modules; build the NDMC RTDS connector.

Phase 3: access & automation (months 9-11): develop the data access workflow and data ingestion pipelines.

Phase 4: deployment & handover (months 12-13): conduct UAT, user training, system go-live, and project handover.

4. Timeline and Budget

Timeline

This project will be a 12-month project timeline culminating in a production-ready system and comprehensive knowledge transfer while ensuring scalability.

Budget

The budget will incorporate personnel costs, software & infrastructure (servers, storage, and licenses), and training & capacity-building workshops.

5. Conclusion:

This project is a key investment in Ethiopia's public health and research infrastructure. By creating a unified digital platform that links important processes from project conception to genomic and bioinformatics data archiving, this will empower researchers, ensure compliance with regulations, and build a lasting resource for genomic big data that will foster public health innovation for years.

References