One health surRveillance Initiative On harmonization of data collection and interpretatioN (ORION)

WP3 – One Health Surveillance Harmonisation Infrastructure

The ORION project

The ORION project, launched in 2018, aims at establishing and strengthening inter-institutional collaboration and transdisciplinary knowledge transfer in the area of surveillance data integration and interpretation, along the One Health (OH) objective of improving health and well-being.

Through three main work packages (WP), ORION's specific goals can be summarized as the delivery of three main resources:

- a "OH Surveillance Codex" (WP1) - a high level framework for harmonised, cross-sectional description and categorisation of surveillance data covering all surveillance phases and all knowledge types;
- a "OHS Knowledge Hub" (WP2) - a cross-domain inventory of currently available data sources, methods / algorithms / tools, that support OH surveillance data generation, data analysis, modelling and decision support; and
- "OHS Infrastructural Resources" (WP3) – that are practical, infrastructural resources forming the basis for successful harmonisation and integration of surveillance data and methods.

Developed solutions will be exemplified and validated during several One Health pilots, which will support the operationalization and implementation of OH surveillance solutions on a national level and provide crucial feedback for future development and dissemination actions.

Trainings and workshops will be offered (WP4) to support and integrate with other EJP projects in their data harmonisation efforts.

WP3 – Data Interoperability resources

The development of a framework of One Health Surveillance (OHS) faces varied **data interoperability** challenges - among institutions, across health surveillance sectors, and among countries. Interoperability is used here to mean "*the ability of different information technology systems and software applications to communicate, exchange data, and use the information that has been exchanged*"[1].

During the first year of work we conducted a review of the data workflows within public health, animal health, and food surveillance, in order to map the opportunities and challenges to support a OHS data workflow. We interviewed different people with experience in collating surveillance data, discussed national workflows, and also the practices of mandatory reporting to the European Centre for Disease Control (ECDC) and the European Food Safety Agency (EFSA).

EFSA and ECDC have done significant work, in their respective domains, to solve the problem of **structural interoperability** among datasets from different countries. As a result, standardised datasets collating surveillance information at the European level already exist, and can be accessed through different resources made available by these agencies. **Semantic interoperability**, on the other hand, is concerned with ensuring the integrity and *meaning* of the data across systems. Semantic interoperability is particularly important in One Health in order to allow *data reuse across sectors*, and even reuse of data for research and knowledge discovery.

[1] HIMSS Dictionary of Healthcare Information Technology Terms, Acronyms and Organizations, 2nd Edition, 2010, Appendix B, p190

A schematic overview of the current data workflows is shown in Figure 1.

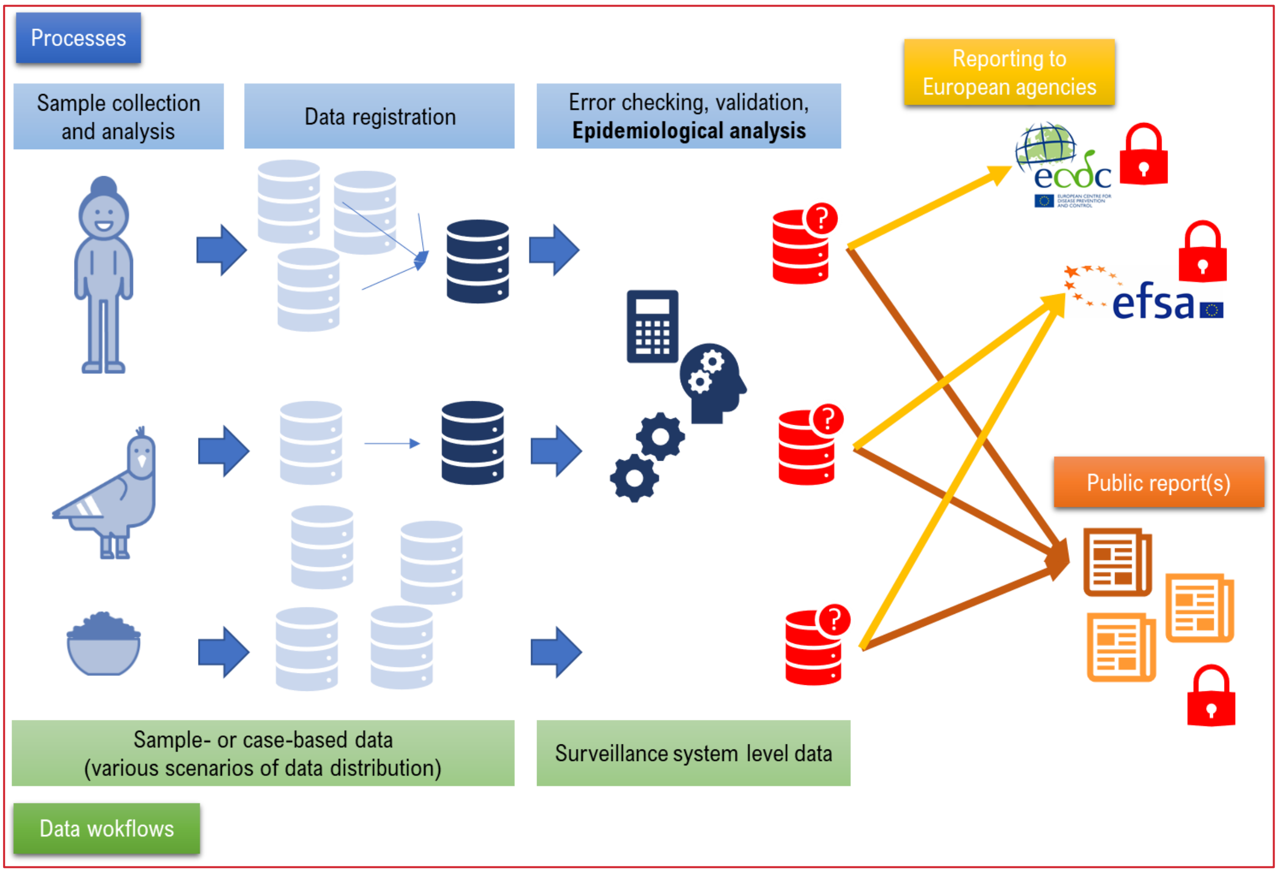
Processes
Sample collection and analysis
Data registration
Error checking, validation, Epidemiological analysis
Reporting to European agencies
ecdc
efsa
Public report(s)
Sample- or case-based data (various scenarios of data distribution)
Surveillance system level data
Data wokflows

Figure 1. *A illustrative view of current data workflows in public health, animal health and food surveillance.*

Focus has been given to the process of collating raw data from various surveillance activities to provide an epidemiological analysis of the hazard situation and the achievement of surveillance goals. The resulting output, here referred to as "***surveillance system level data***", is the data reported to European agencies or made publicly available in yearly reports. We have focused on this level of data aggregation after our initial requirement analysis, which identified that these are the data with the highest potential for sharing across sectors, possessing the highest quality (already validated, error checked and subjected to epidemiological analysis), and yet, the least frequently stored in reusable formats.

Figure 1 depicts no particular country, but a common workflow according to our requirement analysis work[2]. Data workflows are run in parallel in each health sector, with little to no sharing of data across sectors until very late in the process. Various scenarios of data distribution exist, making the job of data collation at the surveillance system level more or less challenging. While raw data are always stored in traditional databases, it is unclear whether countries systematically collect and store surveillance system level data, or whether these data are only collected at the time of mandatory reporting to EFSA and ECDC. Even after reporting, these data are not always (re)stored in databases for reuse at the country level, and a duplication of workflows, for instance between producing data for European agencies and producing data for national reports, is evident.

EFSA and the ECDC work together to analyse the reported data, but the data themselves are not interoperable, as the terminologies adopted across sectors are not semantically integrated. Data from some countries are available on the web through self-published national reports, but these reports are generally available as text/PDFs, so the data behind them is not actually accessible.

[2] ORION's WP3 first deliverable will be published publicly in January 2019. In that deliverable we will detail the results of our requirement analysis, with all the detailed interviews, inventories and a literature review.

Figure 2 depicts a scenario in which surveillance system level data has achieved semantic operability across sectors.

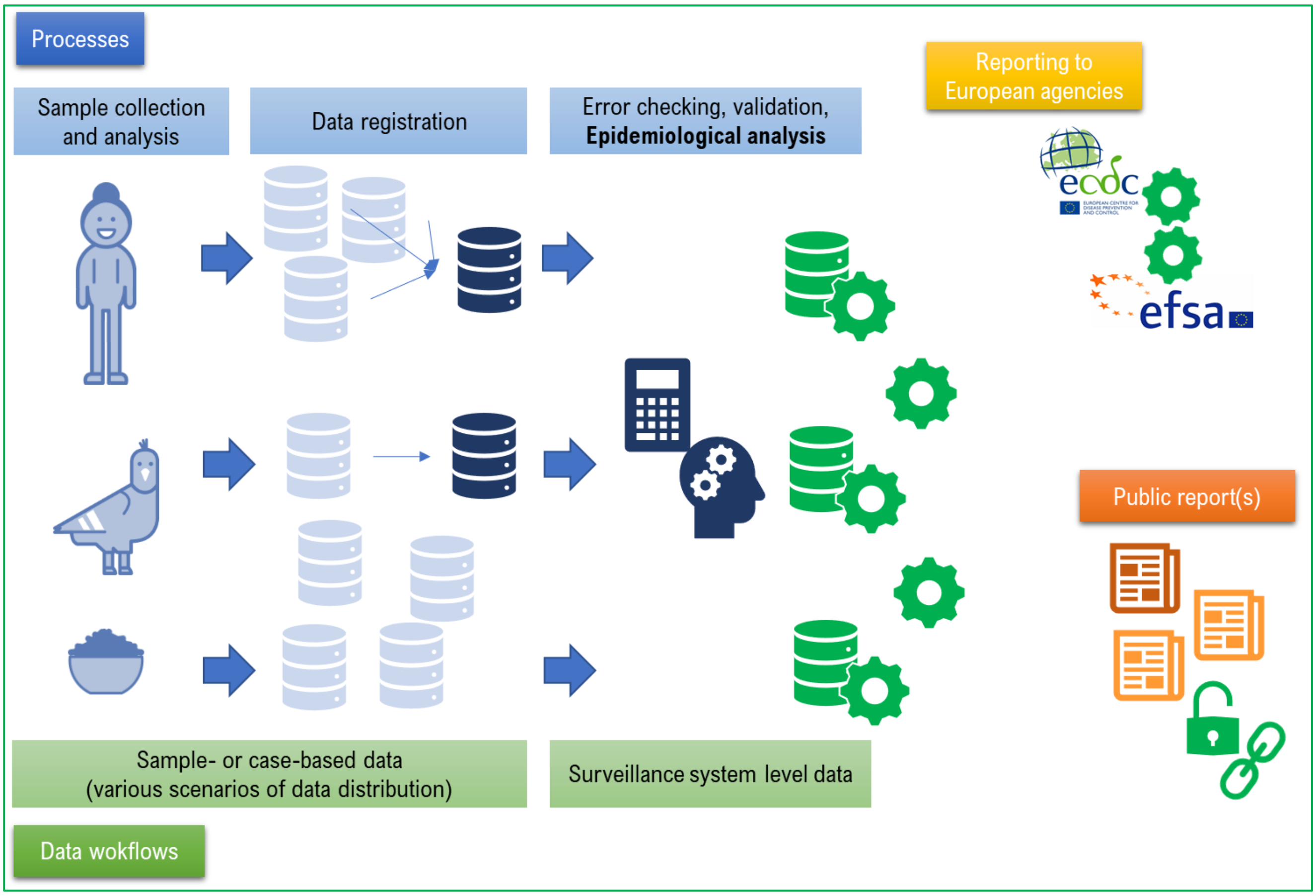
Processes
Sample collection and analysis
Data registration
Error checking, validation, Epidemiological analysis
Reporting to European agencies
ecdc
efsa
Public report(s)
Sample- or case-based data (various scenarios of data distribution)
Surveillance system level data
Data wokflows

***Figure 2.** A data workflow in which semantic data interoperability has been achieved among health sectors.*

In this scenario, no change in the current governance or accessibility to data is assumed. No new systems or extra workflows are put in place. But structures are created to annotate and store existing data, within the current data workflows, in formats that are interoperable. We will detail below *how* this could be achieved, but first we highlight the main **advantages of reaching this scenario of data interoperability**:

1) Surveillance system data is collated, validated, subjected to error checking and epidemiological analysis only once. After that, they are stored in a system agnostic, machine interpretable format, which can be reused for all the purposes these data are needed for: reporting to European agencies, national reports, or analyse of historical trends across years. Time and resources are saved by reducing duplication in workflows.
2) The three sectors become part of the same "surveillance engine". Since data are interoperable, epidemiological analysis that take into consideration the situation across sectors is a matter of strengthening communication. With no barrier to data integration, collaborative work can start early in the process of data collation.
3) Data interoperability across sectors is propagated with the data, so data reported to European agencies is also interoperable, that is, between the ECDC and EFSA.
4) The national surveillance reports can be published as "linked data"[3]. Software agents can be pointed to the address where the report was published to "read data", the same way humans read the PDF.

[3] Linked Data refers not only about datasets, but recommended best practices for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web. The Semantic Web is about making links between datasets that are understandable not only to humans, but also to machines. http://linkeddata.org/

HOW DO WE ACHIEVE SEMANTIC OPERABILITY?

The chosen solution is the development of an ontological framework for health surveillance. "*An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them*"[4].

Several terminology catalogues already exist in health and epidemiology, and we highlight in particular those implemented by EFSA and ECDC to achieve structural interoperability among member states (MS). Ontologies can incorporate these existing resources and re-use all their knowledge. But we move beyond the listing of concepts and include also "relationships" between concepts (semantics), creating a knowledge model for health surveillance. A machine-interpretable version of the domain knowledge offers several advantages, in particular:

- Use of automated reasoners to make inferences and detect errors in the data.
- Flexibility to accommodate to knowledge growth and updates.
- Reuse. Ontologies are meant to model specific pieces of knowledge, in a way that allows linking to complementary pieces. As epidemiology is highly multi-disciplinary, the use of ontologies allows us to piece together expertise from many different domains.
- Interoperability. Terminologies allow humans to understand each other and agree on what things mean. Ontologies allow software to talk to each other.

You can find background materials on ontologies and their use here: http://datadrivensurveillance.org/ontology/.

Two quick examples can demonstrate the power of **semantics** and the result in terms of **interoperability**.

The first is the "knowledge graph" used by google. Try googling a book you like, or a famous person. Try googling "Stephen Hawking". Google's knowledge graph on the bottom right of your screen will display things like their date of birth, family members, relevant work, etc. This is because "Stephen Hawking" is being recognized not just as a string of text, but as the specific concept of a "person". An ontology exists where it has been modelled what are the characteristics of a person: they have professions, they are connected to other persons through relationships such as spouse, parent, child. The machine "understands" you are searching for a person, rather than just looking the text string "Stephen Hawking" all over the web.

The second example is the data about air travel all around the world. Think about how "Expedia.com" is able to retrieve and compare data from a great number of flight providers – this is not backed by data sharing agreements, rather by their adherence to (the same) data annotation model[5]. It is also thanks to schema mark-up that your calendar can automatically recognize and add to your diary that flight event, once you get a confirmation email with your flight details.

The examples above emphasize the main principles we want to address:

1) Build an ontological framework for health surveillance that allows computers to understand and reason with current data terminologies in the same way that humans do, maximising the benefit to cost ratio of the effort put into producing surveillance data;
2) Improve usability of data inside the institutions who own and/or use the data, as well as the potential for reuse by external stakeholders and for research and discovery.

[4] Natalya F. Noy and Deborah L. Mcguinness. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. Available at http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

[5] https://schema.org/

HOW DO WE PLAN TO ACHIEVE THIS WITHIN THE SCOPE OF ORION

We made sure this task is achievable within the time period of ORION by focusing on a very specific part of the surveillance data workflow: the data already shared publicly through national reports. Moreover, we will start from only a few foodborne diseases as a proof-of-concept.

The "**One health surveillance pilot**" that will test the application of the proposed framework in practice will be a joint effort of the Swedish National Veterinary Institute (SVA), and the Public Health Agency of Sweden (FoHM), advised by EFSA and ECDC. SVA and FoHM have chosen the surveillance against ***Campylobacter spp.*** as the primary focus of the pilot, and in secondary focus we will also revise surveillance data for ***Salmonella spp***, and ***E.coli EHEC***.

In this pilot, we will work to improve coordination and cooperation among public health, animal health and food surveillance agencies, ***so that the production of the surveillance report becomes a true one health initiative, with communication and data sharing starting as soon as we start analysing the data***.

Our goal is to support collaborative data analysis among health sectors. At the end of the pilot, we aim to have the Swedish national surveillance report chapters on foodborne diseases published as "linked data", backed up by an ontology (a "key" for software agents to be able to make meaningful use of the data). The knowledge model (data plus "key") will be made publicly available. We will also publish guidelines to allow other MS to adopt the same structure, if desired, to support their internal collaboration across agencies, or even also publish public linked data.

Three parallel working groups were set up to make this possible:

1) **The ontology development track**: advised by ontology experts from the computer science field, and supported by expert elicitation from ORION partners, we are developing an ontology of health surveillance. All resources can be consulted here: http://datadrivensurveillance.org/ontology/
2) **The surveillance practice track**: this team is mapping the data workflows from data collation to production of the Swedish surveillance report, and working to improve inter-agency collaboration as early in the process as possible. This track aims to understand how the technology can support and strengthen the collaboration work. Moreover, it aims to understand how we can support the production of linked data through the existing practices, without adding an extra burden on those who carry our data analysis and publishing.
3) **The technical track**: a team of consultants is developing the necessary tools to support data annotation using the ontology, as well as storage and access to linked data.

By focusing on the surveillance report, we will be able to demonstrate the benefits of interoperable data in a specific part of the surveillance data workflow (Figure 1) which is currently a "dead end": once collated, data are published in PDF reports. By transforming these into "usable" *and* open "linked data", we create a new track for data sharing and interoperability across health surveillance sectors. After this "proof-of-concept" pilot, applying these methods to other parts of the surveillance data workflow, in the future, would lead to ever improving timelines in the production and sharing of data for One Health.

ORION-WP3: HOW TO GET INVOLVED

All resources for the animal health surveillance ontology (AHSO) are available here http://datadrivensurveillance.org/ontology/. Keep an eye on the project website above, or send an email to orion-wp3@datadrivensurveillance.org if you want to remain involved.