D4.2 Roadmap of Database Landscape

Work Package 4: Enhancement & consolidation of digitisation of workflows

Author(s): Michalis Polemis, Kassiani Gkolfinopoulou, Grigorios Spanakos, Ioanna Spiliopoulou

Co-funded by the European Union

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

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	Marie de Looz-Corswarem		
Task	T4.2		
Work Package	4		
Date of Delivery	**Contractual** M14/Nov 2023	**Actual**	M15/Dec 2023
Nature	R – document, report	**Dissemination Level**	PU - Public
Lead Beneficiary	EODY		
Lead Author	Michalis Polemis	**Organization**	EODY
Other authors	Kassiani Gkolfinopoulou, Grigorios Spanakos, Ioanna Spiliopoulou		
Reviewer(s)	Polzer Daniel, Ivana Ferenčak		

Document History

Version	Issue Date	Stage	Changes	Contributor
0.1	02.10.2023	First draft generated in Zadar steering meeting		Whole consortium
0.2	17.11.2023	Updated version	Survey results added	GS, MP
0.3	28.11.2023	Updated version – Draft document		GS, MP
0.4	02.12.2023	Quality checked version		GS, IS, KG, MP
0.5	04.12.2023	Version accepted by CPHL		
1.0	19.12.2023	Consolidated and quality-checked version		Whole consortium

Contents

1. Summary

Molecular sequence data have proved their usefulness in outbreak investigations and are considered essential for Public Health surveillance and intervention. ECDC has set as a priority the implementation of WGS by the EU member states. WGS data are in digital form and consist of raw data, produced by the sequencer and information derived from bioinformatics analysis of the raw data. In order for molecular data to be useful, they must be linked to epidemiological data, stored in databases and be available to the scientific community and competent Public Health authorities for molecular epidemiology investigations.

Databases for molecular sequence data submission by European laboratories are ENA, EFSA One Health WGS System, ECDC WGS Surveillance System, and GISAID. The former is a general-purpose open database for submission of all kinds of sequences, while the latter is an open database for submission of data regarding influenza, RSV, Mpox, arbovirus, and coronavirus sequences. EFSA One Health WGS System and ECDC WGS Surveillance System are official restricted databases, accessible only by nominated National Public Health authorities, and reference laboratories should submit their sequence data on them.

As concluded from a survey, consortium partners currently do not use a database for data storage, and this is an opportunity to design a database that can interoperate with the European databases and has the ability to submit data programmatically. A database that stores raw sequence data (e.g. fastq files), processed data (e.g. assemblies, annotations, strain designations), and a limited number of metadata would be able to fulfil the basic requirements. To be epidemiologically useful however, additional information is necessary. This additional information can be included in the molecular database or linked to a separate epidemiological database.

Depending on the particular needs and the structure of national Public Health agencies, each country should decide the strategy to be adopted.

1.1. Work package 4 – Summary

WP4, Enhancement & consolidation of digitisation of workflows – for this WP the specific needs of partners for analysis, databases, storage, exchange, alerts and connectivity/data exchange/integration with existing local databases, visualisation of WGS data for outbreak monitoring and investigation will be summarised. After a survey of the various national databases and comparison with their European equivalents, a gap and needs analysis will be prepared. This will be done by applying a structured and transparent mixed-methods approach. The results will feed into a road mapping of relevant progress and processes to reach the next level of harmonisation and networking. The documentation includes a matrix for the collection of procedures, risk assessments, biosafety, and biosecurity documentation. This will provide the basis for improved procedures leading to better outbreak detection and preparedness for

future public health threats in line with current European public health best practice structures. This methodological approach was selected because it represents the current state of the art in the field.

2. Background information

2.1. Data

Currently, molecular methods (qPCR in its various forms) for pathogen detection are advancing over classical microbiological methods as they require less laboratory work, provide faster results, and exhibit higher sensitivity and specificity. Furthermore, in the last decade the lower cost of Next Generation Sequencing (NGS) and the ability to sequence hundreds of millions of nucleotide bases have provided the tool for Whole Genome Sequencing (WGS) of pathogens (mainly bacteria and viruses), thereby providing higher discrimination ability for molecular epidemiology. Currently WGS is considered the reference method for outbreak investigation due to its high discriminatory ability. The use of sophisticated equipment for qPCR and WGS, which provide results in digital form, allows for computer-assisted processing and faster dissemination of results. Nowadays, WGS data can be used to closely monitor infectious pathogens, detect outbreaks or the source of infection, and provide valuable information for surveillance, prevention, and control. The high frequency of international travel and food transportation across nations may result in transnational epidemics, thus data sharing between national Public Health authorities is essential for surveillance. WGS data, along with a minimal set of crucial information among partners, may facilitate early detection of Public Health threats. For all these reasons, ECDC has set as priority the implementation of WGS by the member states.

There are several types of WGS data, e.g. fastq files (raw sequence data) and files produced by the bioinformatics pipelines that analyze the raw data. These data can be bam/sam files, fasta files (contigs or genome assemblies, annotations), cgMLST, strain nomenclature, AMR profiles, depending on the degree of data processing and sequencing objective. Inconsistencies have been observed between different sequencing platforms and bioinformatics pipelines, so providing information about the sequencing and analysis process is always necessary.

For WGS data to be useful for molecular epidemiology, at least information about sample type, place, and time should be available.

2.2. European WGS Databases

2.2.1. European Nucleotide Archive Database (ENA)

ENA is a public, general-purpose database for nucleotide sequence data. ENA currently accepts all types of nucleotide sequence data derived from any type of specimen. Since the database is publicly accessible, it is the first place to deposit nucleotide sequences. It offers the ability to search data in several ways and most importantly provides the tools to find and retrieve similar sequences from the vast number of deposited sequences. ENA is one of the three members of the International Nucleotide Sequence Database Collaboration, with NCBI in USA and DDBJ in Japan. All three databases are synchronized regularly, so they contain essentially the same data.

ENA accepts data from sequence experiments at various degrees of processing, e.g. raw NGS data (FASTQ, CRAM or BAM format, provided the data are trimmed and demultiplexed), assemblies, annotations, and simple sequences.

Data submission requires at least the time and country or sea the sample was collected (https://www.ebi.ac.uk/ena/browser/view/ERC000011). Beyond the default data checklist, depending on the type of sample and the purpose, there are 51 different metadata checklists with mandatory and optional fields, many of which have predefined list of entries (https://www.ebi.ac.uk/ena/browser/checklists).

Of special interest for Public Health authorities are the following specific metadata checklists:

ENA Influenza virus reporting standard checklist

ENA virus pathogen reporting standard checklist

ENA prokaryotic pathogen minimal sample checklist

ENA parasite sample checklist

COMPARE-ECDC-EFSA pilot human-associated reporting standard

COMPARE-ECDC-EFSA pilot food-associated reporting standard

Data submission may be performed through a user interface, using Webin-cli or programmatically, as indicated in Table 1.

Table 1. Options for ENA data submission (https://www.ebi.ac.uk/ena/browser/submit).

	Interactive	Webin-CLI	Programmatic
Study	Y	N	Y

Sample	Y	N	Y
Read data	Y	Y	Y
Genome Assembly	N	Y	N
Transcriptome Assembly	N	Y	N
Template Sequence	N	Y	N
Other Analyses	N	N	Y

2.2.2. EFSA One Health WGS System

The European Food Safety Authority (EFSA) is an agency of European Union that aims to provide scientific advice on food safety. The EFSA database is only accessible to the competent national food safety organizations of the European member states. The database currently collects WGS data for *Salmonella enterica*, *Listeria monocytogenes* and *Escherichia coli*, including Shiga toxin producing *E. coli* (STEC), from non-human sources, with the aim of including WGS data for more pathogens such as *Campylobacter*. Interoperability with the ECDC system allows detection of multi-country clusters of isolates of various origin. This is accomplished through automatic exchange and comparison of cgMLST profiles and a limited set of metadata.

Data submission can be performed through a user interface or programmatically. Currently, the database is designed to analyse only Illumina paired-end sequencing data, with other features planned for future releases.

The data listed in Table 2 are collected in the EFSA One Health WGS database.

Table 2. Experimental data elements (https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2022.EN-7413).

Element	Term name	Description	Rule
Local Raw Reads ID	localRawReadId	Unique identification code of the experiment.	Mandatory
Instrument model	instrumentModelCode	the instrument model used for generating the experiment.	Optional
Isolate species	isolateSpeciesCode	the species of the isolate from which the experiment was generated	Mandatory
Subtype	serotypeId	the coding system to describe the subtype of the isolate from which the experiment was generated	Optional

Layout	libraryLayoutCode	report if the experiment generated “single reads” or “paired-end reads”	Mandatory
FASTQ file 1 name	fastQ1FileName	It contains the name of the FASTQ file.	Mandatory
FASTQ file 1 MD5 checksum	fastQ1Md5	It contains the MD5 Checksum of the FASTQ file.	Mandatory
FASTQ file 2 name	fastQ2FileName	It contains the name of the FASTQ file paired to FASTQ file 1.	Conditional
FASTQ file 2 MD5 checksum	fastQ2Md5	It contains the MD5 Checksum of the FASTQ file paired to FASTQ file 1.	Conditional

The EFSA system requires either the upload of fastq files so that data processing is performed on EFSA computational cloud using the EFSA One Health WGS analytical pipeline, or the use of the EFSA pipeline to analyse the sequences in-house, followed by the upload of the results to the EFSA platform. The latter approach also allows data to be uploaded programmatically to the EFSA platform. In case FASTQ files are uploaded, they are removed from the servers after data analysis is complete.

Also, it is required to submit additional data. Some of these data are mandatory, whereas other are optional and are listed in Table 3.

Table 3. Additional data submitted to the EFSA database (https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2022.EN-7413).

Element	Term name	Description	Rule
Sample			
SamplingID	sampling_local_id	The unique identification code of the sample taken in the data provider organisation.	Mandatory
Country of sampling	sampling_country_id	The country in which the sample was taken.	Mandatory
Coded description of the matrix of the sample taken	sampling_matrix_code	Description of the food, feed, animal and environmental category.	Mandatory
Additional information of the matrix of the sample	sampling_matrix_free_text	Additional information on the sample.	Mandatory
Sampling year	sampling_year	The year of sampling.	Mandatory
Month of sampling	sampling_month	The month of sampling.	Optional
Day of sampling	sampling_day	The day of sampling.	Optional
Programme type	programme_type_id	The type of programme as part of which the sample has been collected.	Optional
Programme info	programme_info_id	Report additional information on the sampling programme.	Optional
Sampler	sampler_id	The type of body that performed the sampling.	Optional
Sampling point	sampling_point_id	The point of sampling in the food chain.	Optional
Country of origin of the sample taken	sampling_country_origin_id	The country of origin of the sample taken for laboratory testing.	Optional
Area of sampling	sampling_area_id	The area, or region, or province in the country.	Optional

Element	Term name	Description	Rule
Isolate			
Isolate ID	isolation_local_id	The unique identification code of a culture of a biological agent, isolated from a specific sample taken.	Mandatory
Isolation year	isolation_year	The year of isolation of the biological agent from the culture.	Optional
Isolation month	isolation_month	The month of isolation of the biological agent from the culture.	Optional
Isolation date	isolation_day	The day of isolation of the biological agent from the culture.	Optional

2.2.3. ECDC WGS Surveillance System

The European Centre for Disease Prevention and Control (ECDC) is the Public Health agency of the European Union. ECDC has integrated WGS surveillance data into the already existing TESSy (The European Surveillance System) database. Due to the multifactorial nature of this surveillance system, all WGS-related fields are optional. Access in the database is restricted and data can only be submitted by the competent authorities of the member states of the European Union.

Based on the latest TESSy metadata report, ECDC has taken the necessary provisions to be able to collect WGS data / results for *Salmonella*, *Listeria*, *E. coli, N. meningitidis*, *Campylobacter* & *Mycobacterium*.

As a representative example of the database structure, we present the one related to *Salmonella* isolates (table 4).

To upload WGS data to TESSy, ECDC has developed a multi-platform (Windows, MacOS or Linux) desktop client application which runs on Java (JRE of at least version 8). It is intended for isolate-based reporting to TESSy, for isolates with associated WGS data. It can be configured to submit data to TESSy (mandatory), ENA (optional), and ECDC SFTP server (optional). It can submit assemblies as well as raw read data.

The application requires some initial setup where the parameters for importing and uploading are defined. After the initial setup, data are filled in an isolate table either by configured import, copy/paste or manually. Sequence data are linked to the entries in the table using a link function and the data can be submitted to TESSy and one or both of ENA and ECDC SFTP.

The app comes bundled with configurations for LISTISO (*Listeria*), SALMISO (*Salmonella*) and ECOLIISO (*E. coli*) TESSy subjects; more will eventually be supplied when WGS-based surveillance for more pathogens is implemented at ECDC.

Table 4. Data submitted to the ECDC WGS Surveillance System (https://www.ecdc.europa.eu/en/publications-data/tessy-metadata-report).

Variable	Full name	Description	Required
Age	Age	Age of patient in years as received or at the date of sampling.	True (Warning)
AgeMonth	Age in months	Age of patient in months as reported in the national system for cases < 2 years of age at the time of disease onset.	False
AntigenH1	AntigenH1	Only flagellar (H) antigen - phase 1- of the antigenic formula of the pathogen which is the cause of the reported disease. Free text but need to follow the Kauffmann-White scheme.	False
AntigenH2	AntigenH2	Only flagellar (H) antigen - phase 2- of the antigenic formula of the pathogen which is the cause of the reported disease. Free text but need to follow the Kauffmann-White scheme.	False
AntigenOText	AntigenOSALM	Only somatic (O) antigen of the antigenic formula of the pathogen which is the cause of the reported disease. Free text but need to follow the Kauffmann-White scheme.	False
CarbaGene	Carbapenemase Gene	Carbapenemase gene(s) identified	False
CaseId	Case identifier	A unique identifier for each case within the data source / surveillance system related to the isolate, so that isolate records can be linked to case records.	False
DataSource	Data source	The data source (laboratory) that the record originates from.	True (Error)
DateOfReceiptReferenceLab	DateOfReceiptReferenceLab	Date of receipt in reference laboratory or typing laboratory with reference function.	True (Error)
DateOfReceiptSourceLab	DateOfReceiptSourceLab	Date of receipt in source laboratory, i.e. the laboratory the sample was first sent to.	False
DateOfSampling	Sample date	Date the sample from which the isolate was derived, was taken.	True (Warning)
DateUsedForStatistics	Date used for statistics	The most epidemiologically relevant date for the isolate. Equal to the date of sampling if available. If not, equal to the date of receipt in the source lab, and if that is not available, the date of receipt in the reference lab.	True (Error)
ECDCCaseId	ECDC Case identifier	TESSy globally unique identifier assigned upon upload of the case, so that isolate records can be linked to case records.	False
ESBL	ESBL	ESBL and/or AmpC confirmed with phenotypic or genotypic tests.	False
ESBLGene	ESBL Gene	Extended-spectrum beta-lactamase gene(s) identified	False
Gender	Gender	Gender of the reported case.	True (Warning)

Variable	Full name	Description	Required
GenoSerotype	GenoSerotype-SALM	Salmonella serotype predicted from molecular methods.	False
Imported	Imported	Having been outside the country of notification during the incubation period of the reported disease.	False
PlaceOfResidence	Place of residence	Place of residence of patient at the time of disease onset. Select the most detailed NUTS level possible. UNK is allowed.	False
ProbableCountryOfInfection	Probable country of infection	If Imported=Yes: One entry for each country/region visited during the incubation period of the disease should be provided. The list can be empty even if the case is known to be imported. If there is more than one country N/A should be used in the empty repeated fields.	False
RecordId	Record id	Unique identifier for each isolate within the data source / lab system.	True (Error)
RecordType	Record type	Structure and format of the data (case based reporting or aggregate reporting).	True (Error)
RecordTypeVersion	Record type version	Indicates the version of the Record type used in the reported batch. If no RecordTypeVersion is provided in the batch, it is set automatically with current version of the Record type. RecordTypeVersion is required when no metadata set is provided at upload or when a RecordTypeVersion, other than the current one, needs to be used.	False
ReportingCountry	Reporting country	The country reporting the record.	True (Error)
SampleId	Sample identifier	Unique identifier for each sample within the lab system, allowing to link isolates derived from the same sample.	False
SampleOrigin	Sample source	Sample source: human, food, feed, animal, environment, ...	False
SequenceType	Sequence type	MLST 7 gene sequence type	False
Serotype	SerotypeSALM	Serotype, Group or Subspecies of Salmonella which is the cause of the reported disease.	False
Specimen	SpecimenSALM	The relevant specimen type used for diagnosis of the case	False
Status	Status	Status of reporting NEW/UPDATE or DELETE (inactivate). Default if left out: NEW/UPDATE. If set to DELETE, the record with the given recordId will be deleted from the TESSy database (or better stated, invalidated. If set to NEW/UPDATE or left empty, the record is newly entered into the database.	False
Subject	Subject	Isolate.	True (Error)
VNTRProtocol	VNTR Protocol used	VNTR Protocol used.	False
WgsAssembler	Wgs assembler	The assembler used for sequencing, optionally including parameter settings.	False

Variable	Full name	Description	Required
WgsAssembly	Wgs assembled genome	The assembled genome, as a zipped FASTA file. The file contents are subsequently converted into a Base64-encoded string for inclusion into either the XML or CSV data for the isolate.	False
WgsEnaId	Wgs ENA identifier	European Nucleotide Archive (ENA) run identifier, based on which the sequence read data can be retrieved. Starts with ERR or SRR, i.e. not the sample or experiment which ERS/ERX or SRS/SRX.	False
WgsProtocol	Wgs protocol	Protocol used for sequencing, limited to the sequencing technology used (today Illumina or IonTorrent) and the read length.	False
WgsRawReads	Wgs raw reads	Wgs raw read files (to be used for cloud services)	False
WgsSequenceId	Wgs SRA identifier	Sequence Read Archive (SRA) run identifier, based on which the sequence read data can be retrieved. Starts with ERR or SRR, i.e. not the sample or experiment which ERS/ERX or SRS/SRX.	False

2.2.4. Global Initiative on Sharing All Influenza Data (GISAID)

The Global Initiative on Sharing All Influenza Data (GISAID) aims to promote research and rapid sharing of data about pathogenic influenza and coronavirus viruses, and is recognized by the European Commission as a research organization. To serve this purpose, GISAID launched a database that collects molecular data on these viruses. The database is open and facilitates the submission of molecular data, so that these data are rapidly available to the scientific community.

Submission of molecular sequences to the GISAID database includes the data listed in Table 5.

Table 5. Data submitted to GISAID database (https://www.protocols.io/view/sars-cov2-gisaid-submission-protocol-kqdg35oy1v25/v3).

GISAID metadata	Requested information	GISAID requirement status
Submitter	GISAID-username	mandatory
FASTA filename	Filename of sequence fastafile	mandatory
Virus name	The sequence name on fasta file	mandatory
Type	betacoronavirus	mandatory
Passage details/history	e.g. Original, Vero	mandatory
Collection date	Date in the format YYYY or YYYY-MM or YYYY-MM-DD	mandatory
Location	e.g. Europe /Germany/ Bavaria/ Munich	mandatory
Additional location information	e.g. Cruize ship, Convention, Live animal	
Host	e.g. Human, Canine, Environment	mandatory
Additional host information	e.g. Patient infected during travel	
Sampling strategy	e.g.Sentinel surveillance (ILI)	
Gender	Male, female or unknown	mandatory
Patient age	Same.g 65 or 7 months or unknown	mandatory
Patient status	e.g. Hospitalised, released, live, diseased, or unknown	mandatory
Specimen source	e.g. Sputum, Oropharyngeal swab, Blood, Faeces, Other	
Outbreak	Date, location (e.g. type of gathering, family cluster etc)	
Last vaccinated	Provide details if applicable	
Treatment	Include drug name, dosage	
Sequencing technology	e.g. Illumina MiSeq, IonTorrent, Sanger etc	mandatory

GISAID metadata	Requested information	GISAID requirement status
Assembly method	e.g Geneious 10.2.4, SPAdes, MEGAHIT etc.	
Coverage	e.g 70x, 1000x etc.	
Originating LAb	Where the clinical specimen or virus isolate was first obtained	mandatory
Address	Originating lab address	mandatory
Sample Id given by the originating lab	Original sample Id	
Submitting lab	Where sequence data were originated and submitted	mandatory
Address	Submitting lab address	mandatory
Sample Id given by the submitting lab	Submitting lab sample Id	
Authors	Comma separated list of Authors (Complete First followed by Last name)	

3. Survey

A survey was used in order to map the current molecular data collection and sharing capabilities of the consortium partners. The questionnaire was disseminated electronically (via Google Forms) and all partners answered.

3.1. Results

The results of the survey were the following:

- Three out of four participants stated that they collect WGS molecular data for national use on SARS-CoV-2. One of the countries is also collecting nationwide WGS data on Enterobacterales (including CRE), *B. pertussis*, VRE, MRSA & *C. difficile*.
- All countries use MS Excel as the main tool for data collection. Ridom SeqSphere+ is also used by one of them.
- The main data fields of the existing databases are the following:
  sample information (id/barcode, source, date, targeted or random sampling), patient data (id, sex, age, region), test information (date, results (pre-screening, WGS))
- All countries contribute WGS data to the European Centre for Disease Prevention and Control (ECDC). In addition to data from SARS-CoV-2 isolates, the partners reported that they are also uploading WGS data for *Legionella pneumophila*, *Listeria monocytogenes*, CRE clinical isolates, Mpox and Influenza.
- Programmatic upload does not appear to be used, as countries report using file upload (2 out of 3) or manual data entry (1 out of three).

3.2. Conclusions

The results of the survey indicate that currently no database is used by any of the participating national organizations. All participants reported using excel files for data storage, except for Austria which additionally uses the Ridom SeqSphere+ software.

All participants upload data to ECDC. Greece and Croatia also upload data to GISAID. Apart from SARS-CoV-2 WGS data, Austria uploads data on *Legionella pneumophila* and *Listeria monocytogenes*, and Hungary uploads data on CRE clinical isolates, Mpox and Influenza.

4. Discussion – Future perspectives

As pointed out in the survey, none of the partners currently uses a database, which gives the opportunity to construct a database adapted to the needs of molecular epidemiology, and at the same time compatible with existing international databases, that will promote the rapid availability of molecular data to local and international Public Health agencies. Molecular databases are useful for detecting:

1. Identical and/or similar genomes among pathogens, which is an indication of an outbreak
2. The so called "high risk clones", i.e. clones known to be epidemiologically fit, highly virulent, and/or resistant to antimicrobials
3. New pathogen variants that should be evaluated for epidemiological fitness and antimicrobial resistance.

In all the above cases, sequence comparisons trigger epidemiological investigations and/or the implementation of appropriate control measures to protect Public Health. Consequently, the minimum requirements of a molecular database are to store sequencing data (raw data, assemblies, and annotations), and metadata related to the sequencing process, time, and location. The database should also store strain characterization results, which could be cgMLST, MLST, genotype, AMR etc, depending on the pathogen and the public health interest. These data are also sufficient for data submission to ENA. However, in Europe, Public Health reference laboratories should provide data to the databases of the official European Union agencies (ECDC and EFSA) and the GISAID database, therefore a database suitable for storing molecular data should at least provide the mandatory metadata required by these databases (Table 6) to facilitate automated data submission.

Table 6. Mandatory metadata for submission of molecular data to GISAID, EFSA One Health WGS system and ECDC WGS Surveillance System.

Requested information	EFSA	ECDC	GISAID
Collection date/Date of sampling/Sampling year	+	+[*1]	+
Isolate species / Subject / Virus name	+	+	+
Record id / Sampling ID	+	+	
Additional information of the matrix of the sample	+		
Coded description of the matrix of the sample taken	+		
FASTQ file 1 MD5 checksum	+		
FASTQ file 1 name	+		
FASTQ file 2 MD5 checksum	+		
FASTQ file 2 name	+		
Isolate ID	+		
Layout	+		
Local Raw Reads ID	+		

Gender		+*1	+*1
Patient age		+*1	+*1
Data source / Submitting lab		+	+
Date of receipt in reference laboratory		+	
Record type (ex "case based reporting or aggregate reporting")		+	
FASTA filename			+
Host			+
Passage details/history			+
Patient status			+*1
Sequencing technology			+
Type of record (ex "betacoronavirus)			+

*1 Fields are recommended by the surveillance systems but could be omitted as systems allow missing or "unknown" values as valid input.

This minimal database configuration is advantageous because it is simple and avoids GDPR compliance issues. In this case, the database should provide the ability to link molecular data of the samples with detailed epidemiological information available in national epidemiological databases or archives to facilitate the epidemiological investigation.

Each partner however, depending on available infrastructure, administrative structure, and local legislation, should decide whether a minimal database configuration is suitable, or a more complex database should be sought.