Deliverable D 3.2 GenDAI Safe component, LTAS, and Access Control Services

Work Package	**WP 3 - GenDAI Safe & Cloud Computing Platform - Data Management & Long Term Archiving**
Task(s)	**T3.1 Cloud and Infrastructure Support** **T3.2 GenDAI Safe** **T3.3 ENS Integration** **T3.4 Reproducibility** **T3.5 Long-Term Archiving System (LTAS)** **T3.6 Attribute-Stream Based Access Control**
Deliverable	**D3.2 GenDAI Safe component, LTAS, and Access Control Services:** A secure and efficient data management solution to be integrated with the GenDAI Diagnostics processing pipeline. In addition, the delivery of the LTAS, integrated with sophisticated access control services, ensures secure storage, retrieval, and management of project data over an extended period with full reproducibility.
Author(s)	Paolo Buono, Francesca De Luzi, Thomas Krause, Philippe Tamla, Flavia Monti, Ruben Riestra, Andrea Molinari, Massimo Mecella, Matthias L. Hemmje
Version	1.0
Dissemination Level	PU

Table of Contents

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

Executive Summary

Overall assessment

The GenDAI project aims to develop a cloud-enabled, AI-driven diagnostic platform that leverages metagenomic data to deliver precision medicine solutions. Central to this vision is the design and implementation of a secure, interoperable, and standards-compliant cloud infrastructure that supports all stages of the diagnostic pipeline, from data ingestion and processing to storage, reporting, and compliance.

This deliverable presents the first implementation of the core of the GenDAI Safe components. The LTAS is designed in conformance with the OAIS standard and deployed on Google Cloud Platform, providing structured, tiered archival of all project data types, including raw genomic files, AI models, diagnostic outputs, and electronic health records, each accompanied by a rich JSON provenance record. Access to the archive is governed by Google Cloud IAM role assignments.

These results provide the GenDAI project with a secure, reproducible, and standards-compliant data management infrastructure that enables every data to be fully traced and versioned. The persistent identification mechanism assigns a UUID-based PID to every archived entity and integrates with the ENS to ensure global resolvability, while the reproducibility component links all data required to reconstruct any diagnostic result into a navigable provenance graph.

The LTAS design constitutes a reusable archival framework applicable to any clinical or research platform requiring long-term preservation of heterogeneous biomedical data with full provenance traceability. The modular, cloud-native architecture provisioned through Terraform and deployable on any GCP-compatible environment positions these results as a regulatory-ready data management service with immediate scientific and commercial exploitation potential.

Outputs and Project Progress

The results of this deliverable contribute directly to the project´s RIOs by providing a robust, cloud-based foundation for the development of the GenDAI platform that integrates advanced data management and knowledge infrastructure, focusing on security, reproducibility, and long-term archiving.

The components described in this deliverable are designed to be directly integrated into the GenDAI Diagnostics Workflow developed in WP2. Every diagnostic run executed by the pipeline will automatically archive its inputs, intermediate results, ML model versions, and final outputs into the LTAS, with each artifact linked through its PID to form a complete and reproducible provenance chain. These results also provide the data foundation required by WP4 and WP6, where archived outputs and provenance records will be used to validate diagnostic results and gather feedback for the improved iteration delivered in D3.3.

Contributions to Impacts

Results presented along this Deliverable will contribute to GenDAI's expected impacts: a) the data infrastructure supports clinical pathologists and laboratory operators by tracking every diagnostic result for reproducibility, so that it allows them data audit, re-execution and verification against the specific versions of raw data, model weights, and workflow specifications that produced it.; b) Regulatory bodies and ethical supervisory bodies and advisors will benefit from a system that enforces GDPR-compliant access control and ensures

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

Funded by the European Union

data residency within the EU, directly supporting the compliance requirements reported in WP9; c) Research partners across the consortium will benefit from a shared, versioned archive that preserves all intermediate artefacts of the diagnostic pipeline, enabling collaborative validation and iterative improvement of produced artifacts.

These results are value-adding both from a scientific as well as from commercial standpoints. Scientifically, the OAIS-compliant LTAS and provenance framework provide a replicable blueprint for FAIR and reproducible data management in genomics, directly applicable to other academic consortia working on biomedical data preservation. Commercially, the modular, cloud-native architecture, deployable on any GCP-compatible environment through Terraform, is ready to be offered as a regulatory-compliant data management service for clinical laboratories and digital health companies seeking to meet the archiving and traceability requirements of IVDR and future EU health data regulations. The planned evolution towards semantically descriptive PIDs grounded in the GenDAI project ontology, and the deeper integration with the ENS, further positions the system as a foundation for a Linked Data infrastructure for biomedical knowledge, with exploitation potential well beyond the scope of the project itself.

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

Acronyms

Acronym	Full Term
LTAS	Long-Term Archiving System
GenDAI	Genomic Applications for Laboratory Diagnostics Supported by Artificial Intelligence
AI	Artificial Intelligence
IBD	Inflammatory Bowel Disease
GCP	Google Cloud Platform
GCS	Google Cloud Storage
IaC	Infrastructure-as-Code
OAIS	Open Archival Information System
SIP	Submission Information Package
AIP	Archival Information Package
DIP	Dissemination Information Package
PDI	Preservation Description Information
PID	Persistent Identifier
GUID	Globally Unique Identifier
ENS	Entity Naming Service
EHR	Electronic Health Record
IAM	Identity and Access Management
ASBAC	Attribute-Stream Based Access Control

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

1. Introduction

1.1 Project Overview

The GenDAI project (Genomic applications for laboratory Diagnostics supported by Artificial Intelligence) aims to develop a novel medical diagnostics platform that leverages metagenomic data and Artificial Intelligence (AI) to improve the diagnosis and monitoring of inflammatory bowel disease (IBD). The platform will provide clinicians with a comprehensive, personalized view of a patient's microbiome, enabling more targeted and effective treatment strategies.

1.2 Purpose of this Document

This document presents the design, implementation, and first operational results of the GenDAI Safe component, the Long Term Archiving System (LTAS), and the Access Control Services, constituting deliverable D3.2 of WP3.
The document addresses the three objectives of WP3 as stated in the Description of Action: O3.1 (design, implementation, and validation of the GenDAI Safe and cloud infrastructure), O3.2 (definition of policies and workflows to access and use data and metadata), and O3.3 (prototype of a LTAS based on ISO 14721). It builds directly on the outputs of two preceding deliverables: D1.1, which provided the user requirements, data management specifications, and security requirements that guided the architectural decisions described here, and D3.1, which describes the architecture and the cloud infrastructure.

1.3 Document Organization

The document is structured as follows. Section 2 describes the cloud infrastructure and storage design. Section 3 presents the ITAS. Section 4 describes the metadata and the alignment with OAIS standard. Section 5 covers the persistent identification system. Section 6 describes the access control system. Section 7 presents the conclusions and outlines the next steps towards the improved iteration of the LTAS to be delivered in D3.3.

1.4 Methodology

The design and implementation work described in this document followed an iterative, requirements-driven methodology grounded in the outputs of WP1. The architectural decisions were informed by the user requirements, data management specifications, and security requirements captured in D1.1, which established the need for long-term reproducibility, persistent identification, GDPR-compliant access control, and OAIS-standard archival as core requirements of the GenDAI Safe component. The results presented in this document represent the first implementation iteration and will be refined in D3.3 on the basis of operational feedback gathered during the piloting activities in WP6.

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

2. Cloud Infrastructure and Storage Design

2.1 Usage of Google Cloud Platform

The GenDAI infrastructure is built on top of Google Cloud Platform (GCP), selected as the primary cloud provider for the project. GCP was chosen for its strong compliance posture with GDPR when operated from European regions, its native support for containerized workloads through Docker and Kubernetes, and the availability of managed services covering the full spectrum of compute, storage, and data management requirements.

All cloud resources are provisioned and operated exclusively from GCP regions located within the European Union, specifically in Germany (europe-west3, Frankfurt), ensuring that personal and clinical data never leaves EU jurisdiction. This geographic constraint is a deliberate design decision driven by the sensitivity of the metagenomic and health-related data processed by the GenDAI platform and is in full alignment with the requirements set out in GDPR and the project's ethics framework.

The platform leverages the following GCP services as its foundational layer:

- Google Cloud Storage (GCS): object storage for all archival data, including raw genomic files, AI model weights, pipeline artifacts, and diagnostic outputs.
- Google Cloud IAM and Google Groups: for identity management and access control enforcement.

2.2 Storage Buckets and Storage Classes

Data in the GenDAI system is stored in Google Cloud Storage buckets, which act as the primary containers for all archived objects. GCS offers four storage classes with distinct cost and access-latency characteristics, each suited for different data lifecycle phases:

1. Standard Storage: designed for data that is accessed frequently. Offers the lowest latency and highest throughput, with no minimum storage duration.
2. Nearline Storage: optimized for data accessed approximately once per month. Offers lower storage cost with a minimum storage duration of 30 days and a small per-access fee.
3. Coldline Storage: designed for data accessed roughly once per quarter, with a minimum storage duration of 90 days and a higher per-retrieval cost.
4. Archive Storage: the lowest-cost class, intended for data accessed less than once per year, with a minimum storage duration of 365 days and the highest retrieval cost.

Buckets are organized to reflect the type of data they contain, enabling both logical separation and fine-grained access control at the bucket level.

2.3 Geographic Location and EU Compliance

A deliberate design decision was made to restrict all GenDAI Safe cloud resources to the europe-west3 (Frankfurt, Germany) GCP region. This single-region placement was chosen over multi-region or dual-region configurations in the first implementation to avoid the

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

additional complexity and cost associated with cross-region replication while still meeting the project's data residency obligations.

Possible extension of the GenDAI architecture may consider extending the storage architecture to a dual-region or multi-region configuration to improve disaster recovery posture and availability guarantees. In particular, there is an open design question regarding whether medical data should be stored in the specific country where the patient data was collected – for example, storing Italian patient data in Italian GCP regions and German patient data in German regions. This concern is noted and will be addressed in the next iteration of the LTAS.

2.4 Infrastructure Provisioning with Terraform

All cloud infrastructure resources for the system are defined and managed using Terraform, an Infrastructure-as-Code (IaC) tool. Rather than creating and managing buckets manually through the GCP console, Terraform configuration files describe the desired state of all GCS buckets, IAM bindings, lifecycle rules, and network settings in a declarative and version-controlled manner.

This approach provides several important benefits for the project:

- Reproducibility: the entire infrastructure can be recreated from scratch in a new environment by applying the Terraform configuration, which is essential for disaster recovery and for potential migration to alternative cloud providers in the future.
- Auditability: all infrastructure changes are tracked through version control (Git), providing a complete audit trail of who changed what and when.
- Consistency: identical bucket configurations, IAM policies, and lifecycle rules can be applied across development, testing, and production environments without manual intervention.
- Scalability: as new data types or partner organizations are onboarded, new buckets and policies can be added simply by extending the Terraform definitions.

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801. Funded by the European Union

3. Long-Term Archiving System (LTAS)

3.1 Adoption of the OAIS Standard

The LTAS of GenDAI is designed in conformance with the Open Archival Information System (OAIS) reference model, formalized as ISO 14721. OAIS defines a framework for the preservation of digital information over extended time horizons, addressing the challenges of technological obsolescence, data integrity, and long-term accessibility.

The OAIS standard mandates six key responsibilities for a compliant archive:

- Negotiate and accept information to be archived from producers.
- Obtain sufficient control over the information to ensure long-term preservation.
- Define the designated community of users for the archive.
- Ensure that the preserved information is independently understandable by the designated community without recourse to the original producers.
- Follow documented policies and procedures to ensure the information is preserved with reasonable confidence and can be disseminated as authentic copies.
- Make the preserved information available to the designated community.

The OAIS model structures an archive around six functional entities: Ingest, Archival Storage, Data Management, Administration, Preservation Planning, and Access. In the GenDAI Safe implementation, these entities are mapped to concrete GCP services and custom software components, described in the following sections.

Three types of information packages are defined by OAIS and are adopted in the LTAS:

- Submission Information Package (SIP): the package as received from the producer (a GenDAI partner or the processing pipeline). It contains the content data together with metadata fields for reference (PID), provenance, context, and fixity.
- Archival Information Package (AIP): the package as stored in the archive. The SIP is transformed into an AIP by the Ingest function, which adds packaging information and preservation metadata. The AIP is what is stored in GCS.
- Dissemination Information Package (DIP): the package as delivered to a consumer in response to a request. It is derived from one or more AIPs and may contain a subset of the archived information.

3.2 Structure of the Archive

The GenDAI archive is organized into four logical areas, each corresponding to a category of data produced or consumed during the GenDAI diagnostic workflow:

- Data: raw and processed biological data files, including FASTQ sequencing reads, FASTA reference sequences, CSV metadata tables and electronic health data (EHR). These represent the primary scientific inputs and outputs of the platform.
- Models: trained AI model weights stored as .pt files, together with associated configuration files and training provenance records. Archiving model weights alongside the data used to produce them is essential for long-term reproducibility of diagnostic inferences.

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

- Workflows: Dockerfile specifications and any workflow definition files that describe the exact computational environment and processing steps used for each diagnostic run. Crucially, this includes the explicit versioning of all software components involved in a run. Containerization ensures that software dependencies can be reconstructed years after a run was performed. A more comprehensive treatment of workflow provenance capture, covering the full range of configuration artefacts and their relationship to reproducibility, is planned as a focus of the next iteration of the LTAS, to be delivered in D3.3.
- Outputs: diagnostic results, finding reports (HTML, PDF), visualizations (PNG/SVG), and any logs or LLM-generated text associated with a completed diagnostic session.

3.3 Mapping of File Types to Storage Levels

The following table summarises the mapping between the file types present in the GenDAI platform and their corresponding GCS storage class assignments, reflecting the expected access frequency and long-term preservation requirements of each type. It should be noted that these assignments represent the current design decisions based on the access patterns and retention requirements identified during the first implementation phase. As the platform evolves and operational experience accumulates through the piloting activities, the storage class assignments may be revised in the improved version of the LTAS delivered in D3.3. In particular, lifecycle transition rules and storage tier allocations will be updated to reflect real-world access patterns, cost optimisation findings, and any changes in the regulatory retention requirements applicable to specific data categories.

File Type	Description	Initial Storage Class	Long-Term Class
FASTQ	Raw sequencing reads	Coldline	Archive
FASTA	Reference sequences	Coldline	Archive
CSV	Metadata	Coldline	Archive
PDF	Finding reports	Nearline	Archive
HTML	Finding reports	Nearline	Archive
PNG / SVG	Visualizations	Nearline	Archive
TXT	Logs, LLM I/O	Nearline	Archive
.pt	AI model weights	Standard	Archive
Dockerfile	Container specifications	Nearline	Archive

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

XML (EHR)	Electronic health records	Nearline	Archive

The file types currently mapped to archival storage levels represent the initial configuration of the system, defined on the basis of the requirements and access patterns identified during the first implementation phase. This mapping is expected to be reviewed and updated in the final version of this deliverable, following the piloting activities carried out within WP6.

In particular, new file types may be added, removed, or reclassified depending on:

- operational needs emerging during the piloting activities;
- evolution of the formats used by the GenDAI diagnostic pipeline;
- updates to the applicable regulatory requirements (GDPR, IVDR);
- cost and performance optimisations derived from the analysis of real-world access patterns.

Any revision to the mapping will be documented and tracked in the updated version of the system, ensuring full consistency with the OAIS model and continuity of the provenance chain.

3.4 Versioning and Long-Term Storage

Object versioning is enabled on all primary GCS archival buckets. When a new version of an object is uploaded, the previous version is retained as a non-current version and is not deleted automatically. Non-current versions are subject to their own lifecycle policies and may be transitioned to lower-cost storage classes after a configurable period.

Versioning is managed at two complementary levels. At the infrastructure level, GCS natively retains all previous versions of an object within the bucket. At the application level, the GenDAI Safe system maintains an explicit version history within the JSON provenance record associated with each object. Each archived object carries a version number that is incremented on every update, and the provenance record of each version contains a reference to the PID of its immediate predecessor, forming a linked version chain that is fully navigable without relying on the GCS internal versioning mechanism alone.

This dual approach serves two distinct purposes. The GCS-level versioning provides a safety net against accidental overwrites and a low-cost mechanism for retaining historical states of the archive. The application-level version chain, embedded in the provenance records, provides a semantically meaningful and standards-compliant representation of the evolution of an archived object over time, which is accessible to any consumer of the archive independently of the underlying storage infrastructure.

4. Metadata and Provenance Management

4.1 JSON-Based Provenance Records

Each object archived in the GenDAI LTAS is associated with a structured JSON provenance record. This record is stored alongside the object in GCS as a companion metadata file and

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

serves as the primary vehicle for capturing the information required by the OAIS standard for every archived artefact. The record is organised into seven top-level categories: Reference Information, Representation Information, Fixity Information, Provenance Information, Context Information, Access Rights Information, and Descriptive Metadata.

The categories are derived from the PDI model defined by the OAIS standard (ISO 14721) and serialised in JSON for its simplicity. While the record does not formally adopt a dedicated metadata standard, its structure presents non-formal conceptual alignments with several established standards: the fixity, provenance and access rights categories align with PREMIS (Preservation Metadata Implementation Strategies), the natural complement to OAIS at the level of concrete preservation metadata; the descriptive metadata fields align with Dublin Core; and the overall provenance model presents a conceptual affinity with W3C PROV. These alignments will facilitate a future formal mapping towards one or more of these standards.

The full structure of the JSON provenance record is as follows:

```
{
  "reference_information": {
       "reference_pid": "<UUID>",
       "version": <integer, update on each update>,
    "previous_version_pid": "<UUID of the previous version, null if first version>"
  },
  "representation_information": {
       "mime_type": "<standard MIME type, e.g. application/pdf>",
    "original_extension": "<original file extension, e.g. .pdf>"
  },
  "fixity_information": {
    "checksum_algorithm": "<algorithm used, e.g. SHA-256>",
       "checksum_value": "<hash value of the file content>"
  },
  "provenance_information": {
       "filename_original": "<file name at the time of submission>",
       "file_size_bytes": <size in bytes>,
    "system_creation_date": "<file creation date in the operating system>",
    "system_modification_date": "<last modification date in the operating system>",
       "ingest_timestamp": "<exact date and time of archiving, ISO 8601>",
    "submitting_application": "<name of the application that performed the upload>",
       "submitting_user": "<ID or name of the operator who uploaded the file>"
  },
  "context_information": {
    "originating_department": "<department, clinic, laboratory, or institute>",
       "document_type": "<type of document, e.g. report, image, raw data>",
```

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

```
      "relation_id": "<ID of the medical record or event associated with the file>"
  },
  "access_rights_information": {
    "access_classification": "<confidentiality level, e.g. restricted, internal, public>",
      "legal_framework": "<applicable regulatory framework, e.g. GDPR, IVDR>",
    "retention_period_years": <mandatory retention period in years>
  },
  "descriptive_metadata": {
      "filename": "<human-readable file name>",
    "referring_physician": "<person responsible for uploading the file>",
      "notes": "<additional free-text notes>"
  }
}
```

The six categories of the record map directly onto OAIS concepts as follows.

Reference Information provides the Persistent Identifier (PID) of the archived object, implemented as a UUID in the current version. The PID is the stable, permanent key that links the archived object to all downstream references, including those maintained by the ENS service provided by OKK. Once assigned, a PID is never reused or reassigned. The Reference Information block contains three fields: reference_pid, which holds the UUID assigned to this specific version of the object; version, an integer that starts at 1 on first ingest and is incremented on every subsequent update; and previous_version_pid, which holds the UUID of the immediately preceding version, or null if the record represents the first version. Together, these three fields form a traversable version chain: given any version of an archived object, its complete history can be reconstructed by following the previous_version_pid references backwards to the original ingest, without requiring any centralised version registry.

Representation Information captures the technical format of the content object: its MIME type and original file extension. This information allows the archive to understand how to interpret the binary content of the object independently of the software environment in which it was originally created, which is a core requirement of the OAIS standard for long-term intelligibility. This is particularly important for the diversity of file types present in the GenDAI platform, which spans genomic formats (FASTQ, FASTA), health records (XML/EHR), AI artefacts (.pt, Dockerfile), and clinical outputs (PDF, PNG/SVG, CSV).

Fixity Information contains the cryptographic checksum of the file content, together with the algorithm used to compute it. This information is used to verify the integrity of the stored object at any point in the future, detecting silent data corruption or unauthorised modification. GCS provides native checksum verification at upload time; the GenDAI Safe system supplements this with periodic integrity checks that recompute and compare checksums against the stored values, particularly for objects in Coldline and Archive storage classes where silent bitrot over long time periods is the primary risk.

Provenance Information records the full chain of custody for the object from its origin to the moment of archival. It captures the original filename at submission, the file size, the creation and last-modification timestamps as recorded by the operating system, the precise timestamp

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

of ingest into the archive, the name of the application that performed the upload (e.g., the GenDAI Diagnostics pipeline or a partner's transfer tool), and the identity of the operator or service account responsible for the submission. Together, these fields establish who submitted what, from where, and when, which is essential both for scientific integrity and for regulatory compliance under GDPR and IVDR.

Context Information situates the archived object within the broader clinical and organisational context of the GenDAI project. The originating department field records the clinic, laboratory, or research institute from which the object originates. The document type field classifies the object by its role (e.g., raw sequencing data, diagnostic report, visualisation). The relation ID field links the object to the specific medical record or diagnostic event with which it is associated, enabling the archive to reconstruct the full set of artefacts belonging to a given patient episode or experimental run.

Access Rights Information encodes the governance constraints that govern who may access the object and for how long it must be retained. The access classification field specifies the confidentiality level of the object (e.g., restricted, internal, public), which is used by the ASBAC access control layer described in Section 6 to enforce appropriate access policies. The legal framework field records the regulatory regime applicable to the object (e.g., GDPR, IVDR), providing a machine-readable basis for compliance auditing. The retention period field specifies the mandatory retention duration in years, enabling the lifecycle management system to flag objects that are approaching the end of their required retention period rather than deleting them prematurely.

Descriptive Metadata provides human-readable information intended to support the discovery and usability of the archive. The filename field contains the current display name of the object, the referring physician field identifies the clinician or researcher responsible for the submission, and the notes field allows free-text annotations to be associated with the object at ingest time.

4.2 Alignment with OAIS Information Model

The JSON provenance record is designed to satisfy the full Preservation Description Information (PDI) requirements of the OAIS standard. The content information of each package, the file itself, is complemented by the six PDI categories captured in the JSON record. The Reference Information provides the unique identifier. The Representation Information ensures long-term interpretability. The Fixity Information guarantees integrity. The Provenance Information establishes the chain of custody. The Context Information situates the object within its scientific and clinical setting. The Access Rights Information encodes the governance constraints.

The OAIS Data Management entity is realized through the provenance JSON records and the GCS metadata layer, which together allow the archive to respond to queries about stored objects and to return the relevant metadata alongside the content data when a Dissemination Information Package (DIP) is assembled for delivery to a consumer.

5. Persistent Identification

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

5.1 Generation of PIDs via GUID

In the current implementation of the GenDAI LTAS, each archived object is assigned a Persistent Identifier (PID) at the time of ingest. PIDs are generated as Globally Unique Identifiers (GUIDs), using the UUID version 4 standard (randomly generated, 128-bit identifiers). The GUID is included in the PID field of the JSON provenance record and is also stored as a custom GCS object metadata attribute, enabling retrieval of any archived object by its PID without requiring knowledge of the file's original name or storage location.

The PID serves as the stable, persistent reference to an archived object. Once assigned, a PID is never reused or reassigned, even if the underlying object is updated. New versions of an object receive new PIDs, with the provenance record of the new version referencing the PID of the predecessor, forming a version chain.

5.2 Integration with the ENS

The Entity Naming Service (ENS) is provided by OKK (OKKAM) as a project-wide service for assigning and resolving persistent identifiers across all GenDAI data assets and knowledge artefacts. The ENS is designed to ensure the global uniqueness and long-term resolvability of identifiers, going beyond what a locally generated GUID can guarantee.

OKK integrates the ENS with the GenDAI data management process, so that every entity required to be persistently archived, including input data objects, model versions, workflow runs, and output artefacts, is registered in the ENS and receives an ENS-managed PID. These PIDs are then stored in the JSON provenance records alongside the GUID, allowing both local lookups (via GUID) and globally resolvable references (via ENS PIDs).

5.3 Ontology Extension for Descriptive PIDs

A recognized limitation of GUID-based PIDs is that they are opaque: a GUID conveys no information about the nature or content of the identified object. Future development aims to extend the PID scheme to leverage the full expressive power of the ENS and formal ontologies developed within the project.

Specifically, the development of a formal, extensible ontology to represent the relevant concepts and relations of the microbiome research domain enables such. This ontology will allow PIDs to be grounded in semantic descriptions, so that an identifier encodes not just a unique reference but also information about the type of object, its role in the workflow, and its relationships to other archived entities.

This evolution towards semantically descriptive PIDs will enhance both the discoverability of archived objects and the expressiveness of the provenance graph, moving the LTAS towards a Linked Data architecture that can be queried and reasoned over using standard semantic web tools.

6. Access Control System

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

6.1 Access via Google Groups and IAM Roles

Access to the GenDAI archive is managed through Google Cloud Identity and Access Management (IAM), using Google Groups as the organizational unit for role assignment. This approach allows access policies to be defined at the group level (e.g., all members of a given partner organization) rather than for individual users, simplifying administration and reducing the risk of misconfiguration.

The following role categories are defined:

- Producers with write access: users authorized to submit new objects to the archive. Producers are granted the writer role on the relevant interested buckets.
- Consumers with read access: users authorized to retrieve archived objects or query the provenance metadata. Consumers are granted the viewer role on the relevant buckets or at the object level.
- Administrators with full access: users designated as project administrators manage the overall archive infrastructure, IAM policies, lifecycle rules, and audit logs. Administrators hold the Storage Admin role.

Access can be granted at multiple granularities within GCS: at the bucket level (granting access to all objects within a bucket) or at the individual object level (granting access to specific files). This flexibility allows fine-grained control, for example to expose only the anonymized outputs of a diagnostic ru$f$n to a clinical consumer while keeping the raw patient data restricted to authorized research partners.

6.2 ASBAC Extension

The attribute-based access control layer provides a dynamic and policy-driven access control mechanism that operates on top of the IAM role assignments. The Attribute-Stream Based Access Control (ASBAC) system is capable of evaluating access decisions based on a richer set of attributes than static IAM roles, including the semantic properties of the requested object (as described in the project ontology) and the current context of the requesting user.

In the current implementation, the ASBAC system extends the base GCS access control by securing the API endpoints of the LTAS service layer. Access requests to retrieve archived objects or provenance metadata are intercepted by the ASBAC policy engine, which evaluates a set of access policies expressed in terms of the attributes of the requesting subject, the requested resource, and the current environment.

LLM-assisted workflows are being explored to support the authoring of ASBAC policies, reducing the expertise required to define and maintain complex access control rules.

7. Conclusions and Next Steps

This deliverable has presented the first full implementation of the GenDAI Safe component, the LTAS, and the Access Control Services, addressing the three objectives of Work Package 3 as defined in the DOA.
Several directions for further development have been identified and will be addressed in the next iteration of the system, to be delivered in D3.3, following the piloting activities in WP6.

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.

- The storage class assignments and lifecycle transition rules described in Section 3.3 will be reviewed and updated on the basis of real-world access patterns and cost optimisation findings gathered during piloting.
- The workflow provenance capture described in Section 3.2 will be extended to cover the full range of configuration artefacts, including pipeline parameters, reference databases, and LLM prompts, that are necessary for complete computational reproducibility.
- The geographic data residency architecture will be revisited to evaluate whether a multi-region or per-country bucket configuration is required to meet the data residency obligations applicable to each participating institution.
- The ENS deeper integration will be further studied to support the registration of inter-object provenance relationships, moving the LTAS towards a linked data architecture in which the full provenance graph of any diagnostic result can be queried using standard semantic web tools.
- The formal ontology being developed within the project, grounded in authoritative biomedical ontologies, will be used to extend the PID scheme towards semantically descriptive identifiers that encode not just a unique reference but also the type, role, and relationships of each archived entity.
- The ASBAC access control layer will be further developed to allow policies to reference semantic attributes of archived objects as described in the project ontology, enabling richer and more expressive governance rules.

This project is funded by the European Union, MSCA SE, Grant Agreement No 101182801.