



D4.4 Models/guidance for improvement of data logistics

WP4 Enhancement & consolidation of digitisation of workflows

Author(s): Ivana Ferenčák, Irena Tabain



**Co-funded by
the European Union**

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

Copyright message

© HERA 2 Consortium, 2024

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorized provided the source is acknowledged.

Document Information

Grant Agreement Number	101102440	Acronym	HERA 2	
Full Title	Models/guidance for improvement of data logistics			
Call	EU4H-2022-DGA-MS-IBA-1			
Topic	EU4H-2022-DGA-MS-IBA-01-02			
Type of Action	EU4H-PJG			
Start Date	01.10.2022	Duration (in months)	48	
HaDEA Project Officer	Diana Burghardt			
Task	T3.2			
Work Package	3			
Date of Delivery	Contractual	M42/March 2026	Actual	M42/March 202
Nature	R- Document report	Dissemination Level	PU - Public	
Lead Beneficiary	CIPH			
Lead Author	Ivana Ferencak	Organization	CIPH	
Other authors	Irena Tabain			
Reviewer(s)	Cabal Rosel Adriana, Judit Henzko, Grigoris Spanakos			

Document History

Version	Issue Date	Stage	Changes	Contributor
0.1	05.02.2026	First Draft		Ivana Ferencak, Irena Tabain
0.2	11.03.2026		Refined and sent for feedback	All Partners
1.0	20.3.2026	Final Version	All Feedback incorporated	All Partners

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY	4
2. BACKGROUND INFORMATION	5
3. METHOD	6
4. SAMPLE MANAGEMENT	6
6. REPORTING.....	9
7. CONCLUSIONS	9

1. EXECUTIVE SUMMARY

This deliverable evaluates the current data logistic infrastructure, practices and capabilities when supporting polymerase chain reaction (PCR) and the whole genome sequencing (WGS) workflows in each partner's institute. The analysis reveals that even though each partner developed functional system, tailored to their local needs and institutional context, there are significant heterogeneity in data management approaches, computational infrastructure, and standardisation. These differences limit interoperability and make integrated genomic surveillance challenging with the increased risk for data loss or misinterpretation.

There are clear opportunities for harmonisation of identifiers, metadata, quality control (QC) procedures, data storage, and reporting formats, which could strengthen data quality assurance and improve efficiency. This report identifies key gaps and proposes actionable recommendations to establish a more cohesive, standardised, and resilient data logistics framework for WGS-based surveillance across the consortium.

2. BACKGROUND INFORMATION

WGS and PCR have become central tools for infectious disease surveillance and outbreak detection. A key priority of the HERA2 project is to consolidate, sustain and integrate WGS and PCR capabilities built under HERA1 into routine surveillance and outbreak investigation activities beyond SARS-CoV-2 in the EU. To fully exploit WGS data, interoperable databases, stable data exchange platforms, visualization tools, reporting solutions, and automated alert systems are required, all operating on harmonized epidemiological and molecular data.

The work conducted in HERA2 shows that the surrounding digital infrastructure is still fragmented and heterogeneous with substantial differences between partners in how samples, sequencing data and results are handled and integrated into national and EU public health systems. In this context, within work package 4 “Enhancement & consolidation of digitization of workflows” we analysed the implemented processes, identified gaps in data logistics, and summarized partner needs for data analysis, databases, storage, exchange, alerts and connectivity with local databases, and visualization of WGS data for outbreak monitoring and investigation.

Based on this analysis this deliverable assesses the current situation and propose recommendations to improve for data logistic across the consortium. The partners included are Austrian Agency for Health and Food Safety GmbH (AGES, Austria), National centre for Public Health and Pharmacy (NNGYK, Hungary), Croatian Institute for Public Health (CIPH, Croatia), and four laboratories from Greece: KEDY-AMR Bacteriology Laboratory, Regional Public Health Laboratory of Thessaly, Regional Public Health Laboratory of Crete- Bacteriology and Regional Public Health Laboratory of Crete- Virology.

3. METHOD

Information on data logistics was obtained from structured questionnaire in a table format, completed by all consortium partners, and returned to the authors by email. The questionnaire captured end-to-end workflow from sample recipient through laboratory processing, WGS data integration, analysis, storage, and reporting, with specific emphasis on identifiers, metadata, systems integration, and governance.

4. SAMPLE MANAGEMENT

4.1. Current status and Gap Analysis

Across the partner network, different coding and sample identification systems are used, with no shared identifier standard. This heterogeneity creates interoperability challenges when cross-institutional data sharing or joint analyses are required. Storage solutions range from centralized database systems to hybrid paper-digital archives and purely paper-based filing supplemented by spreadsheets, leading to differences in data security, searchability, and archival consistency.

Austria uses two different integrated laboratory information management systems (LIMS) platforms depending on the reference laboratory. However, the software is outdated and cannot support a fully digital workflow from sample reception to the generation of the individual WGS reports. Next-generation sequencing (NGS) library management is not integrated into the LIMS, the spreadsheets are used instead, increasing the risk of human error. After NGS processing, results must be transferred manually into the LIMS, even when they are generated automatically by bioinformatics pipelines, which complicates integrated genomic surveillance and delays analysis. Case-based epidemiological data are managed in the Electronic Reporting System (EMS), that is not connected to the LIMS, further hindering data consolidation. Current LIMS cannot handle minimum spanning tree figures, so key genomic relationship information must be transcribed into text, increasing workload and the likelihood of inconsistencies.

Ongoing development of an NGS- specific solution will improve cluster detection and does not address the underlying issues in individual sample tracking and end-to-end digitalization.

Hungary also uses a LIMS, but NGS data are not managed within this system. Instead, NGS data are stored and processed on a separate dedicated server, resulting in a fragmented data environment without a single end-to-end audit trail. Reports are prepared and handled manually, which increases the risk of human error and reduces efficiency. Case-based epidemiological data are managed in a national EMS that operates independently of the LIMS and has no direct interoperability, complicating data consolidation, cross-referencing, and timely outbreak assessment. Although the development of a dedicated NGS integration solution is currently underway, it does not yet resolve the fundamental challenges across the full sample processing workflow from receipt through laboratory procedures to data analysis and reporting.

Croatia uses LIMS for routine laboratory work, and similarly to Hungary, NGS data are managed on a separated server and SQL database. This database currently manages sample and sequencing results but does not support library management or any wet lab procedures. There is no library management in the LIMS, only sample and sequencing results. Case-based epidemiological data are managed within the electronic reporting system (NAJS), which runs independently of the LIMS and has no direct interoperability again complicating data consolidation, cross-referencing, and timely analysis.

Greece currently relies on manual data entry but is in the process of implementing a LIMS at the Central and Regional Public Health Laboratories, an effort that has been ongoing for the past year. Similar to the other partner networks, NGS analysis results are not yet integrated into the LIMS, though there are plans for future incorporation. Currently, sequencing data generated across multiple platforms is uploaded directly to a robust in-house computational cluster. While this dedicated infrastructure allows in-house pipelines to efficiently and automatically generate analysis reports in spreadsheets and HTML format, the separation of NGS data from the LIMS maintains a fragmented overarching data environment. Furthermore, case-based epidemiological data are collected and managed independently by the epidemiological department. Because there is no direct interoperability between the laboratory and epidemiological systems, data consolidation and cross-referencing remain challenging.

Most partners use email to request missing documentation from sending institutions, but there is no standardized communication protocol or escalation pathway for data inconsistencies detected at sample intake. Overall, the lack of unified identifiers, incomplete LIMS coverage of NGS workflows, and weak integration with epidemiological reporting systems limit the ability to perform robust, integrated genomic surveillance.

4.2. Proposed Improvements

To strengthen the molecular and epidemiological surveillance framework based on WGS, the development of standardized, harmonized, and interoperable procedures for sample management is strongly recommended. A consortium-wide unique sample identifier (USI) scheme should be established, incorporating, for example, partner institution code, sample year and sequential number (e.g., AT-2026-001234). This would enable seamless cross-institutional tracking while allowing partners to keep local numbering for internal operations, with each partner maintaining a local-to-USI mapping table in their LIMS or database.

Given that a single common LIMS for all partners is unlikely, a shared relational database schema for core sample metadata should be agreed. This schema should at least cover patient and case identifiers (pseudonymised when required), material type, collection and arrival dates, processing status, responsible staff and key NGS-specific fields such as platform, run ID, library preparation protocol, and reference genome used. Clear data dictionaries and controlled vocabularies should accompany this schema to prevent semantic inconsistencies.

To standardise sample quality assessment and make data comparable, a central data quality protocol should be established, ideally in the form of a shared checklist that covers both laboratory and bioinformatics checkpoints. Guidance should define critical steps and QC metrics within the bioinformatic workflow, including QC of raw sequencing data and explicitly filtering criteria for data evaluation. Standardized documentation forms should capture each step of the bioinformatic analysis, including tools and parameters used for mapping, genome assembly, variant calling, clustering, and classification. This will improve transparency, reproducibility, and cross-laboratory comparability.

A harmonized reporting template with predefined fields and selectable options should be developed for: molecular-level results (WGS-derived genomic data) and associated epidemiological metadata. Standardization of these templates will improve consistency and comparability of results across institutions and countries and will facilitate automated data ingestion into surveillance systems.

The documentation management process for sample consignments should be formalised by defining a standard workflow for handling missing or incomplete documentation. This could include automated email notifications to the sending institution within a defined time window after sample arrival, mandatory response deadlines, and clear interim sample holding protocol.

From a governance and sustainability perspective, it is advisable to establish a consortium-level data stewardship framework defining roles, responsibilities, and data ownership along the data lifecycle. Regular inter-laboratory audits and external proficiency testing should be implemented, alongside structured and recurring training programmes for laboratory and bioinformatics personnel. A long-term sustainability plan should address infrastructure maintenance, software updates, and financial planning. Data protection and regulatory compliance (for example, GDPR) must be embedded into system design, including pseudonymization, role-based access control, encrypted data transfer, and secure storage policies. Automated validation scripts that check sample data against predefined rules (such as identifier format, completeness of mandatory metadata, and logical date relationships) are recommended to reduce manual data entry errors and to ensure consistency before processing begins.

As a complementary measure focused on pathogen-specific cluster detection and genomic surveillance, an integrated genomic surveillance framework is proposed, combining an R-based analysis pipeline together with a dashboard for real-time cluster monitoring. This approach can improve situational awareness and allow rapid identification of emerging clusters, even if some limitations of the current LIMS solutions remain.

5. DATA ANALYSIS AND BIOINFORMATICS

5.1. Current status and Gap Analysis

Partners currently employ heterogeneous bioinformatics pipelines, with differing tools, parameters, and workflows, and there is no cross-institutional comparison of pipeline performance. This variability can lead to reproducibility issues and difficulties when comparing results across laboratories. Storage solutions for analysis outputs are also diverse, with some partner relying on local drives and unclear or undefined retention policies which increases the risk of data loss or untraceable analyses.

There is no standardisation in QC parameters and tools used for WGS data, and requirements for sample metadata differ between partners. As a result, aggregating data across partners and performing joint outbreak analyses is challenging. All partners upload data to public repositories, but this is done inconsistently and mainly in connection with scientific publications rather than a routine component of genomic surveillance.

5.2. Proposed Improvements

It would be beneficial to develop consortium-wide, largely automated WGS analysis pipeline framework that can be implemented locally but adheres to common specifications. A shared minimal specification should define required processing steps, (for example, QC, trimming, mapping or assembly, variant calling, and typing/lineage assignment), expected outputs, and basic parameter ranges, while allowing flexibility for local optimisation. This should be coupled with versioning and documentation strategy that records pipeline versions, parameter files, and change logs.

A centralised or logically centralised data repository for WGS analysis outputs should be established to facilitate data sharing and joint analyses. The repository should support role-based access, contain harmonised data, and link raw data, intermediate files, and final results for each sample. To ensure comparability, common QC standards and thresholds should be defined, along with a set of mandatory metadata elements that must accompany each dataset when shared within the consortium or submitted to public repositories.

6. REPORTING

6.1. Current status and Gap Analysis

Reports currently leave laboratories in different formats and with varying levels of detail. Combined with the lack of standardised metadata, this heterogeneity makes integrated surveillance at national or EU level very difficult. Approval processes and verification steps differ between partners, and often rely on manual procedures, which are time-consuming and error prone. In addition, some partners do not bill for their services, which may complicate sustainability planning for WGS-based surveillance activities.

6.2. Proposed Improvements

Standardised report templates should be created for WGS-based results, with clearly defined sections for laboratory findings, genomic interpretation (for example, lineage or cluster assignment and their limitations), and associated epidemiological information. Whenever possible, report generation and approval workflows should be integrated into the LIMS, with electronic sign-off and automatic transfer of key data to surveillance systems. Providing both human-readable formats (such as PDF) and machine-readable outputs (for example, structured CSV or JSON) would facilitate automated aggregation and analysis.

7. CONCLUSIONS

Joint efforts are needed to move towards a single, harmonised sample identification system across European health agencies to support integrated genomic surveillance. While such a fully unified system is an ideal long-term objective is unlikely to be realised in the immediate future, important progress can be made through incremental harmonisation. As a pragmatic, short-term measure, samples that are handled collaboratively between partners should at least carry a shared international or Europe-wide identifier that can be mapped to local systems. This, combined with harmonised metadata, QC standards, and reporting templates, will significantly improve the robustness and interoperability of WGS-based surveillance workflows.