



EC mandate on 'One Health' system for the collection and analysis of WGS data from food/animal isolates

EURL Salmonella meeting

28/05/2021

Trusted science for safe food

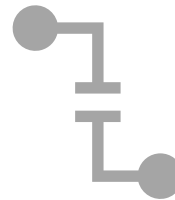
- Received December 2019 – Joint ECDC and EFSA
- **Term of references**
 - **ToR 1:** set up in ECDC and EFSA **two interoperable systems for the collection and sharing of WGS data** provided by Member States, allowing the joint analysis of WGS data for at least *Salmonella*, *L. monocytogenes*, and *E. coli* for the purpose of **multi-country outbreak detection and assessment**;
 - **ToR 2:** deliver services allowing data providers to interact and query the **systems, according to the agreed provisions on the management of data** on molecular testing of food, feed and animal isolates of selected foodborne pathogens and their use together with molecular typing data on isolates from human infections for public health purposes
- Deadline June 2022



Two interoperating systems

(EFSA and ECDC)

Each system collects and stores the data (i.e. allelic profiles and descriptive data) of the respective data domain.



Cross-sector matches

Databases will be queried, and comparison will be performed live to the data stored, returning any matches (according to business rules)

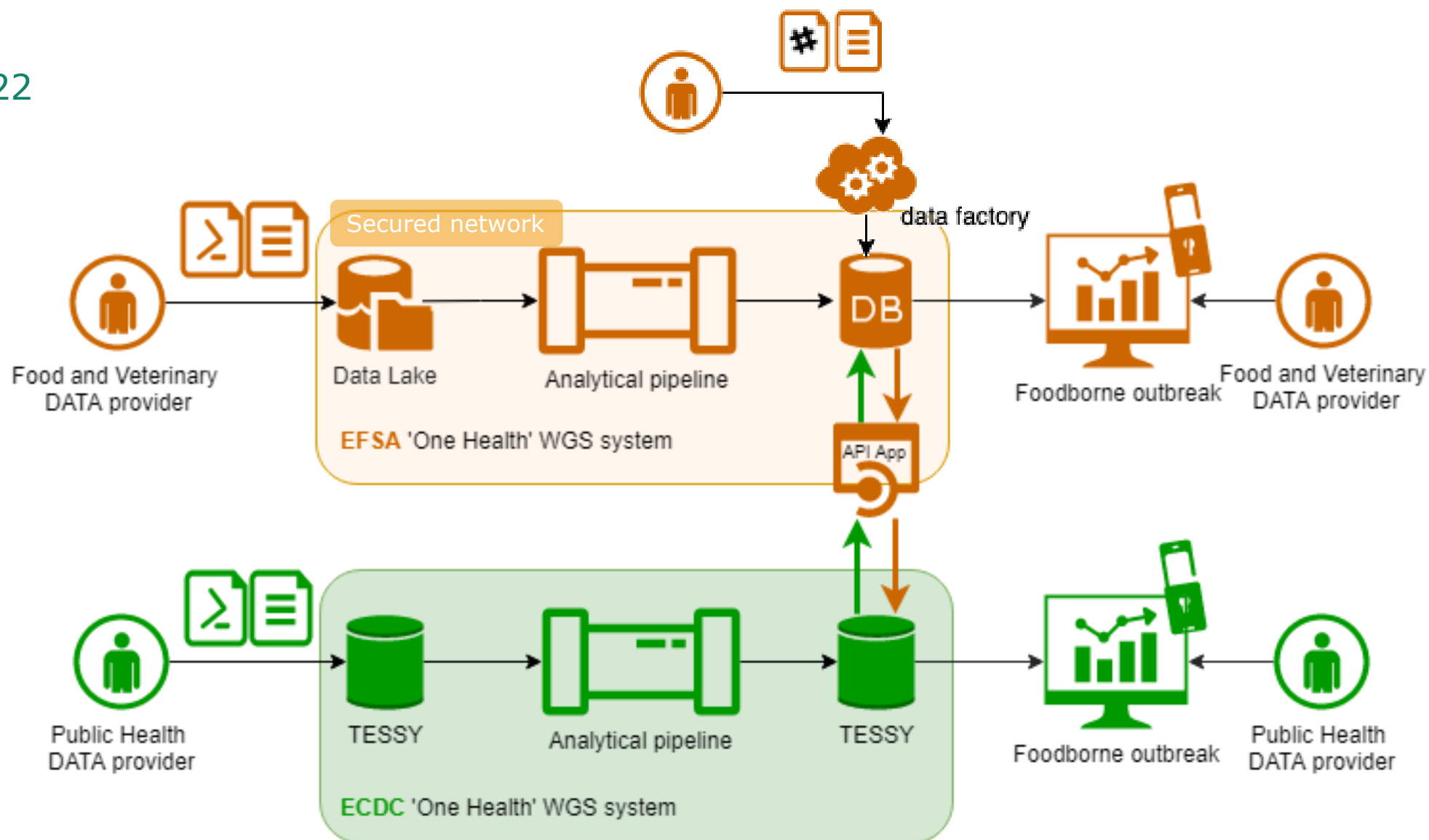


Machine-to-machine

Automatic exchanging of allelic profiles and descriptive data as established in the Collaboration Agreement

The foreseen 'One Health' system

June 2022



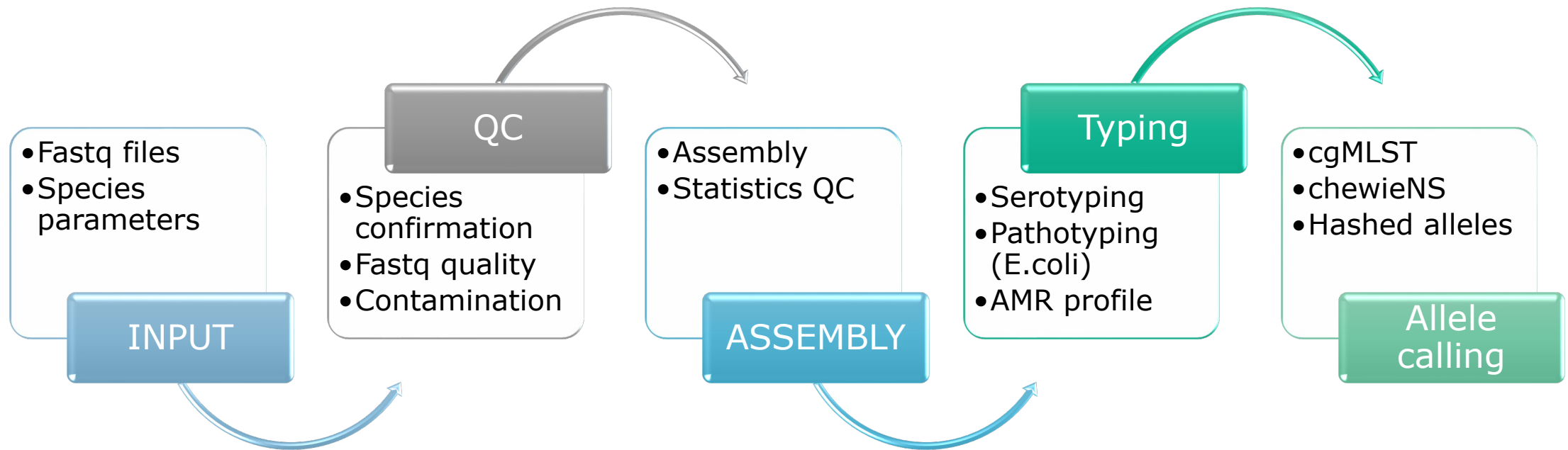
- The WGS Portal will be **integrated in the EFSA cloud Azure system**
- **Azure Data lake** located in Western (primary site) and Northern Europe (secondary site)
- the system will not allow re-submission of the same raw data (check for MD5 checksum)
- The **analytical pipeline** in Azure (i.e. from FASTQ to typing) will be **open source** and distributed in GitHub
- Data will be stored and analysed in a system within **secured network**

- Data provider **is always owner of the submitted and transformed** data
- The **data providers are allowed to withdraw data** for future use at any time
 - **Physical deletion** of *FASTQ* and *assembly* → remove the data from the data-store
 - **Logical deletion** of the *allelic profile* and *typing data* → data remains in the data-store, but invisible for future research
- Access policy still under discussion

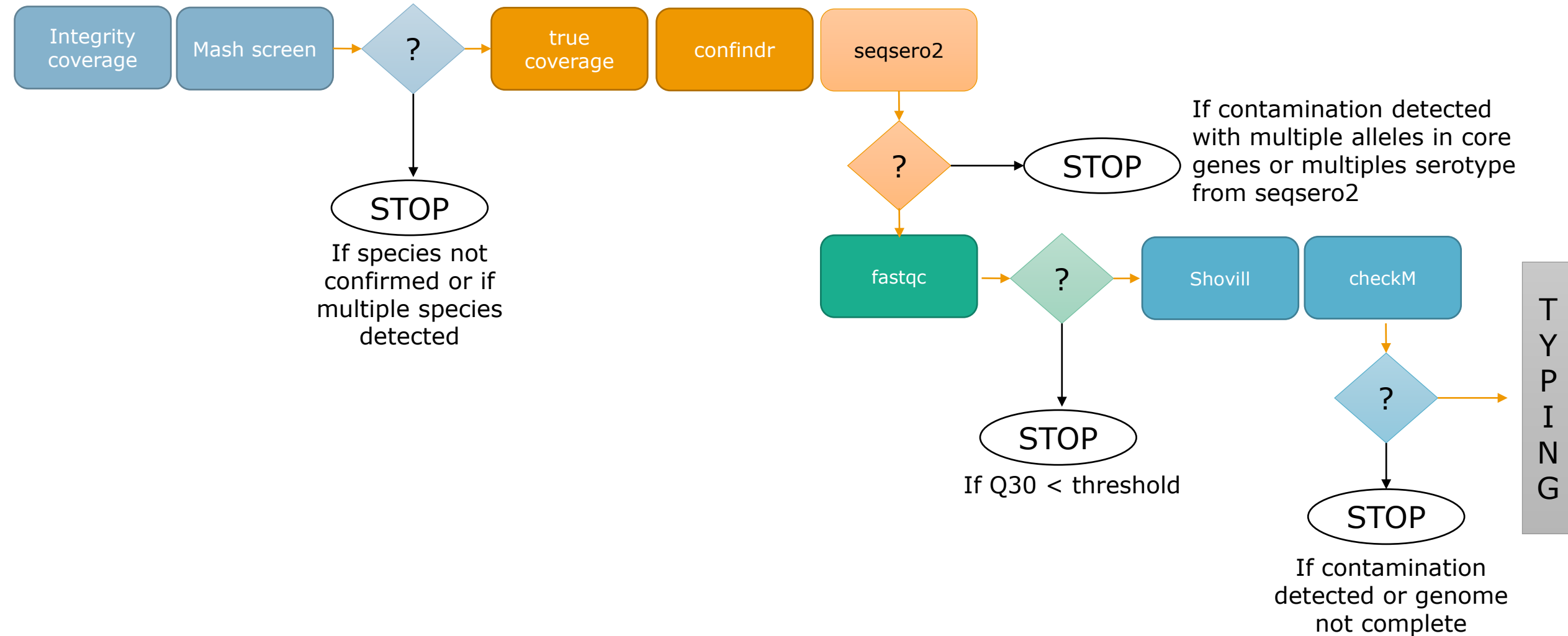
- **Raw Reads:** FASTQ(s) submitted by data-providers or downloaded from public repositories
- **EpiData:** epidemiological data linked to one Raw Reads submission; always submitted by the data-providers
 - **Contextual data:** metadata available in the public repositories and linked to publicly available raw reads
- **Analytical Pipeline Results:** containing data extracted from the raw reads by the analytical pipeline in the EFSA system or at the data-provider premises

- *Nextflow* can be used on any POSIX compatible system (Linux, OS X, etc)
- It requires Bash 3.2 (or later) and Java 8 (or later, up to 11) to be installed
- Implementation with Azure Batch possible

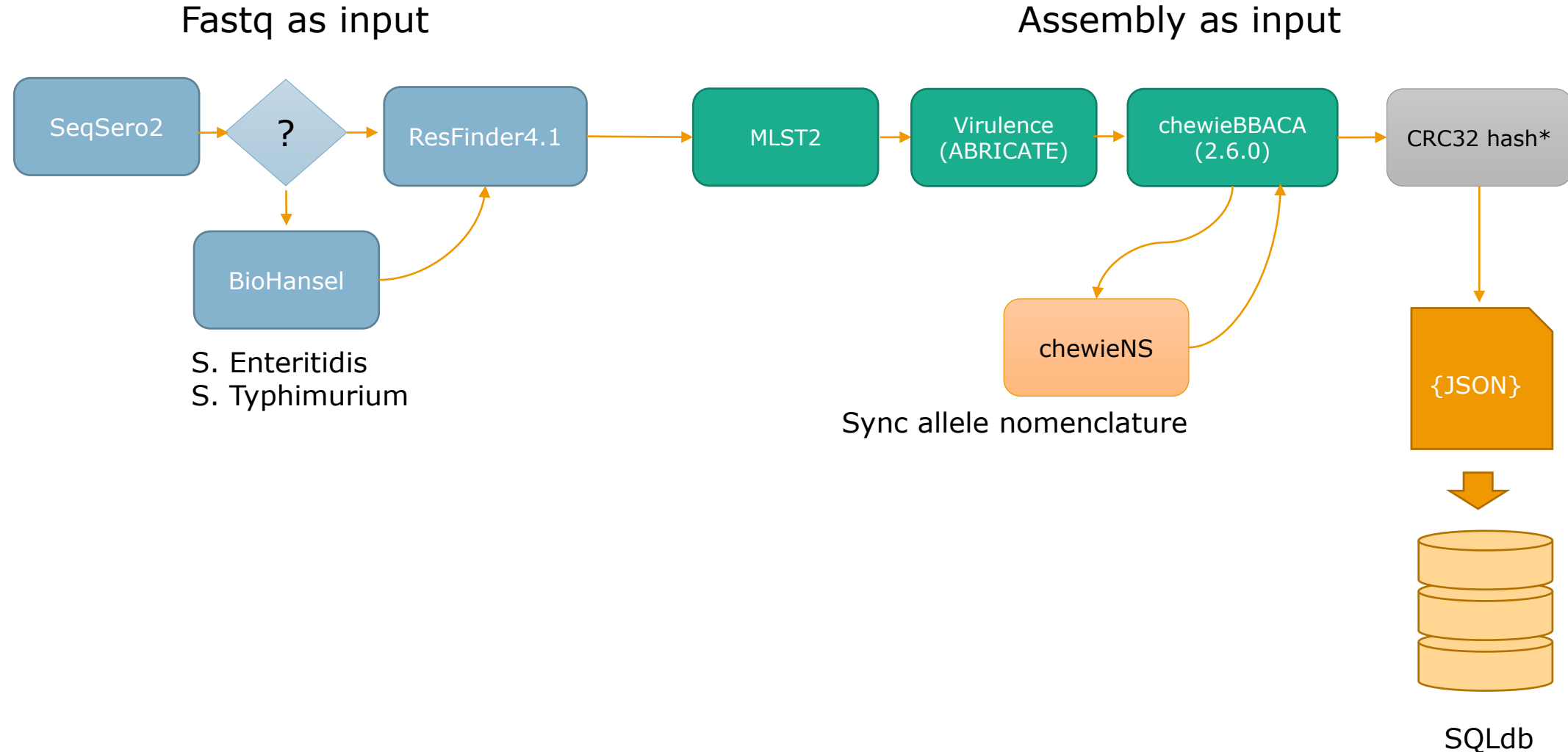
Nextflow pipeline



Details of QC block (*Salmonella*)



Details of typing and Allele Calling (*Salmonella*)



*NOTE: CRC32 hash function was chosen to be interoperable with BfR (<https://pubmed.ncbi.nlm.nih.gov/33643254/>; <https://www.frontiersin.org/articles/10.3389/fmicb.2021.649517/full>)

[illegible]

- cgMLST schema of **3,255 loci**
- Tested on outbreak of several outbreak dataset from different serotypes including Enteritidis, Typhimurium, Dublin, Havana, ... (further testing is ongoing)
- Clearly separate outbreak strain from unrelated isolates; cut-off for cluster definition might be slightly larger (depending on serotype)
- NOTE: Doesn't work for non-enterica subspecies
- Already available cluster nomenclature (<https://efsa.onlinelibrary.wiley.com/doi/abs/10.2903/sp.efsa.2018.EN-1498>)



July 30, 2018

Dataset Open Access

INNUENDO whole genome and core genome MLST schemas and datasets for Salmonella enterica

 Mirko Rossi;  Mickael Santos Da Silva;  Bruno Filipe Ribeiro-Gonçalves;  Diogo Nuno Silva;  Miguel Paulo Machado;  Mónica Oleastro; Vítor Borges; Joana Isidro; Luis Viera; Jani Halkilahti; Anniina Jaakkonen;  Federica Palma; Saara Salmenlinna; Marjaana Hakkinen;  Javier Garaizar;  Joseba Bikandi; Friederike Hilbert;  João André Carriço

Publication date:

July 30, 2018

DOI:

DOI [10.5281/zenodo.1323684](https://doi.org/10.5281/zenodo.1323684)

Communities:

A cross-sectorial platform for the integration of genomics in surveillance of food-borne pathogens

License (for files):

 Creative Commons Attribution 4.0 International

- Mirroring ENA submission model for fastq
- Raw Reads submitted to Data Lake through *SFTP or VPN or WEB-IN*
- **Raw reads linked to “Local Raw Reads ID” unique for OwnerOrganization**
- LocalRawReadsID for public data == ENA sample ID
- EpiData submitted linked to raw reads using LocalRawReadsID
- EpiData can be submitted asynchronously



ExpData to submit with FASTQ

LocalRawReadsID	data provider's sequence identification code or ENA Accession Number	provided
OwnerOrganization	the organization who owns the RawReads data; it should match to the one derived from the organization according userID who has logged on to provide the data	provided
InstrumentModel	the model of instrument used for the analysis (controlled vocabulary https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html#platform)	provided
Species	species describing the isolate species: one out of following three possible values are allowed: <ul style="list-style-type: none">• "Salmonella enterica"• "Listeria monocytogenes"• "Escherichia coli"	provided
LibraryLayout	SINGLE or PAIRED	provided
RawReads	Names of the FASTQ-files	provided

LocalRawReadsID	data provider's sequence identification code or ENA Accession Number	M/Exp
OwnerOrganization	The organization who owns the RawReads data; Linker to FASTQ submission	M/Exp
Sample ID	ID of the sample from which the bacteria has been isolated and sequenced and from which the WGS files have been derived.	M
Country of sampling	Country of sampling	M
Date of sampling	Year/month/day	M
Sample Matrix	Description of the sample taken based on FoodEx2 catalogue	M
Isolate ID	ID of the bacteria isolate sequenced and from which the WGS files have been derived.	M
Sampling point	Point where the sample which generates the isolate has been sampled	R
Country of origin	Country of origin of the sample taken	R
Date of isolation	Year/month/day of isolation	O
Area of sampling	Area where the sample was collected	O
Programme type	"Outbreak investigation" (K032A) or "RASFF alert notification" (K033A)	O
Sampler	i.e. "Official sampling" (CX02A)	O
Additional sampling programme information	RASFF notification number, if available.	O

- Users will access “Organization” page
- User of one organization for any species has the right to:
 - Submit/maintain entries
 - Execute/monitor/visualize analytical pipeline
 - Search EFSA database for similar matches
 - Visualize in ViZ the results from the matches
 - Download report and full results (only if published in DB)
- User will be able to search ECDC database based on profile similarity (not direct access of the system)
- User cannot download any data from EFSA system (only the data owned by the user)

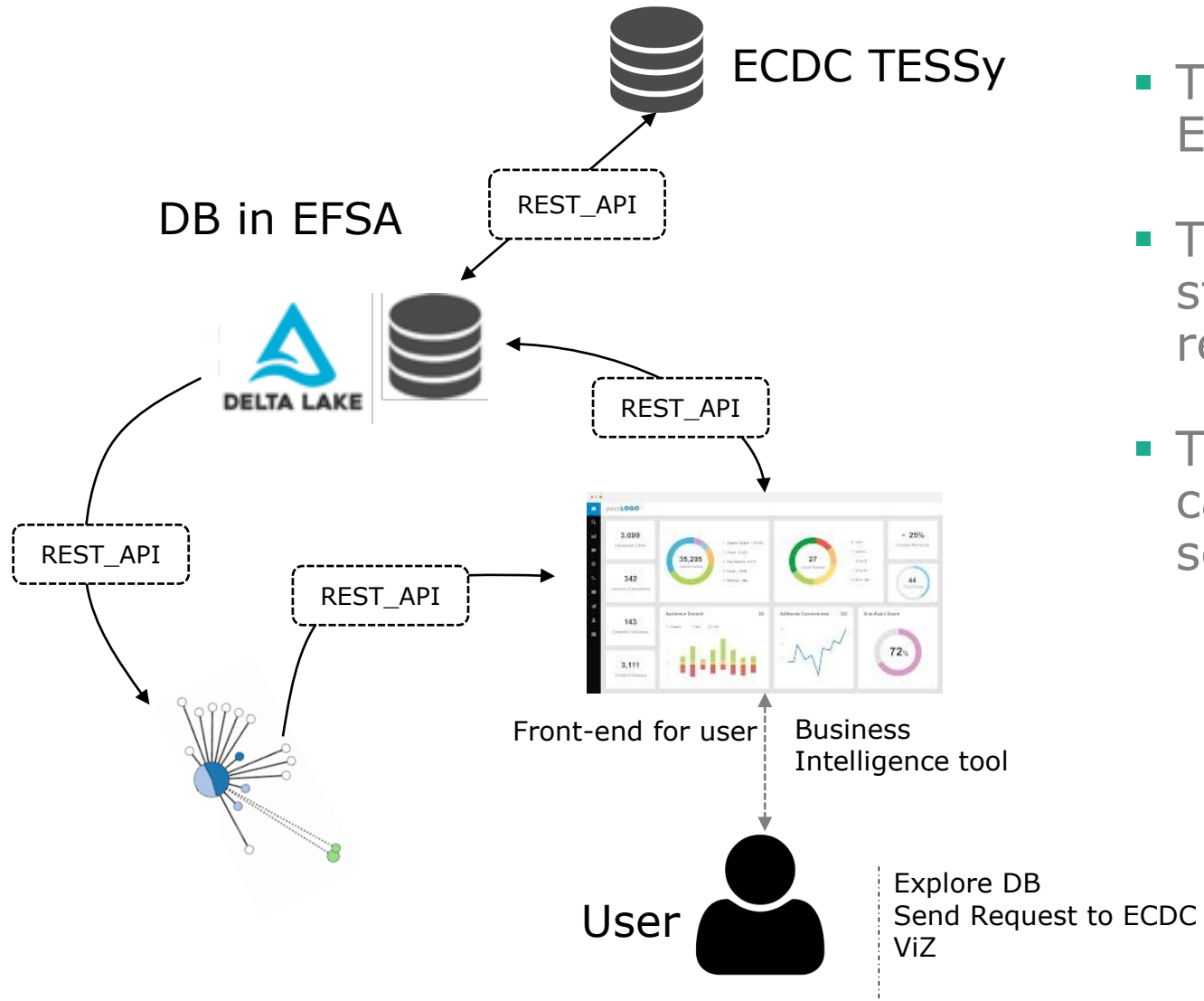
The WGS Portal

<div><div>▼</div><div>+ Upload ▼</div><div>+ Edit ▼</div><div>🗑 Delete ▼</div><div>⌫ Stop</div><div>▶ Execute</div><div>↺ Start release</div><div>🌳 GrapeTree</div><div>🔍 Query ECDC</div><div>⬇ Download</div></div> <div>Last grid refresh: 27/05/2021 - 15:49</div> <div>No new rows added. </div>								
Experimental data >					Last actions >			
Entry ID	Local raw reads ID	Status	Last modified on	Species	EpiData upload status	Pipeline run status	Pipeline version	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text" value="dd/mm/yy"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	
 <input type="checkbox"/> EFSA-S-2021-000001	2021-Giani1	Succeeded	27/05/2021 - 13:48	Escherichia coli		Requested	AP_Version_1.0	

Demo

- An UX testing was performed in October 2020 for validating the design
- UX testing involved several users (NRLs, EURLs and EFSA user)
- Several iterations was needed, still work is ongoing

Interaction with system



- The user can explore the content of EFSA DB based on specific requests
- The user can search for similar strain in EFSA DB and visualize relationship based on GrapeTree
- Through the EFSA system the user can submit request to ECDC DB for searching similar human isolates

Some examples of on going designs

Comparison Epidata

5
In number countries present

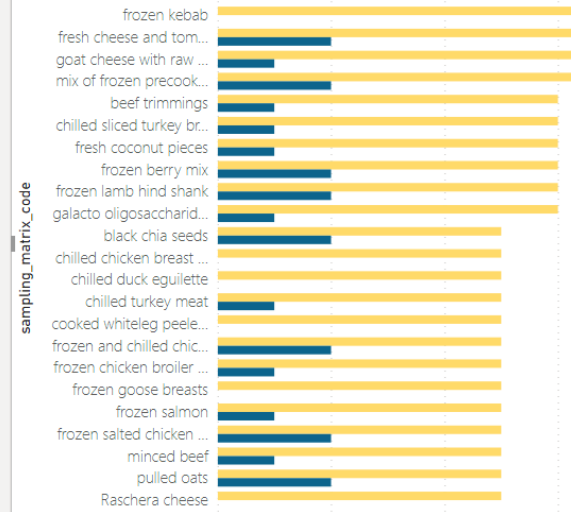
L. monocytogenes

Austrian Food Authority



Comparison

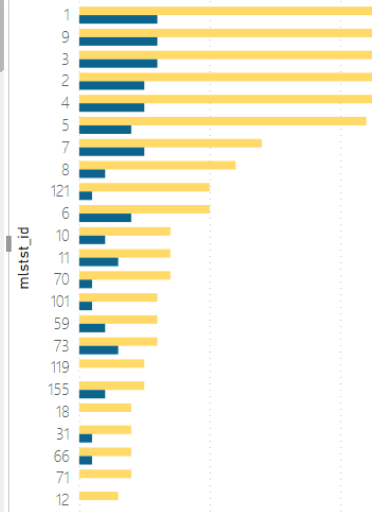
● Data All ● Data Owner Organisation



Data All and Data Owner Organisation

Comparison

● Data All ● Data Owner Organisation



Data All and Data Owner Organisation

Comparison Epidata

4
In number countries present

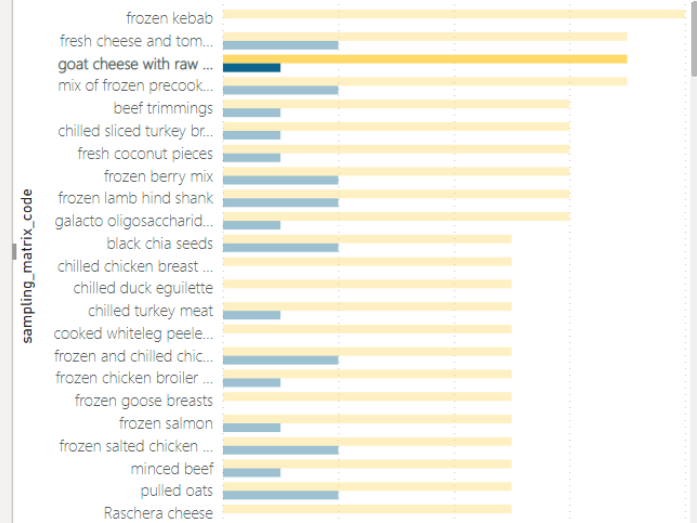
L. monocytogenes

Austrian Food Authority



Comparison

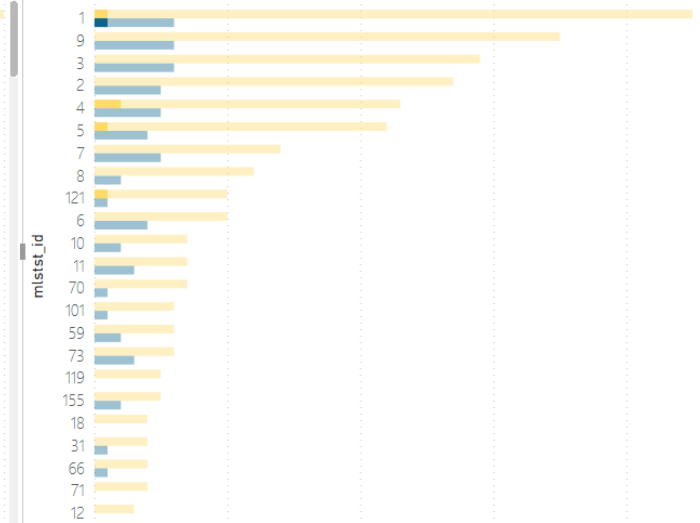
● Data All ● Data Owner Organisation



Data All and Data Owner Organisation

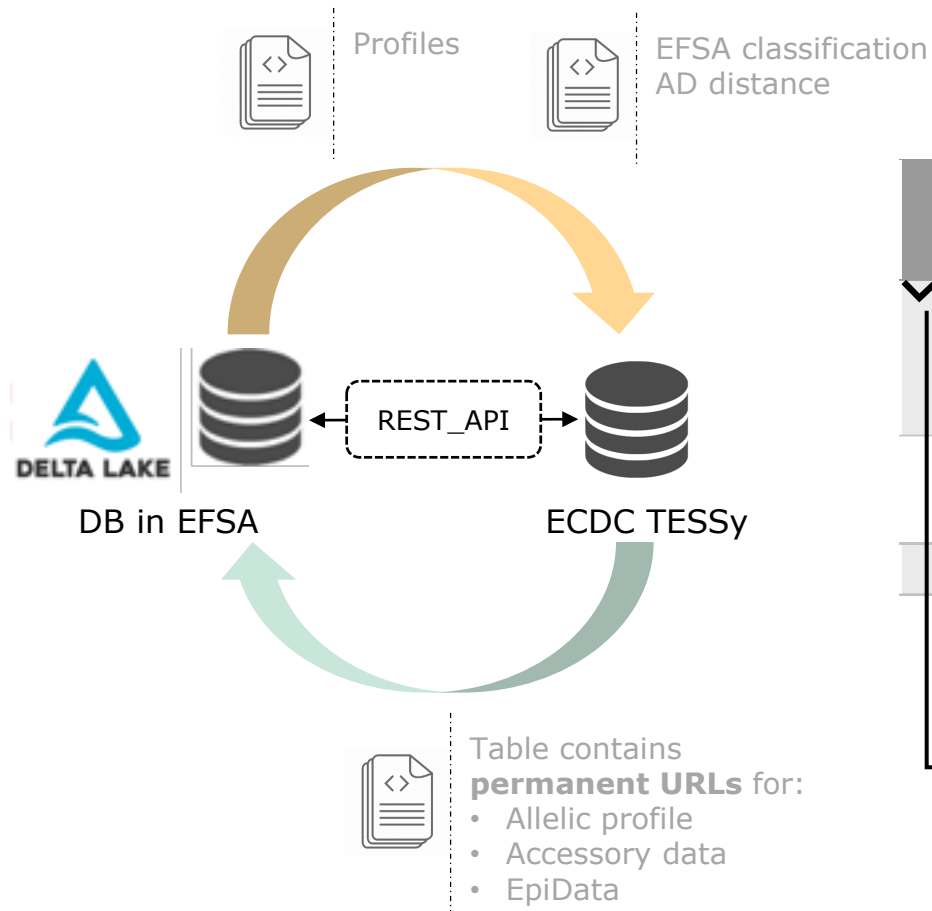
Comparison

● Data All ● Data Owner Organisation



Data All and Data Owner Organisation

ECDC integration: query to ECDC DB



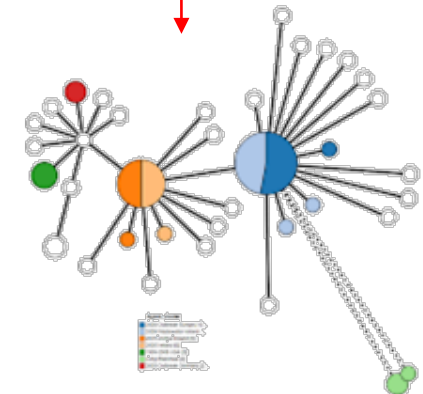
Monitoring calls

→ Each user sees own calls; EFSA SO and Admin see all calls

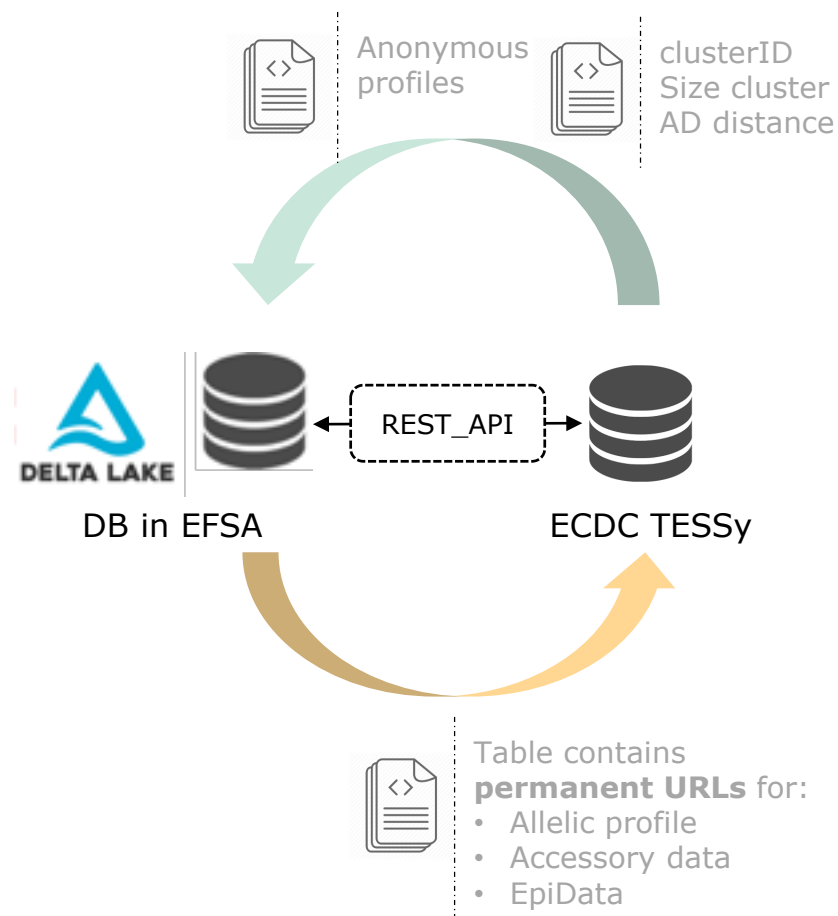
QueryID	Schema	EFSA user	Timestamp	Match	Value K	EFSAIDs	ECDC clusterID	Total matches at ECDC	ECDC entries	ValidTo	VIZ
✓ F	EC_v1	mario.bianchi@efsa.europa.eu	2021-01-19 03:14:07	yes	5	658364503 693435545	EC1412	3	XXxxX YYyyYY ZZzzYY	2021-04-19 03:14:07	Link
20210002	LM_v1	hans.muetermann@bfr.de	2021-01-19 09:10:56	no	7	652667603	-	-	-	-	-

Actions:

- Delete result of the query
- Download the ECDC entries and run a ViZ
- Rerun a new query with the same EFSA isolates



ECDC integration: query from ECDC



Monitoring calls

→ Only visible to the EFSA SO and Admin

QueryID	Schema	clusterID	Size ECDC	ECDC system user	TimeStamp	Match	Value K	UI	EFSAIDs	Report
20210001	EC_v1	EC1412	14	joe.smith@ecdc.europa.eu	2021-01-19 03:14:07	yes	5	724	658364503 693435545	link
20210002	LM_v1	LM2534	3	matti.hanninen@thl.fi	2021-01-19 09:10:56	no	7	n.a.	-	-

QUERY ID	EFSAID	Country of sampling	Reporting year	Year of sampling	Type of matrix	Description of the matrix of the sample taken	Description of the isolate species	Average AD	Max Ad	Min Ad
20200001	658364503	UK	2019	2019	FOOD	salmon smoked	Listeria monocytogenes	3	1	4
	693435545	PL	2018	2018	FOOD	salmon smoked	Listeria monocytogenes	5	3	7

- Data provider will have the opportunity to interact with the WGS portal programmatically
- Submission of allele profiles and typing data (in standardized JSON format)
- User would be able to submit data either by using the public available nextflow pipeline or by using in house method
 - Need to be interoperable with the EFSA data model
 - Need to be comparable (critical component such as assembly, allele calling, hash function and schema)

Thanks for your attention



EFSA is committed to:

**Excellence,
Independency,
Responsiveness and
Transparency**

www.efsa.europa.eu

Contact:
mirko.rossi@efsa.europa.eu



Subscribe to

www.efsa.europa.eu/en/news/newsletters
www.efsa.europa.eu/en/rss



Engage with careers

www.efsa.europa.eu/en/engage/careers



Follow us on Twitter

@efsa_eu
@plants_efsa
@methods_efsa