

PROGRESS REPORT (MSCA-SE)

PROJECT	
Project number:	101182801
Project acronym:	GenDAI
Project name:	Genomic applications for laboratory Diagnostics supported by Artificial Intelligence
Project starting date:	01/09/2024
Project duration:	36 months


PERIOD COVERED	
 Please note that this is only a progress report. The information in this report must also be included in the next periodic report/final report.	
Period	from 01/09/2024 to 31/08/2025

TABLE OF CONTENTS

PROGRESS REPORT (MSCA-SE)..... 1

0. EXECUTIVE SUMMARY 3

1. MILESTONES, DELIVERABLES AND CRITICAL RISKS 4

2. OVERVIEW OF THE PROGRESS AND ACTIVITIES 5

0. EXECUTIVE SUMMARY

Executive summary

The project's main achievement at this point is the delivery of the foundational elements that enable the further development of the components of the GenDAI platform: user requirements, a complete suite of formal conceptual models, the ethical framework and data management plan required by the future usage of the platform, and an initial real-world dataset prepared for development and validation of the project outputs.

The availability of these results has enabled the proper launch of all the technical development tasks. By the end of the reported period, all specific work packages were underway, alongside the activities related to project management and dissemination.

We have faced no unexpected problems in our work so far. As foreseen, there were typical delays related mainly to the formal launching of secondments of people who are usually involved either in academic or business routines. As part of our mitigation plans, we addressed these delays involving a larger number of secondees in performing shorter stages and increasing the partners visited in certain secondments. Another source of delays was related to the gathering of users / customers' requirements. Despite the strong commitment of our exemplary power user, practicalities related to the common understanding of the needs/expectations and the diversity of workflows/processes to be investigated/reviewed, slowed the preparation and execution of the interviews needed to discuss requirements and expectations, thus delayed tasks downstream to produce the first outputs. By the end of the period, catch-up has been achieved, and the next set of deliverables should be on time.

Regarding implementation of secondments, we have started 15 out of a project total of 31 secondments. Secondments already completed or ongoing by the time of this report represent a total effort of 105 PM (42% of project PM), involving 18 staff members. Women performed 34% of PM). So far, there is no need to make any major adjustments to the original secondments plan.

Based on these facts, the Consortium is fully confident in completing the project within the planned timeline and budget.

1. MILESTONES, DELIVERABLES AND CRITICAL RISKS

Deliverables and milestones (outputs/outcomes)	YES/NO
<p>We confirm that we updated the following Continuous Reporting screens:</p> <ul style="list-style-type: none"> • Deliverables • Mobility declaration • Gender • Dissemination and Communication Activities • Publications (Open Access compliant and repository) • Summary for publication <p><i>If there are issues, identify them and explain the reasons why</i></p>	YES
<p>All relevant Deliverables for this period have been submitted, with the exception of D9.2. This is currently still with the Ethical Advisor and will be submitted shortly.</p>	

Critical risks	YES/NO
<p>We confirm that we updated the following Continuous Reporting screen:</p> <ul style="list-style-type: none"> • Critical risks <p><i>Please provide an update of the foreseen risks (if applicable).</i> <i>In case of unforeseen risks, please provide a short summary.</i></p>	NO
<p>No need for update. Risks identified in DOA are valid and no other new risks were identified so far.</p>	

2. OVERVIEW OF THE PROGRESS AND ACTIVITIES

Overview of the progress and activities	YES/NO
<p>We confirm that the project activities are progressing as planned and that there are no major issues that would put the project implementation in danger. <i>If there are issues, identify them and explain the reasons why.</i></p>	YES
<p>Overall progress of each active WP (as per DOA):</p> <p>WP1 Clinical Requirements, Pre-analytics, sample Collection, Wet Lab Processing and Ethics</p> <p>Status: ONGOING (Month 1 - 36)</p> <p>Tasks Performed:</p> <p>The initial year of WP1 has successfully set the foundation for all subsequent technical development. Key tasks performed include:</p> <p>T1.1 Identification of Existing Body of Knowledge and Detailed User Requirements: A multi-faceted approach was employed, including extensive desk research, socio-economic modelling of the clinical diagnostics ecosystem, and direct fieldwork. A cornerstone of this task was a series of structured interviews with the clinical end-user, biovis Diagnostik MVZ GmbH, to gather specific requirements for AI, data processing, and visualization.</p> <p>T1.2 Conceptual Modelling and Formal Requirement Specification: The gathered user requirements were translated into a full suite of formal conceptual models using User-Centered System Design (UCSD) principles and specified in UML. This included the creation of Use Context, Use Case, Information, Component, Architecture, and UI Models.</p> <p>T1.3 Sample Collection: This task has successfully commenced with the internal delivery of the first batch of 100 patient samples. These samples are crucial for the initial development and validation of the platform's components.</p> <p>T1.4 Ethical Proposal, Approval and Clearance: All necessary ethical and legal questions regarding sample data processing and machine learning applications have been addressed, ensuring a responsible-by-design approach.</p> <p>T1.5 ENS Integration: Foundational work began on integrating the Entity Name System (ENS) to uniquely identify all project artifacts, linking them to research outputs and archived data to create a comprehensive web of knowledge.</p> <p>Main Outputs/Results:</p> <p>The work in WP1 has produced the foundational assets upon which the entire GenDAI platform will be built:</p> <ul style="list-style-type: none"> • A comprehensive analysis of user requirements and challenges across clinical workflows, data security, AI, and visualization. 	

- A complete suite of formal conceptual models (Use Case, Information, Architecture, UI) that serve as the technical blueprint for all subsequent development in WP2, WP3, WP4, and WP5.
- An initial dataset of 100 high-quality IBD patient samples, prepared for use in the development and validation of the platform's analytical pipelines.
- An established ethical framework and a formal Data Management Plan (DMP), ensuring all project activities are compliant with GDPR, FAIR principles, and best practices for responsible research.

These results provide the essential steps in achieving the project's primary goals:

- The sample dataset is the first tangible result towards RIO 1 (Create new metagenomic datasets).
- The conceptual models provide the agreed-upon, traceable framework that directly informs the architecture and implementation of RIO 2 (Diagnostics Workflow), RIO 3 (Cloud Platform), RIO 4 (AI Discovery), and RIO 5 (Interactive Reporting).
- The established ethical and data management plans are fundamental to achieving RIO 6 (Deliver a regulatory compliant tool suite).

Secondments:

Sabina Akram, Paolo Buono, Rosa Lanzilotti, Pasquale Ardimento from UNIBA visited ICT to work on the user requirements for the GenDAI platform as part of T1.1.

Kim Mertens (FTK) visited SAP to work on the detailed user requirements in T1.1, with a special focus on cloud computing.

Ana Maria Perna from OKK visited MTU and FTK to work on long-term-reproducibility and the Entity Naming System (ENS), also part of T1.1.

Haithem Afli and Mike Leske (MTU) stayed at ICT to collaborate on applying genomic foundation models to IBS diagnostics, gathering requirements (T1.1) and working on the formal specifications (T1.2).

Ruben Riestra (ICT) stayed at UNIBA and SAP to support the conceptual modelling of interactive visualizations and cloud deployment (T1.2).

In his secondment at MTU, Thomas Krause (ICT) worked on the requirements for the used AI models including ethical aspects, thus contributing to T1.4.

Matthias Hemmje (FTK) worked together with OKK on the integration of the ENS into the knowledge management platform (KM-EP) as required for T1.5.

Deviations

Typical delays related to a) the formalization and actual launching of secondments of people who are usually involved either in academic or business routines. Mitigation actions: we involved a larger number of secondees in performing shorter stages and increasing the partners visited in certain secondments. B) Delays related to the gathering of users / customers' requirements. Despite the strong commitment of our exemplary power user, practicalities related to the common understanding of the needs/expectations and the diversity of workflows/processes to be investigated/reviewed, slowed the preparation and execution of the interviews needed to discuss requirements and expectations, thus

delayed tasks downstream to produce the first outputs. By the end of the period, catch-up has been achieved, and the next set of deliverables should be on time.

WP2: GenDAI Diagnostics – Advanced Genomics Data Processing Workflow

Status: ONGOING (Month 6 - 30)

Tasks performed:

T2.1 Reproducible genomics data processing pipeline: the design of the microbiome annotation pipeline, as the core of the GenDAI Diagnostics Workflow, has been initiated using an event-based architecture to maximize transparency and reproducibility.

T2.2 GenDAI Discovery Model Integration: preparatory work has begun on the integration of AI models to be included in the Diagnostic Workflow.

T2.3 GenDAI Interactive Reporting Integration: early design discussion to align on requirements for embedding interactive reports into the Diagnostics Workflow has taken place.

T2.4 GenDAI Safe Integration Task lead: Work has been initiated to link the Diagnostic Workflow with the GenDAI Safe long-term storage archival services.

Main Outputs/Results:

Preliminary design of the workflow and annotation pipeline management system.

Scientific insight on how to build novel systems and application based on AI, also contributing to the book De Luzi, F., Monti, F., & Mecella, M. (2025). Engineering Information Systems with Large Language Models.

<https://link.springer.com/book/10.1007/978-3-031-92285-5>

Secondments:

Activities included a joint work session between ICT and the secondees of SAP (Massimo Mecella, Matteo Marinacci, Jacopo Rossi, Flavia Monti).

- Massimo Mecella contributed, in the context of T2.1, to the design of the microbiome annotation pipeline, focusing on system architecture alignment and integration within the cloud-based Workflow Management System.
- Matteo Marinacci supported, in the context of T2.1, the draft of reproducible genomics workflows, working on event-based orchestration and pipeline transparency features, by also adopting GenAI-based methodologies.
- Jacopo Rossi, in the context of T2.4, worked on establishing technical connections between the Diagnostics pipeline and the GenDAI Safe archival system, contributing to the traceability of data, intermediate outputs, and model versions.
- Flavia Monti, in the context of T2.4, focused on defining integration protocols for archiving inputs and outputs within the Diagnostics Workflow, supporting the alignment of storage services with ENS-based knowledge management.

WP3 GenDAI Safe & Cloud Computing Platform – Data Management & Long-Term Archiving

Status: ONGOING (Month 6 - 30)

Tasks Performed:

T3.1 Cloud & Infrastructure Support: Established cloud resources on Google Cloud (Compute Engine, Kubernetes, Vertex AI) to provide scalable compute, storage, and workflow management.

T3.2 GenDAI Safe: Designed and prototyped a secure, ontology-driven data management and archiving system ensuring reproducibility and cyber-resilience.

T3.3 ENS Integration: Progress made on assigning Persistent Identifiers (PIDs) via ENS to guarantee persistence and traceability.

T3.4 Reproducibility: Developed initial metadata extraction and reproducibility workflows.

T3.5 Long-Term Archiving System (LTAS): Design of an OAIS (ISO 14721) compliant archiving system, designed to capture raw data, workflows, ML models, and training datasets, ensuring full transparency and reproducibility.

Main Outputs/Results:

The work on WP3 has already produced several tangible results contributing to the development of a robust, secure, and reproducible data management infrastructure.

Operational cloud infrastructure established: a first version of the GenDAI cloud environment has been devised, integrating container orchestration (Kubernetes), scalable storage solutions, and GPU/TPU-based compute clusters. This provides partners with a functioning environment to run diagnostics pipelines and AI training at scale.

Secure and compliant data framework: the infrastructure incorporates GDPR- and OAIS-aligned policies for data handling, ensuring that clinical data can be managed in a way that is reproducible, auditable, and future-proof. This directly supports the long-term goal of delivering a regulatory-compliant diagnostics platform

Deliverable D3.1 – Cloud Infrastructure Resources documents these results and ensures their reproducibility across the consortium. Work is progressing towards the next major output: the GenDAI Safe component with integrated LTAS and access control services (D3.2, due M20), which will consolidate these achievements.

Draft implementation of D3.2 GenDAI Safe component, Long-Term Archiving System and Access Control Services (due M20) is on track.

Secondments:

Activities included joint design sessions between SAP (lead), ICT and secondees from Sapienza University of Rome (Massimo Mecella, Francesco Leotta, Andrea Marrella, Francesca De Luzi).

- Massimo Mecella focused on system architecture alignment with OAIS standards, via joint design sprints with ICT to translate OAIS roles (ISO 14721) into the project context, defining SIP/AIP/DIP profiles for the four classes of resources we will preserve.

- Francesco Leotta led metadata modelling and reproducibility strategies, ensuring that all diagnostic resources—including raw data, workflows, ML models, and training data—can be fully traced and reproduced.
- Andrea Marrella contributed to workflow management and integration with cloud orchestration tools, enabling scalable execution and archiving of end-to-end processes.
- Francesca De Luzi supported requirements engineering and usability considerations of LTAS, collecting ICT feedback to inform the next, improved iteration of the LTAS.

WP4 GenDAI Discovery - Biomarker prospection in the microbiome using AI

Status: Ongoing (Month 6 – 30)

Performed Tasks)

T4.1 Microbiome Feature Selector using Genetic Algorithm and ML: WP4 broadened its scope by advancing from BiGAMi-based feature selection towards Genomic Large Language Models (LLMs) for biomarker discovery and metagenomic analysis. This evolution aligns with the most recent advances in AI-driven genomics and positions the consortium at the forefront of the field. During his stay at MTU, ICT researcher Thomas Krause actively developed bioinformatics and genomic language model expertise and supported the integration of LLM-based approaches in T4.1.

T4.2 NLP for Metagenomics Analysis:

- Literature mining: NLP pipelines were applied to large microbiome corpora to extract links between microbial taxa and clinical phenotypes.
- Genomic NLP: Transformer models (DNABERT-2, DNABERT-S) were applied to 16S rRNA datasets, delivering alignment-free classification with performance advantages over BLAST+ and VSEARCH. These results were shared via a bioRxiv preprint (Alignment-Free Bacterial Taxonomy Classification with Genomic Language Models, June 2025).
- Conference success: “GASE: Generatively Augmented Sentence Encoding” by Manuel Frank and Haithem Afli (MTU) was accepted at EMNLP 2025 (Class A*). The GASE framework will be tested to optimize metagenomic embeddings within GenDAI.

T4.3 Model Configuration and Operationalisation: Foundations for reproducibility, monitoring and version control have been outlined to streamline deployment.

T4.4 Training UI Design planning progressed for a user interface enabling clinical partners to retrain or fine-tune models on updated datasets.

Main Outputs / Results

- Evolution from BiGAMi-centred feature selection to Genomic LLMs, fully aligned with state-of-the-art scientific directions.
- Alignment-free bacterial taxonomy framework disseminated via bioRxiv.
- GASE (Frank & Afli) accepted at EMNLP 2025 (Class A*), with direct application to metagenomic embeddings.
- Planning established for deployment (T4.3) and Training UI (T4.4).
- Broad dissemination via invited talks (UKCI 2024), publications and the forthcoming IEEE BIBM 2025 workshop.

Secondments Involved

- Thomas Krause Developed bioinformatics and genomic language model expertise at MTU and contributed to T4.1 by supporting the integration of LLM-based approaches.

Deviations with Respect to the DOA and Mitigation

- Enhancement to the research plan: WP4 expanded from a primary focus on BiGAMi to Genomic LLMs in line with rapid progress in the domain. This enhancement unlocks greater opportunities for biomarker discovery and strengthens clinical relevance.
- Resulting benefits: The approach has already delivered high-impact outputs (bioRxiv preprint; EMNLP 2025 acceptance) and increased international visibility, while maintaining full alignment with WP4 objectives.

WP5 GenDAI Interactive Reporting - Visual Data Analysis

Status: ONGOING (Month 6-30)

Performed tasks

T5.1 Identification of data of interest and visual structure. The identified data have been mapped with visual structures for the goal of visualizing such data and providing insights to the observer.

T5.2 Addressing user needs. This task is oriented to the identification of tasks performed by the users with the system.

T5.3 Interactive reporting. This task is focused on the development of interactive reporting. Starting from early prototypes, the system will be evaluated with end users who will contribute to the development of a user interface that fits the user's needs.

T5.4 Visual Analytics tools offering end-user composition paradigms. This task aims to allow the end users to customize the generated report according to their needs.

T5.5 Usability and UX. The assessment of the usability quality of the development system will be performed through a User-Centered Design (UCD) approach; according to UCD, the system is evaluated by the end users since the early phases to increase the possibilities of having a system that is efficient, effective, and satisfactory for the user.

Output/results

- Initial prototype with basic visualizations and basic interaction.
- Progress in preparing the first set of publishable outputs on GenDAI Interactive Reporting Visualizations (D5.1). This comprises information about the state of the art in visualizing genomic data and about the early prototypes of the interactive report visualizations. It takes into account the literature, the already existing reports, and the interactions with the end users.

<p>Secondments involved</p> <p>Sabina Akram mainly contributed to the activities of gathering information about the state of the art related to genomic analyses and existing visualization techniques that show the data gathered in the considered domain.</p> <p>Philippe Tamla (ICT) has collaborated in identifying visual structures for data visualization.</p> <p>Rosa Lanzilotti provided the consortium with her expertise in human-computer interaction, more specifically in running usability activities (interviews, user observation, heuristic inspections).</p> <p>Paolo Buono's main focus has been related to the state of the art of visualizations in the genomic context and, more specifically, in the IBD context.</p> <p>Andreas Hundsdörfer (ICT) started the exploration of the possibilities for the integration of the interactive reporting</p> <p>Pasquale Ardimento performed a study on the existing possibilities to develop interactive PDF files. This is useful for the project since interactivity in reports is a requirement.</p> <p>WP7 Project Management</p> <p>Tasks performed</p> <ul style="list-style-type: none"> • T7.1 Project Data, Content, and Knowledge Management Infrastructure: Setup of a comprehensive Data Management, Document Management, and Knowledge Management Ecosystem with Workflow, Audit Trail, and Long-Term Archive for the project to ensure full compliance along the project timeline with all ethics regulations. • T7.2 General operational and financial management. Weekly virtual meetings involving all WP leaders plus staff involved in key tasks secure continuous monitoring of project progress, identification of issues to be solved and decision making on mitigation measures, etc. • T7.3 Technology and scientific management. Weekly virtual meetings involving all WP leaders plus staff involved in key tasks secure continuous monitoring of project progress, identification of issues to be solved and decision making on mitigation measures, etc. • T7.4 Open Science and Data Management: Deliver a FAIR model for sharing of research data and regularly update the DMP. • T7.5 Quality and Risk Management: Creation and implementation of the Quality & Risk Management Plan <p>Main Outputs / Results</p> <p>Data Management Plan, Knowledge Management Resources and Quality & Risk Management Plan, published in D7.1</p> <p>WP8 Dissemination, Communication and Exploitation</p> <p>Progress and outputs according to plan. See table on Communication, Dissemination, Open Science and Exploitation below.</p>	
--	--

<p>Works in WP8 will be enlarged in the next reporting period as planned following availability of project results</p> <p>Work Package 9 Ethics requirements</p> <p>Tasks performed: ensure compliance with the 'ethics requirements' set out in this work package.</p> <p>Main outputs/results</p> <ul style="list-style-type: none"> • D9.1 H - Requirement No. 1 • D9.3 POPD - Requirement No. 4 • D9.4 POPD - Requirement No. 5 • D9.5 OEI - Requirement No. 6 <p>Deviations: We already submitted a draft of Requirement No. 3 (D9.2 H) to our External Ethical Advisor, and we will deliver it along the next few weeks as we get his approval.</p>	
--	--

Implementation timetable	YES/NO
<p>We confirm that the project activities are on schedule and that there are no significant delays. <i>If there are delays, identify them and explain the reasons why.</i> <i>In particular, please provide (1/2 page max):</i></p> <ul style="list-style-type: none"> - updates to the secondment plan and to the corresponding activities 	YES

Communication, Dissemination, Open Science and Exploitation
<ul style="list-style-type: none"> • Progress as planned. Neither changes nor deviations to be reported. Main activities: • Creation and submission of D8.1 Dissemination, Communication and Exploitation Plan • Launch of the project website: https://project-gendai.eu/ • Design of scripts for promotional short videos, written statements and press releases • Creation and execution of 1 Seminar and 1 workshop aimed at PhD candidates on academia-industry cooperation using innovation driven research projects. 10 participants from UNIBA (Bari, feb2025) and 15 participants from La Sapienza (Rome, June 2025), external to the GenDAI project team • Potential allies identified among MSCA SE “sibling” projects: EVEREST, EngVIPO, BIOREM, i-GREENPHARM, IDPfun2, EXPAND-EV, AHEAD and GRASSHOPPER.

Ongoing: Getting in contact to explore cooperation/shared impact generation opportunities

- Contributions to the book De Luzi, F., Monti, F., & Mecella, M. (2025). Engineering Information Systems with Large Language Models. <https://link.springer.com/book/10.1007/978-3-031-92285-5>
- GASE paper (Frank & Afli) accepted at EMNLP 2025 (A conference), a flagship NLP output with direct application to metagenomics.
- Dissemination through invited talks (UKCI 2024), publications, and active participation in the upcoming IEEE BIBM 2025 Workshop on Genomic Foundation Models for Diagnostic Innovation, 15–18 December 2025, Wuhan, China. Six (6) papers in preparation, with the preliminary titles: a) A General Overview of GenDAI project and its vision, b) Responsible use of AI in Genomics and ethical implications c) Advances in Genomic Foundation Models, d) Revisiting data visualizations in diagnostic reports, e) Technical infrastructure for GFM in the GenDAI Project, f) LLM-driven cloud-based design.
- Further outreach and impact generation: GenDAI team members Binh Vu (FTK) and Thomas Krause (ICT) have been invited to become permanent members of CCSDS MOIMS-DAI which is responsible for maintaining the ISO Archival and Long-Term preservation standard ISO 14721 Open Archive Information Systems (OAIS).