



D5.4 Report on sharing the generated sequencing data in the public domain

Work Package 5: Implementation and Harmonization of Alert systems and workflows

Author(s):

Daniel Polzer, Adriana Cabal



This project has received funding from the European Union's Health and Digital Executive Agency (HaDEA) under Grant Agreement No 101102440.

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This deliverable contains original unpublished work except where clearly indicated otherwise.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	Consolidation of WGS/RT-PCR and infrastructure processes in surveillance and outbreak investigation activities			
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)	42	
HaDEA Project Officer	Diana Burghardt			
Task	T5.1			
Work Package	5			
Date of Delivery	Contractual	M34/07.2025	Actual	M34/07.2025
Nature	R - Document report	Dissemination Level	PU - Public	
Lead Beneficiary	AGES			
Lead Author	Daniel Polzer	Organization	AGES	
Other authors	Adriana Cabal			
Reviewer(s)	Melissa Hagendorn			

Document History

Versio	Issue Date	Stage	Contributor
0.1	11.06.2025	Draft	Daniel Polzer
1.0	13.06.2025	Draft ready for partners to give feedback	all
1.1	04.07.2025	Feedback by all partners given and integrated	all
2.0	12.01.2026	Revision after feedback from project officer	Jennifer Rao

TABLE OF CONTENTS

Executive Summary 5

Background Information 5

Current Data Sharing Policies 6

 List of Platforms data is uploaded to..... 6

Conclusion 8

Executive Summary

Sequencing data can be used for many purposes. While some of them only require data from a small number of samples, the vast majority analyses profit from a large number of sequences from a variety of countries. These analyses for example allow for investigation of cross-border outbreaks, cross-domain outbreaks (One Health) as well as monitoring of general evolutionary trends of pathogens. Therefore, it is a necessity for countries to share their generated sequencing data. This deliverable aims to give an overview of the data shared by this consortium as well as the challenges that stand in the way of sharing data.

Background Information

The sharing of sequencing data is fundamental to advancing our understanding of pathogens and improving global public health responses. Genomic data provides critical insights into pathogen evolution, transmission dynamics, drug resistance, and outbreak detection. Timely and transparent sharing of this information enables coordinated surveillance and response efforts, supports research and innovation, and enhances preparedness for current and future health threats.

A unique advantage of integrating sequence data from multiple countries and across sectors, such as human health, veterinary medicine, and agriculture, is the ability to conduct One Health analyses. For example, understanding the transmission of zoonotic diseases like avian influenza, antimicrobial resistance genes, or foodborne pathogens requires data from both human and animal populations, as well as environmental sources. Such integrated datasets allow for:

- Detection of cross-species transmission events.
- Identification of shared or divergent evolutionary patterns in pathogens across regions or hosts.
- Mapping of global transmission routes and identifying hotspots of emergence.
- Coordinated risk assessments that inform policy and intervention across borders and sectors.

This comprehensive approach would not be possible with data confined to a single country or domain. Therefore, fostering international collaboration and encouraging cross-sectoral data sharing are not just best practices, they are essential steps toward building resilient, informed, and proactive public health systems worldwide.

However, sequencing data alone is not sufficient. The value of such data is significantly enhanced when it is accompanied by relevant metadata, such as the date and location of sample collection, host species, clinical outcomes, and environmental context.

The requirements for metadata vary between the platforms for sharing sequencing data. This leads to additional overhead in adapting the metadata format for different use cases. Harmonization efforts between these platforms would be appreciated.

Sharing metadata must be approached with care however, as it raises important considerations regarding privacy, data ownership, ethical use, and compliance with national and international regulations. Anonymization, informed consent, secure data platforms, and clear data governance frameworks are essential to ensuring responsible data sharing. This part is often challenging for sequence providers. A clear legal framework and guidelines in laymen's terms from the European authorities would be a great contribution to make sharing of data more accessible.

Current Data Sharing Policies

The members of this consortium follow different policies in regards to data sharing. Additionally, these policies also vary by pathogen. Generally speaking: however data is shared in case of outbreak investigations and surveillance purposes as well as for publications. Depending on the purpose and the pathogen an appropriate platform is chosen, on which the data is shared. In regards to metadata published in addition to the sequence, the members of this consortium report on varying granularity depending on the pathogen and the purpose. However, often it is not clear for us if we are legally allowed to share some information that we might have. For instance, when uploading the sequence of the CoVid genome from a patient, we are unsure if we are legally allowed to share/publish/upload to public databases some of this metadata, even if it is anonymized. Now while we suspect that the full name and place of living might fall under special categories of personal data Article 9 (<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e2051-1-1>), in regards to health data, it is not so clear for us how to proceed if the combination of multiple metadata fields may identify a person, even if the metadata is “abstracted” e.g. place of living not as a concrete address but maybe a ZIP Code or province. It could very well be that only 1 or 2 people were infected at the time of sampling in this ZIP code in total making it obvious to which patient the sample belonged. So ideally we would hope for some kind of document that tells us which metadata we are always allowed to share (maybe age, sex, sampling time, country of sampling), some that we are never allowed to share (maybe full name, place of living, sexual orientation, religious beliefs, immigration background or sex work), and which we are allowed to share if some criteria are fulfilled (for example if there are at least 10 potential people/patients in a geographical resolution we can share this resolution, else we need to share a more zoomed out one).

In the following, we give an overview of the platforms used by this consortium.

List of Platforms data is uploaded to

There are numerous platforms available for sharing data with the scientific community. Below, we outline the platforms currently used by members of this consortium. While each partner selected their preferred option independently, for certain pathogens there is a well-established consensus within the scientific community regarding which database should be used. In those cases, we adhered to that consensus and chose the recommended databases.

- **Epipulse** (<https://epipulse.ecdc.europa.eu/>) is the official ECDC’s online portal for European public health authorities and partner organisations to collect, analyse, share, and discuss infectious disease data for threat detection, monitoring, risk assessment and outbreak response. We use Epipulse to respond to comments posted by other EU/EEA countries in regards infectious disease threats. We use it for data information exchange, including WGS. This is particularly useful for cross-border outbreaks. WGS data can be uploaded with anonymized metadata.
- **MYGenomeSurv**: <https://www.medizin.uni-muenster.de/migenomesurv/home.html>) Austria is part of this network integrating partners in the DACH region for some years now and we upload WGS data for certain pathogens of interest, namely, foodborne bacteria, VRE and M. tuberculosis. Alerts are received by the consortium partners once a match between related isolates is found.

- **PubMLST** (<https://pubmlst.org/>): a website hosts a collection of open-access, curated databases that integrate population sequence data with provenance and phenotype information for over 130 different microbial species and genera. The system is based on the gene-by-gene analysis of microbial genomes, with each deposited sequence annotated and curated to identify the genes present and systematically catalogue their variation. We use PubMLST to share new Sequence Types of bacterial pathogens. This specific database was used since it is the de facto standard for sharing this kind of data. We also use it to access cgMLST schemes and other typing schemes, which can be used to categorize sequenced isolates.
- **European Nucleotide Archive (ENA)** (<https://www.ebi.ac.uk/ena/browser/home>): ENA is an open, supported platform for the management, sharing, integration, archiving and dissemination of sequence data. When sharing raw data from a sequencing run there are two commonly used databases to share them. One of them is the European Nucleotide Archive, the other is the Sequence read Archive (SRA). Both are commonly used and also were used in this consortium. The choice when to use which of these two to use depends on personal preferences as well as requirements of some journals. Since both databases mirror their raw read data, the choice does not have a significant impact in the long run. However, ENA also allows the upload of processed data from a sequencing run (e.g. assemblies) which are not mirrored.
- **Sequence read Archive (SRA)** (<https://www.ncbi.nlm.nih.gov/sra>): The NCBI's (National Center for Biotechnology Information) counterpart to the ENA. It is run by the United States' National Institute of Health. Since it is run outside of Europe this can be on consideration to choose the previously mentioned ENA instead. However as also mentioned above already the choice between these two databases is more a matter of taste/convenience and since they mirror each other does not impact data sharing in the long run.
- **Pathogenwatch** (<https://pathogen.watch/>): it provides a platform for comparing pathogen genome assemblies from around the world, integrating diverse data sets with rich representations. While also sometimes used to share data with the scientific and public health community, in this project it was mainly used to analyze generated genomes, and in particular to obtain plasmid replicon data.
- **GISAID** (<https://gisaid.org/>): it is a database for sharing assembled viral genomes of Influenza, CoVid, RSV, Monkey Pox, as well as a few Arboviruses. It is the de facto standard for sharing Influenza and CoVid genomes. Therefore, it was used to share genomes of these pathogens by members of this project consortium.
- **Genbank** (<https://www.ncbi.nlm.nih.gov/genbank/>): it is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Conclusion

Using data from different sources and fields allows for a broader understanding of situations and is fundamental for improving global public health responses as well as for research purposes. This requires that generated sequences are also shared within the public domain. Sequencing data without metadata is however without any value and the legal framework for sharing this metadata varies between disciplines, countries and pathogens, leading to a complex situation. While the members of this consortium share sequencing and metadata to the best of their abilities, clear guidelines and a framework – formulated in laymen's terms – from the relevant European authorities would be greatly appreciated.

Currently this consortium shares data mainly for surveillance, outbreak investigation as well as research purposes. For this, different platforms, with different metadata requirements are used. A possible goal for the future would be an automated upload to these databases based on pre-approved criteria.