COLLABORATION AGREEMENT

on cross-sectoral sharing and use of Whole Genome Sequencing and related data of foodborne pathogens for public health purposes between ECDC and EFSA

Annex II.
Technical specification of the ECDC-EFSA One Health WGS System

Content

The ECDC – EFSA One Health WGS system ..... **Error! Bookmark not defined.**

EFSA and ECDC WGS and metadata exchange 3

General interaction ..... **Error! Bookmark not defined.**

Sharing of metadata ..... **Error! Bookmark not defined.**

Core genome MLST schemas and allele calling procedure 5

System logging 12

Application programming interface call content 5

Query process from ECDC WGS system to EFSA WGS system ..... **Error! Bookmark not defined.**

Query 6

Response ..... **Error! Bookmark not defined.**

Query process from EFSA WGS system to ECDC WGS system 6

Query ..... **Error! Bookmark not defined.**

Response ..... **Error! Bookmark not defined.**

Audit trail 10

Performance indicators 12

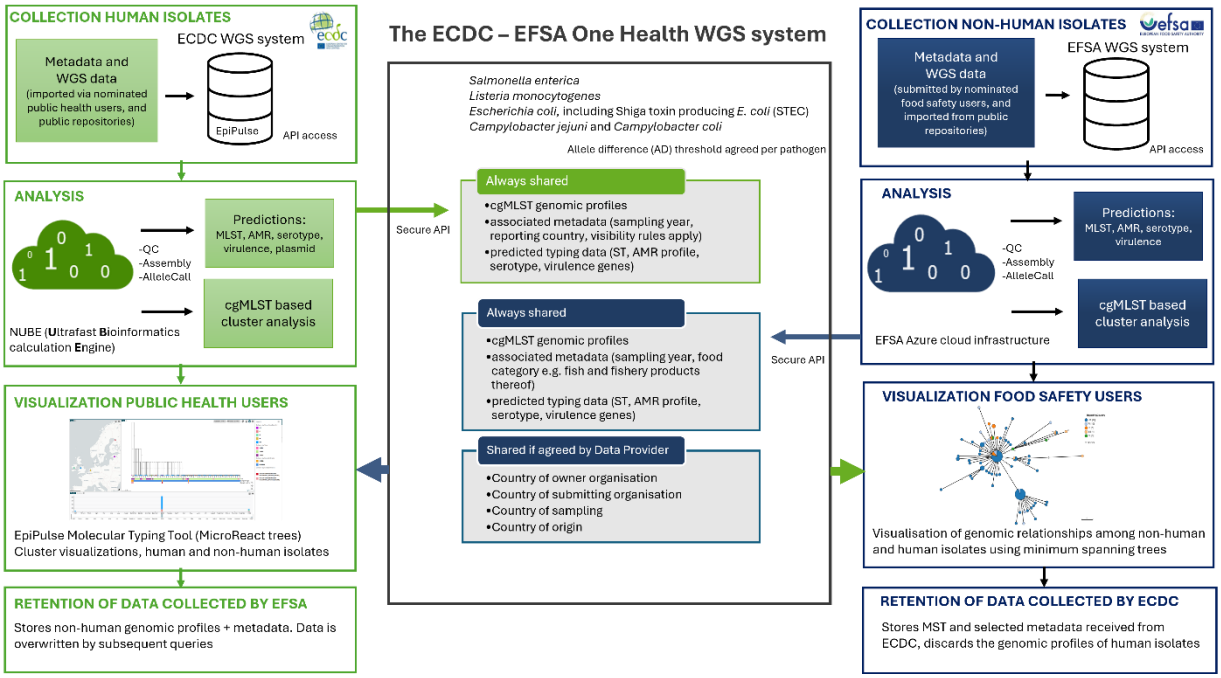
Exceptional data sharing protocol during API downtime ..... **Error! Bookmark not defined.**

Abbreviations 13

The ECDC – EFSA One Health WGS system

This system consists of two interconnected IT platforms, one hosted by ECDC (i.e. EpiPulse Molecular Typing Tool and EpiPulse Cases, hereinafter "ECDC WGS System") and the other by EFSA (hereinafter "EFSA WGS System"), each dedicated to collect and analyse Whole Genome Sequencing (WGS) data of *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli*, including Shiga toxin producing *E. coli* (STEC), *Campylobacter jejuni* and *Campylobacter coli* obtained from humans (in ECDC) and food, feed, animals and related environment (in EFSA).

Figure 1. Schematic description of ECDC-EFSA One Health WGS system.



Left panel indicates the workflow of ECDC WGS System, the pipeline is publicly available at this link. The right panel indicates the workflow of EFSA WGS system and is publicly available at this link. The centre panel shows the interaction between the systems and the data shared between the sectors

WGS and metadata exchange

WGS data are shared between ECDC and EFSA through a query–response mechanism. This does not include raw sequencing reads (e.g. FASTQ files) and is limited to processed, sequence-derived data, such as cgMLST allelic profiles and sequence-derived typing information, together with associated metadata.

Query–response mechanism

Either ECDC or EFSA may initiate a query by submitting reference cgMLST genomic profiles and their associated metadata. The responding system then performs a pairwise comparison against all cgMLST profiles in its database and returns all profiles that fall within the genetic distance threshold set by the querying party for at least one reference profile. The response includes the matching genomic profiles and their associated metadata.

This mechanism ensures that data are exchanged based on genetic proximity to the reference set and applies in the same way regardless of which party initiates the query.

Data exchange follows two main workflows:

- **ECDC routine cluster analysis and match detection**: during routine operations, ECDC initiates queries as part of its cluster analysis workflow.
- ***Ad hoc* analysis and match detection**
    - **ECDC-initiated workflow**: ECDC may submit queries to EFSA outside routine activities at any time, usually in relation to specific EpiPulse events.
    - **EFSA-initiated workflow**: EFSA or users of the EFSA WGS system at Member State level may initiate queries to ECDC at any time.

ECDC routine cluster analysis and match detection

When at least two genomic profiles from isolates of human origin form a cluster, the workflow proceeds as follows:

1. **Query submission (ECDC → EFSA)**
   ECDC submits a query containing:
    - cgMLST genomic profiles from the isolates of human origin forming the cluster; and
    - the associated metadata (i.e. sampling year, reporting country)
    - sequence-derived typing data (i.e. ST, AMR profile, serotype, virulence genes)
    - predefined species-specific genetic distance thresholds

   ECDC may submit either the complete cluster or a defined subset of profiles from isolates of human origin from the cluster for further analysis.
2. **Pairwise comparison and match identification (EFSA)**
   EFSA compares the received genomic profiles from isolates of human origin pairwise with genomic profiles from isolates of food/non-human origin stored in the EFSA database, using the thresholds specified in the query payload. Matching genomic profiles are then identified.
3. **Return of matching results (EFSA → ECDC)**
   EFSA returns the following information to ECDC:
    - the cgMLST profiles of the matching isolates of food/non-human origin; and
    - their associated metadata (sampling year, food category e.g. fish and fishery products thereof)
    - predicted typing data (ST, AMR profile, serotype, virulence genes)
    - during the production of a joint Rapid Outbreak Assessment, EFSA transfers to ECDC the entire set of collected epidemiological data of the matching profiles.
4. **Data handling and retention at ECDC**
    - genomic profiles received from EFSA are stored and integrated with the dataset of isolates of human origin for calculation of distances and cluster analysis;
    - EFSA isolates are assigned to clusters using the species-specific cluster definitions established by ECDC, in line with the approach used for isolates of human origin; and
    - these datasets are periodically refreshed through subsequent routine queries, with existing records being updated or overwritten as necessary to ensure data to be consistent with what is stored in the EFSA WGS System.
    - Users of the ECDC WGS system who have access to a cluster can view all metadata associated with that cluster.
5. **Data handling and retention at EFSA**
    - EFSA uses the genomic profiles and metadata received from ECDC, together with the matching profiles from isolates of food/non-human origin from its own database, to construct and store a minimum spanning tree (MST) showing the genetic relationships among the isolates.

- After processing the original query payload received from ECDC, EFSA discards the payload, including the genomic profiles from isolates of human origin. EFSA does not maintain a repository of genomic profiles from isolates of human origin originating from ECDC.
- When reviewing query results, country-level metadata is visible only to users at Member State level from the corresponding country. Users cannot view country-level metadata associated with other Member States.

Ad hoc analysis and match detection

ECDC-initiated workflow

1. **Query submission (ECDC → EFSA)**
   ECDC submits a query containing:
   - one or more cgMLST profiles of isolates of human origin, not necessarily belonging to the same cluster but usually linked to the same event; and
   - the associated metadata.
   - sequence-derived typing data
   - genetic distance thresholds range from zero to a pre-agreed species-specific maximum value
2. **Pairwise comparison and match identification (EFSA)**
   Matches are identified by EFSA in the same way as during routine ECDC cluster analysis, see above.
3. **Return of matching results (EFSA → ECDC)**

   EFSA returns to ECDC the same information as during ECDC routine cluster analysis, see above.
4. **Data handling and retention at ECDC**

   The profiles of isolates of food/non-human origin received from EFSA are handled in the same way as during ECDC routine cluster analysis, see above, with the following exception when a higher cut-off distance than the predefined routine cut-off is being used:
   - Those matching profiles that are neither assigned to a cluster nor associated with a joint Rapid Outbreak Assessment (ROA) are automatically deleted after three months.
5. **Data handling and retention at EFSA**

EFSA handles and retains the data in the same way as during routine ECDC cluster analysis, see above.

EFSA-initiated workflow

1. **Query submission (EFSA → ECDC)**
   EFSA submits a query containing:
   - one or more cgMLST profiles of isolates of food/non-human origin.
   - genetic distance thresholds range from zero to a pre-agreed species-specific maximum value

2. **Pairwise comparison and match identification (ECDC)**
   ECDC performs a pairwise comparison between the received profiles of isolates of food/non-human origin and the profiles of isolates of human origin stored in the ECDC database, using the genetic distance threshold requested in the query. Matching profiles are then identified.

3. **Return of matching results (ECDC → EFSA)**
   ECDC returns to EFSA:

   - the cgMLST profiles of matching isolates of human origin; and
   - their associated metadata.
   - sequence-derived typing data

4. **Data handling and retention at ECDC**

   ECDC does not retain the query payload, including the profiles of food and other non-human origin isolates submitted as part of the query.

5. **Data handling and retention at EFSA**

   EFSA handles and retains the ECDC data in the same way as during routine ECDC cluster analysis, see above.

Query–response content

Each Application Programming Interface (API) call, including both queries and responses, uses a JSON file containing the fields described below. For queries sent from the EFSA WGS system to the ECDC WGS system, the ECDC API uses pathogen-specific end points. For queries sent from the ECDC WGS system to the EFSA WGS system, the EFSA API uses a single endpoint, with the target species specified as a query parameter.

Query from ECDC WGS system to EFSA WGS system

Query Payload Structure

The request payload contains a query header, which defines the scope and configuration of the query, and two arrays: “reference”, which contains the cgMLST profiles used as query references, and “metadata”, which provides contextual information for each profile included in the “reference” array.

The query header includes the following elements.

Element	Type	Example	Description
query_id	String	1	Unique identifier for the query.
cluster_id	String	2025-10.LIST.18.CC1	Identifier of the cluster/event analysed in the ECDC WGS system.
threshold	Integer	7	Genetic distance threshold applied in the analysis.
isolate_species_code	String	RF-00000251-MCG	Code identifying the species-specific cgMLST schema in the EFSA WGS system.

In the “reference” array, each object represents a cgMLST profile and includes the following elements:

Element	Type	Example	Description
id	String	ECDC_820	Pseudonymised identifier of the isolate of human origin from which the profile was generated.
Locus 1	Integer	3691085794	Allele designation for the first locus (locus name depending on the cgMLST schema used).
Locus N	Integer	2691085794	Allele designation for the N-th locus (locus name depending on the cgMLST schema used).

Each object in the “metadata” array must match an object in the “reference” array through the “id” field.
The array provides contextual information and includes the following elements:

Element	Type	Example	Description
id	String	ECDC_820	Pseudonymised identifier matching the corresponding profile in the reference array.
epidemiological_data	Object	{"sampling_year": "2025", "sampling_country": "Sweden", "sampling_matrix_type": "HUMAN"}	Metadata related to sampling and origin.
experimental_data	Object	{"instrument_model": "IONTORRENT", “submitting_organization_country": "Sweden"}	Information about the sequencing experiment.
microbiological_data	Object	{"sequence_type": "1", "serotype": "4b", "clonal_complex": "CC1", "cluster_code": "2025-10.LIST.18.CC1"}	Microbiological typing information.
extra_data	Object	{"event_code": "2025-FWD-0007"}	Additional metadata, including the event code where applicable.

Response Payload Structure

After receiving a query, the EFSA WGS system immediately returns the query ID and provides the endpoint where the response will be made available. The ECDC WGS system is responsible for retrieving the response from that endpoint.

The response payload includes the original request from ECDC, followed by the “output” object. This object contains a general response header, composed of the “query” and “response” objects, and the results in the “matching_profiles” array.

Element	Type	Example	Description
query	Object	{"query_id": 7235, "query_date": "2025-11-05T11:37:32.352Z"},	Metadata describing the received query ID and timestamp.
response	Object	{"response_id": 7235, "response_date": "2025-11-05T11:37:32.403Z", "response_match_found": true, "response_profile_number": 1}	Response metadata, including the response ID, timestamp and whether matching profiles were found. ***NOTE**: if no profiles match the reference profiles included in the original query, "response_match_found" is set to "false" and "response_profile_number" is set to 0.*

The “matching_profiles” array provides the cgMLST profile, additional typing information and contextual information for each profile in the EFSA database that matches the reference profile(s) included in the query. It includes the following elements:

Element	Type	Example	Description
entry_id	String	EFSA-S-2025-013886	Identifier of the matching profile in the EFSA WGS system.
entry_public	Boolean	FALSE	**TRUE** if the cgMLST profile was generated from a genome sequence not assigned to an EFSA data provider organisation, generally because it was imported from INSDC. **FALSE** if the cgMLST profile was generated from a sequence assigned to an EFSA data provider organisation.
epidemiological_data	Object	{"sampling_year": 2011, "sampling_country": "France","sampling_matrix_type": "Food","sampling_matrix_category": "Milk and milk products (dairy)", **"sample_country_of_origin":** null, **"sampling_point"**: null, **"sampler"**: null, **"programme_type"**: "Outbreak investigation",**"programme_info":** "2018-	Metadata related to sampling and origin. ***NOTE**: elements shared only during assessment are shown in bold.*

Element	Type	Example	Description
		FWD-00038","sampling_country_origin": null, **"sampling_area":** null, "sampling_date": "2011-06-19T00:00:00",**"sampling_matrix_code":** "Cheese",**"isolation_year":** 2011, **"isolation_month":** 6, **"isolation_day**": 19}	
experimental_data	Object	{"instrument_model": null, "submitting_organization_country": "Italy","owner_organization_country": "Italy","layout": 2},	Information about the sequencing experiment.
microbiological_data	Object	{"mlstst": 8, "predicted_serotype": null,"qc_cgmlst_pass": true, "qc_cgmlst_missingloci": 4, "gene_list": [{"amr_gene_name": "fosX", "amr_gene_identity": 98.76 "amr_gene_coverage": 100}],"seq_variations": [],"ecoli_virulence_genes": null, "ecoli_stxsubtypes": null}	Microbiological typing information.
ena_data	Object	{"ena_study_accession": null, "ena_sample_accession": "SRS18002094", "ena_experiment_accession": "SRX20709187", "ena_run_accession": "SRR24951062","center_name": "FDA", "collection_date": "2023-05-02", "geographical_origin": "Lithuania"	Applicable only if the cgMLST profile was generated from a genome sequence that was imported from INSDC.
allelic_profile	Object	{"locus1": "138852938", ..., "locusN": "782725232"}	Allele designations for the N loci, with locus names depending on the cgMLST schema used.

The epidemiological data included in the EFSA response to ECDC may vary depending on data availability in the EFSA WGS system and on the EFSA data provider's decision to share country information, as shown in the example below.

Data element	Example of payload based on different options			
Available Epidemiological data in the EFSA MTS	*Yes*	*Yes*	*No*	*No*
Decision of the EFSA Data Provider regarding sharing "Country information"	*Yes*	*No*	*Yes*	*No*
SubmittingOrganizationCountry	Germany	*null*	Germany	*null*
SamplingYear	2018	2018	*null*	*null*

SamplingCountry	Germany	*null*	*null*	*null*
SamplingMatrixType	Food	Food	*null*	*null*
SamplingMatrixCategory	Milk and milk products (dairy)	Milk and milk products (dairy)	*null*	*null*

Query from EFSA WGS system to ECDC WGS system

Query payload structure

The request payload consists of a query header, which defines the scope and configuration of the query, and one "reference" array containing the cgMLST profiles used as query references.

The query header includes the following elements:

Element	Type	Example	Description
query_id	String	1	Unique identifier for the query.
threshold	Integer	7	Genetic distance threshold applied in the analysis.

In the "reference" array, each object represents a cgMLST profile and includes the following elements:

Element	Type	Example	Description
id	String	EFSA-S-2026-000244	Identifier of the profile of an isolate of food/non-human origin in the EFSA database.
Locus 1	Integer	3691085794	Allele designation for the first locus (locus name depending on the cgMLST schema used).
Locus N	Integer	2691085794	Allele designation for the N-th locus (locus name depending on the cgMLST schema used).

Response payload structure

After receiving a query, the ECDC WGS system returns the result synchronously. The response payload includes the original request from EFSA, followed by the "output" object. This object contains a general response header, composed of the "query" and "response" objects, and the results in the "matching_profiles" array.

Element	Type	Example	Description
query	Object	{"query_id": 7235, "query_date": "2025-11-05T11:37:32.352Z"},	Metadata describing the received query ID and timestamp.

Element	Type	Example	Description
response	Object	{"total_matches": 15, "execution_time_ms": 250},	Response metadata, including the total number of matches and execution time in milliseconds.
success	Boolean	true	true when no errors are identified.
message	String	Match query completed successfully	Message returned by the system, particularly when an error occurs.

The “matching_profiles” array provides the cgMLST profile, additional typing information and contextual information for each profile in the ECDC database that matches the reference profile(s) included in the query. It includes the following elements:

Element	Type	Example	Description
profile_id	String	ECDC_9752	Identifier of the matching profile in the ECDC WGS system.
distance	Integer	5	Allelic distance to the closest reference.
epidemiological_data	Object	{"sampling_year": "2025", "sampling_country": "Sweden", "sampling_matrix_type": "HUMAN"}	Metadata related to sampling and origin.
experimental_data	Object	{"instrument_model": "IONTORRENT", "submitting_organization_country": "Sweden"}	Information about the sequencing experiment.
microbiological_data	Object	{"sequence_type": "1", serotype: "4b", "clonal_complex": "CC1", "cluster_code": "2025-10.LIST.18.CC1"}	Microbiological typing information.
extra_data	Object	{"event_code": "2025-FWD-0007"}	Additional metadata, including the event code where applicable.
alleles	Object	{"locus1": "138852938", ..., "locusN": "782725232"}	Allele designations for the N loci, with locus names depending on the cgMLST schema used.

System logging

Security controls are applied to all API-based exchanges between the EFSA and ECDC WGS systems. Access is restricted to the agreed system-to-system communication channels and is protected through authentication and network-level controls defined bilaterally by EFSA and ECDC.

When the ECDC WGS system submits queries to the EFSA WGS system and retrieves responses from dedicated EFSA endpoints, access is performed through the EFSA partner API using an authentication mechanism managed by EFSA. This mechanism is used to verify the identity of authorised systems, ensure secure API communication, and prevent unauthorised access to the EFSA endpoints.

When the EFSA WGS system submits queries to the ECDC WGS system and retrieves the corresponding responses, EFSA accesses the ECDC API directly. Access to the ECDC API is restricted to the whitelisted IP address of the EFSA WGS system. This network-level control ensures that only requests originating from the authorised EFSA system can reach the ECDC API.

Both organisations maintain logs of API activity to support traceability, troubleshooting and monitoring of the exchange process. Logged information is limited to the elements required for operational oversight, auditability and incident investigation, in line with the agreed data handling and retention arrangements.

Audit trail

For each query sent to or received from the ECDC WGS system, EFSA stores the following information in the EFSA WGS system database: query submission date, response time, Internet Protocol (IP) address of the ECDC WGS system, and returned status code.

For each query received from the EFSA WGS system, the ECDC logging system stores the hashed IP address, query date and time, and request and response bodies.

Performance indicators

EFSA and ECDC monitor system performance over time by recording response times from the queried system and noting any downtime or other deviations from normal operations.

Core genome MLST schemas and allele calling procedure

EFSA and ECDC perform cgMLST analysis using agreed species-specific schemas and allele-calling procedures implemented in their respective WGS systems. Allele calling is performed using chewBBACA version 3 or later (https://github.com/B-UMMI/chewBBACA), with parameters and quality criteria agreed between EFSA and ECDC. The schemas, analytical settings and profile formats used for data exchange are aligned to ensure that profiles can be compared consistently across systems.

The cgMLST schemas used for the exchange are defined as follows:

- For *Salmonella enterica* and *Escherichia coli*, including STEC, the cgMLST schemas are those defined by the INNUENDO consortium (Llarena et al., 2018). The schemas are available in chewBBACA online (https://chewbbaca.online/), and allele calling follows the instructions provided on the relevant schema homepage.
- For *Listeria monocytogenes*, the cgMLST schema is the one defined by Institut Pasteur (Moura et al., 2016), adapted for chewie-NS (Mamede et al., 2021). The schema is available in chewBBACA online (https://chewbbaca.online/), and allele calling follows the instructions provided on the relevant schema homepage.
- For *Campylobacter jejuni* and *Campylobacter coli*, EFSA and ECDC apply the Oxford cgMLST V2 schema, downloaded from the PubMLST website (Jolley et al., 2018) and adapted for use with chewBBACA 3.3.10; this schema is not implemented on chewBBACA online.

Profiles exchanged between the two systems consist of allele designations expressed as CRC32 hashes generated from the corresponding allele nucleotide sequences. Any transformation required to harmonise profiles across schema versions or naming conventions is applied before comparison, according to bilaterally agreed specifications. EFSA and ECDC exchange and review the detailed technical specifications, including schema versions, locus mappings, translation rules, allele-calling settings and genetic distance thresholds, on a quarterly basis.

Exceptional data-sharing protocol during API downtime

If the API service is unavailable for more than five consecutive working days, EFSA and ECDC will activate an alternative procedure for sharing WGS data and metadata in exceptional circumstances. This procedure applies only when both organisations need to publish results or respond to requests from risk managers, including in the following situations:

- joint assessments;
- urgent requests; and
- food crisis situations.

The alternative procedure becomes active from the sixth working day of downtime and remains in place until the API service is restored. During this period, the exchange is handled manually. cgMLST profiles and associated metadata are transferred to an *ad hoc* folder in the ECDC SharePoint area jointly accessible by ECDC and EFSA.

Under this procedure:

- ECDC shares the payload for the relevant cluster; and
- EFSA performs the analysis manually and provides the individual matching profiles.

Abbreviations

API	Application Programming Interface
cgMLST	core genome Multilocus Sequence Typing
CRC32	A 32-bit hashing method that generates a checksum from input data (Cyclic Redundancy Check)
EpiPulse	European surveillance portal for infectious diseases
ECDC	European Centre for Disease Prevention and Control
EFSA	European Food Safety Authority
INSDC	International Nucleotide Sequence Database Collaboration (ENA, NCBI and DDBJ)
IP	IP address (Internet Protocol)
MLST	Multilocus Sequence Typing
MST	Minimum Spanning Tree
MTS	Molecular Typing System
MTT	EpiPulse **M**olecular **T**yping **T**ool
NUBE	ECDC Bioinformatics system (**N**UBE **U**ltrafast **B**ioinformatics calculation **E**ngine)
PubMLST	Public databases for molecular typing and microbial genome diversity (hosted at Oxford university)
STEC	Shiga toxin-producing *Escherichia coli*

WGS	Whole Genome Sequencing