

**Terms of reference for EPHI National Genomics Data Engineering and Machine learning Collaborative research project.**

**Prepared By:**

Mihret Kitaw

Frew Legese

December, 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 4](#_Toc219127642)

[1.1. Overview of the Genomics and Bioinformatics Core Facility of EPHI 4](#_Toc219127643)

[1.2. PURPOSE AND BACKGROUND 4](#_Toc219127644)

[1.3. RESEARCH OBJECTIVES 5](#_Toc219127645)

[General Objectives: 5](#_Toc219127646)

[Specific Objectives 5](#_Toc219127647)

[1.4. RESEARCH QUESTIONS 6](#_Toc219127648)

[2. FRAMEWORK AND APPROACH 6](#_Toc219127649)

[2.1. Project overview 6](#_Toc219127650)

[2.2. Purpose and benefit: 7](#_Toc219127651)

[2.3. Project Scope: 7](#_Toc219127652)

[2.4. Benefits of the project 8](#_Toc219127653)

[3. Existing work setting (AS-IS) 8](#_Toc219127654)

[3.1. Existing system 8](#_Toc219127655)

[3.2. Problem statement. 9](#_Toc219127656)

[3.3. Focus areas 9](#_Toc219127657)

[4. Methodology 10](#_Toc219127658)

[4.1. Overview: 10](#_Toc219127659)

[4.2. System Architecture 10](#_Toc219127660)

[4.3. Process work flow 12](#_Toc219127661)

[4.4. Deliverables 14](#_Toc219127662)

[4.5. System Functionalities 14](#_Toc219127663)

[4.6. Non-Functional Requirements 15](#_Toc219127664)

[4.7. Third-party integration requirement 15](#_Toc219127665)

[4.8. Resource requirement 15](#_Toc219127666)

[4.9. Expectation or outcome: 16](#_Toc219127667)

[4.10. Implementation strategy 16](#_Toc219127668)

[5. Timeline and Budget 16](#_Toc219127669)

[References 18](#_Toc219127670)

**Table of figures:**

[Figure 1. Architecture of the system 11](#_Toc219127566)

[Figure 2. Process workflow. 13](#_Toc219127567)

# Introduction

## 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[1].

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.

## PURPOSE AND BACKGROUND

The rapid growth of genomic data has created unprecedented opportunities for public health research, but it also poses significant challenges in terms of data engineering, quality management, and analytical scalability. Modern approaches increasingly integrate artificial intelligence (AI) and machine learning (ML) into genomic data pipelines, enabling automated data processing, intelligent pipeline optimization, and advanced quality monitoring[2]. Studies highlight that AI-driven systems can reduce manual data cleaning efforts by up to 45% and improve diagnostic accuracy in healthcare genomics by more than 30% [3] . Furthermore, machine learning–based governance frameworks have demonstrated enhanced compliance monitoring and security, reducing data-related incidents by over 40%[3]. For institutions such as the Ethiopian Public Health Institute (EPHI), adopting genomic data engineering practices combined with ML-driven analytics is essential to strengthen national health research capacity, improve disease surveillance, and ensure interoperability with global genomic standards. This project therefore seeks to establish a robust foundation for genomic data engineering while adopting a machine learning to unlock actionable insights for public health decision-making.

One of the primary advantages of automation in genomic data engineering is the increased efficiency it brings to data handling [4]. Students and researchers often work with vast amounts of data that require cleaning, transformation and analysis. Automation tools such as ETL (Extract, Transform, Load) frameworks facilitate these processes, allowing users to set up workflows that can run independently [5]. This not only saves time but also ensures that data is consistently processed in an accurate manner. Mastering these tools gives students and researchers a competitive edge, preparing them for the demands of the job market where data skills are increasingly sought after [5].

## RESEARCH OBJECTIVES

# General Objectives:

* To build an effective and resilient data engineering platform to manage, acquire and share the genomic metadata, and sharing a sequencing result.
* To build a foundation of data engineering for a genomic data analysis, machine learning and AI agents inside EPHI.

# Specific Objectives

* To establish a centralized web application through digital platform that automates the archival, secure storage, and management of genomic data and its associated metadata while enforcing the minimal metadata standards.
* To digitize and automate the project initiation and governance workflow by developing a centralized online portal for digital Project Information Form (PIF), and integrated Data Management Plan (DMP) submission.
* To implement real-time API integration with the EPHI Institutional Review Board (IRB) system for automated ethical clearance validation.
* To develop a secure, role-based digital data and metadata access and sharing hub that automates controlled access to raw data and analytical results.
* To develop automated pipelines for the ingestion, registration, real-time status for the service requests, analyses, and transfer of sequencing data from core facility instruments to designated user workspaces that enhancing transparency and user engagement
* To build a secure, tiered data access control system with an auditable request-and-approval workflow for raw data and analysis results.

## RESEARCH QUESTIONS

What tools are being utilized to effectively manage, share, acquire, and process the genomic metadata data and sequencing data, and for sharing the result in EPHI Bioinformatics unit?

## FRAMEWORK AND APPROACH

## 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.

## Purpose and benefit:

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 the main purpose of the development is and implementation of the GenDE platform for EPHI is. Some, but not limited to the following, are the purposes of the 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.
* Simplify 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.

## Project Scope:

* Development of a web portal for user registration, assessment, 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.

## Benefits of the project

* Simplified 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.

# 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.

## 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 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.

## 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.

## 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).

# Methodology

## 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, simplified workflow. We will implement the Security-First Data Engineering Framework (SF-DEF) architecture which is a system that provides a systematic method of incorporating security into the data processing processes [6] . 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[7].

## 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.

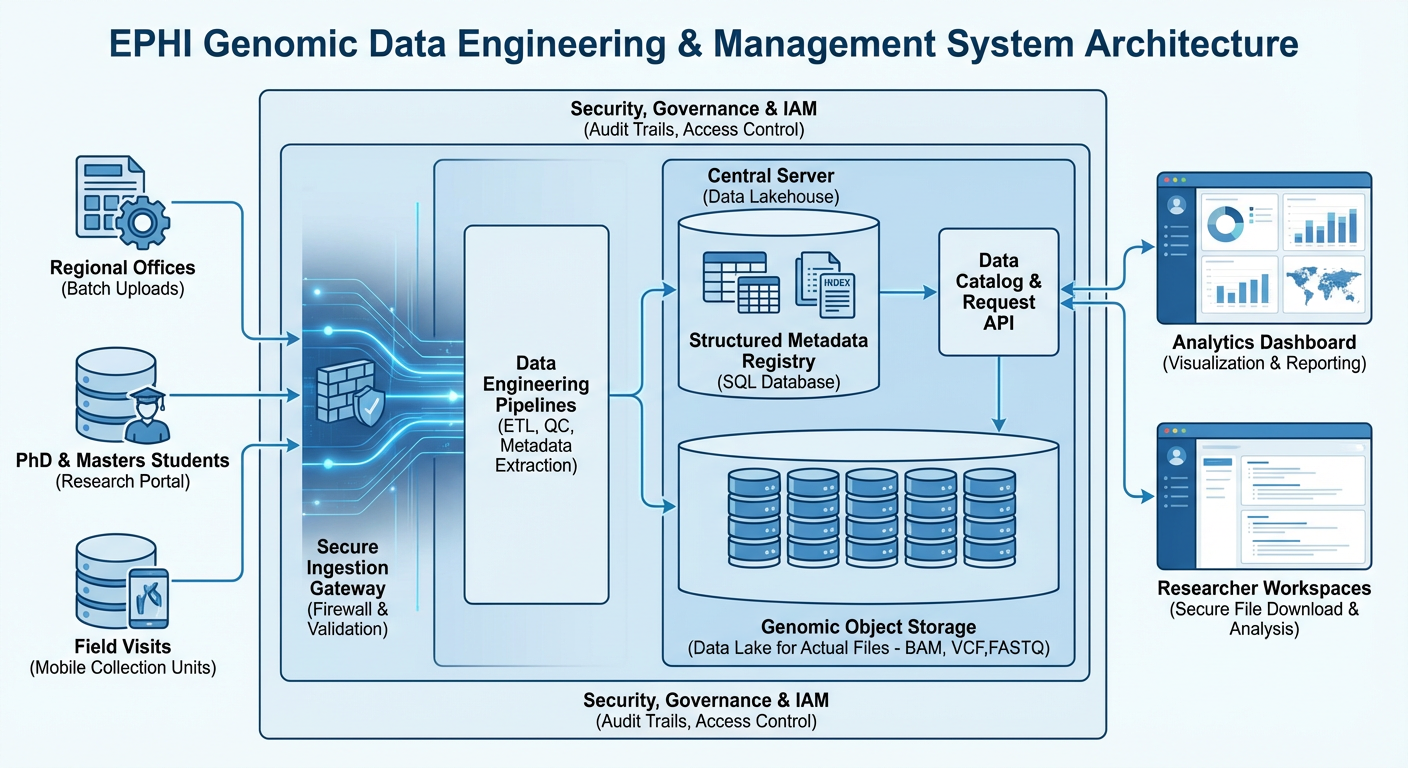


Figure 1. Architecture of the system

* + 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 shall support resumable uploads, rate limiting, and automated quarantine of suspicious submissions.

* + 1. **Validation and De‑identification Layer**

After the ingestion, 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.

* + 1. **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[8].

* + 1. **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[9]. Where appropriate, notarization or tamper-evident checksums are used to strengthen integrity guarantees. The data will be shared among inter-departments, students and researchers with adopted genomic data governance and sharing rule developed with alliance of Global Alliance for Genomics and Health [10][11].

## Process work flow

As shown in the figure 2 below, the process shall start from the metadata submission where the data is reside either from student, or researchers. The process follows metadata validation by lab technicians and by the system itself, and sequence generation. The system will be designed to serve students (MSc and PhD), researchers, laboratories, and external data providers who submit biological samples along with structured metadata through a secure EPHI Genomics web portal. Once a project is registered, the metadata is stored in a relational database while the physical sample is submitted to a sequencing facility. During the sequencing phase, the project status is tracked within the system to reflect its real-world progress. After sequencing is completed, the resulting genomic files (such as FASTA, BAM, or VCF) are uploaded into the platform by lab tech from EPHI, automatically indexed, and linked back to the original sample metadata without storing raw sequences in the database. Django, which is a Python web framework acts as the central orchestration layer, managing authentication, authorization, metadata persistence, file registration, and API access, while large genomic files are efficiently served from file storage using indexed, region-based access. Users can browse projects and samples through a hierarchical interface, visualize genomic regions directly in the browser using an embedded genome viewer, and explore analytical dashboards that show where data is stored, how it is distributed across projects, and how resources are being utilized. Once sequencing results are released, the system updates the project status to “result released” and automatically notifies the submitting user via email, providing a transparent, traceable, and scalable workflow that integrates laboratory operations, data engineering, and user-facing genomic visualization into a single cohesive platform.

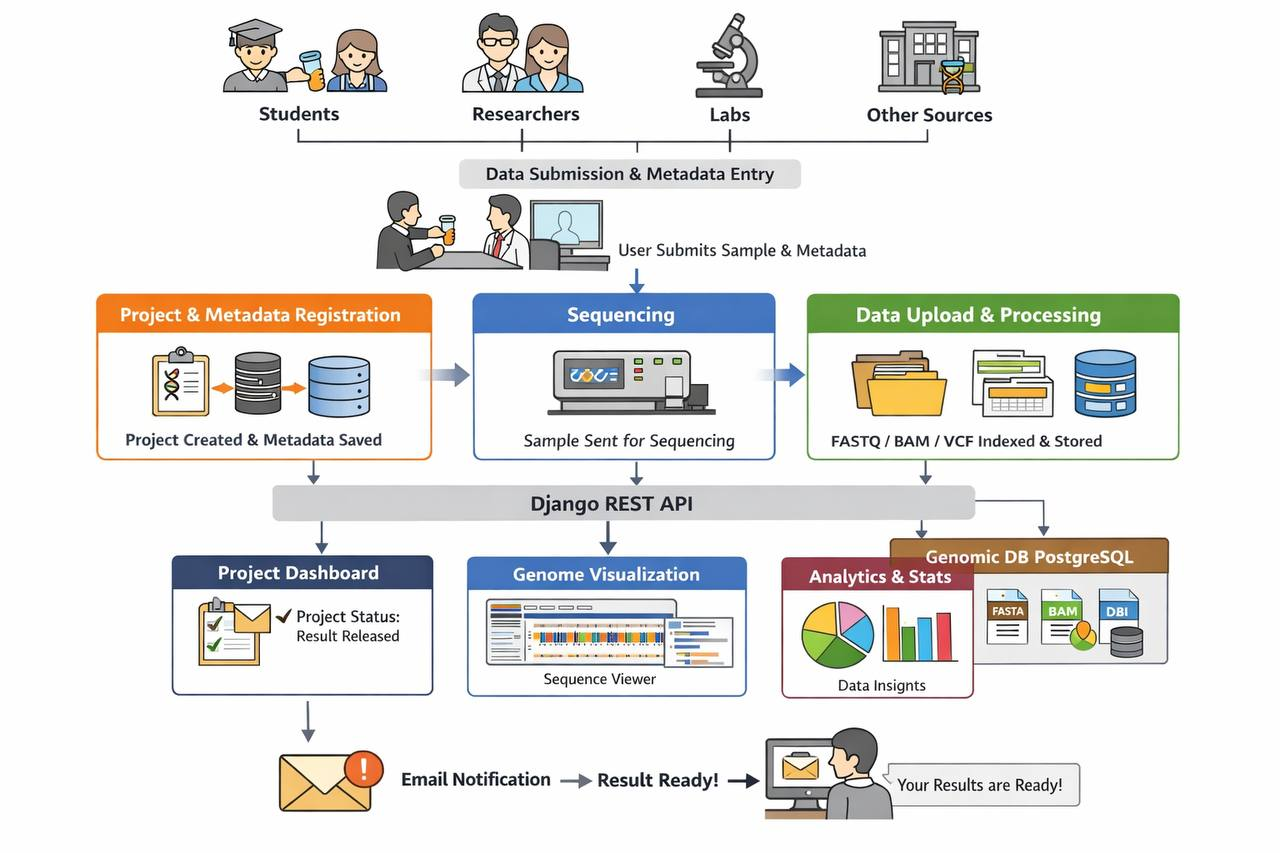


Figure 2. Process workflow.

## Deliverables

For the successful completion of this project, it is expected that 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.

## System Functionalities

The system functionality will be varied according to the changes in the requirements and needs which would be raised 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.

## 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.

## Third-party integration requirement

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

* + 1. EPHI IRB System: For ethical clearance validation.
    2. EPHI NDMC: As the primary endpoint for data archival, cataloging, and persistent identification (DOI minting).
    3. DataCite API (via NDMC): For registering datasets and obtaining DOIs.

## Resource requirement

The minimal resource requirement for this project to come to life, the following, but not limited to be 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.

## 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).

## 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.

# 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.

|  |  |  |
| --- | --- | --- |
| **Item** | **Type** | **Price (USD)** |
| Perdiem and transportation payment for coding at workstations and transportation from workstations to EPHI Every month, one expert | Air ticket: 80 USD per trip  Perdiem per trip per 12 months | 19,200 |
| Communication | 50 USD per developer workstation and fares for total of 12 months for 10 developer | 6,000 |
| Regular visit on a persona and system testing with real-time scenario at workshop. | 50 USD per person for total of 12 months 6 times | 6,000 |
| Discussion workshop on the system progress and real data testing every 2 month. | 50 USD per person for total of 12 months 6 times | 6,000 |
| Deployment and training and launching workshop on the system. | 50 USD per person for total of 12 months 6 times | 6,000 |
| Contingency (5%) | Contingency fund for extra payments, cost changes | 2,160 |
| Grand Total | | **45,360** |

# References

[1] G. Shabir, “Bioinformatics and Big Data Analytics in Genomic Research Qaiser asad Department of health science , university of Public Health , Gujrat , India,” *Med. Pap.*, vol. 3, no. December, pp. 165–179, 2023.

[2] S. Jafarbeiki, R. Gaire, A. Sakzad, S. K. Kermanshahi, and R. Steinfeld, “Collaborative analysis of genomic data: vision and challenges,” *Proc. - 2021 IEEE 7th Int. Conf. Collab. Internet Comput. CIC 2021*, pp. 77–86, 2021, doi: 10.1109/CIC52973.2021.00020.

[3] G. Govindarajan, “The Future of Data Engineering: AI and Machine Learning Integration,” *Eur. J. Comput. Sci. Inf. Technol.*, vol. 13, no. 11, pp. 79–86, 2025, doi: 10.37745/ejcsit.2013/vol13n117986.

[4] R. Gopalan, H. V. Parameswaran, and G. Kalpinagarajarao, “Journal of Artificial Intelligence , Machine Learning and Data Science Data Engineering Mastery : Automation Essentials for Researchers and Students,” pp. 1–9, 2023.

[5] R. Gopalan, H. V. Parameswaran, and G. Kalpinagarajarao, “Data Engineering Mastery: Automation Essentials for Researchers and Students,” *J. Artif. Intell. Mach. Learn. Data Sci.*, vol. 1, no. 3, pp. 1505–1513, 2023, doi: 10.51219/jaimld/ranjith-gopalan/339.

[6] Sushil Kumar Tiwari, “Security-First Data Engineering: Best Practices for Compliance in Healthcare and Financial Data Pipelines,” *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 61s, pp. 763–770, 2025, doi: 10.52783/jisem.v10i61s.13438.

[7] C. R. Nwokocha, O. O. Ogundolapo, and M. M. Eji, “Using Power BI and SQL for Predictive Health Data Analytics to Improve Decision-Making and Performance Tracking in Public Health Systems,” 2025.

[8] D. M. Saini, “Essentials of Data Engineering,” 2024.

[9] I.-S. Sivakamavalli, Jeyachandran; Park, Kiyun ; Kwak, “Genome Databases, Types and Applications: An overview,” *J. Diabetes Investig.*, vol. 3, no. 6, pp. 558–567, 2012, doi: 10.1111/jdi.12039.

[10] G. A. for G. Health, “Framework for Responsible Sharing of Genomic and Health-Related Data.” GA4GH, 2014.

[11] W. H. Organization, *Guidance for Human Genome Data Collection, Access, Use and Sharing*. Geneva: World Health Organization, 2024. [Online]. Available: https://www.who.int/publications/i/item/9789240102149