Data interoperability in One-Health Surveillance – a report from the ORION project

This is a summarized version of Deliverable JIP1-3.3, intended for a broader audience

DRAFT

Responsible Partner: 38-SVA

Contributing partners: 9-BfR; 10-FLI, 12-DTU; 13-SSI; 20-APHA; 21-PHE; 31-RIVM; 37-FOHM

This meeting is part of the European Joint Programme One Health EJP.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

GENERAL INFORMATION

European Joint Programme full title	**Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards**
European Joint Programme acronym	**One Health EJP**
Funding	**This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.**
Grant Agreement	**Grant agreement n° 773830**
Start Date	**01/01/2018**
Duration	**60 Months**

DOCUMENT MANAGEMENT

Deliverable	JIP1-3.3 - Revised OH Harmonisation Infrastructure Hub, including lessons learned from the OH pilots
WP and Task	JIP1 – WP3
Leader	Fernanda Dórea (SVA)
Other contributors	**Taras Günther (BfR), Mia Holmberg (SVA), Cecilia Jernberg (FoHM), Marika Hjertkvist (FoHM), Matthias Filter (BfR), Alessandro Foddai (DTU), Johanne Ellis-Iversen (DTU), Charlotte Cook (APHA), Joanna Lawes (APHA), Lesley Larkin (PHE); Ingrid Friesema (RIVM); Maria-Eleni Filippitz (sciensano); Mickael Cargnel (sciensano); Geraldine Boseret (sciensano); Karin Lagesen (NVI)**
Due month of the deliverable	**M36 (later postponed to M42 with agreement of the coordination)**
Actual submission month	**M42**
Type ***R: Document, report*** ***DEC: Websites, patent filings, videos, etc.*** ***OTHER***	R
Dissemination level ***PU: Public*** ***CO: confidential, only for members of the consortium (including the Commission Services)***	PU

DRAFT

This meeting is part of the European Joint Programme One Health EJP.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

ONE-HEALTH DATA INTEROPERABILITY INFRASTRUCTURE IN THE EUROPEAN UNION

WP3 in ORION was planned to tackle infrastructural resources related to data harmonization in One-Health surveillance (OHS). During the requirement analysis carried in year 1 (and reported in the deliverable JIP1-3.1 (available here), it became clear that to support collaborative data analysis across health sectors this WP should focus on solutions to document context and preserve meaning of surveillance data across health sectors – that is, solutions to promote ***semantic interoperability***.

> **Semantic interoperability** is concerned with ensuring the integrity and *meaning* of the data across systems[1]. This is particularly important in health surveillance in order to allow *data reuse across sectors*, and even reuse of data for research and knowledge discovery[2].

Working in three parallel working groups (*knowledge modeling*, *technical develop*ment and *surveillance practice*) this WP has focused on two overarching goals:

1) Build a knowledge model for one-health surveillance that allows computers to understand and reason with current data terminologies in the same way that humans do, maximizing 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.

The focus on semantic interoperability has allowed this WP to seek to add value to the integratory activities already performed by EFSA and ECDC, and to develop tools for data interoperability that can be implemented in any scenario of data governance, that is, respecting current data sharing barriers. None of the outcomes of this WP rely on, or promote changes in current data sharing practices. Rather, they support and promote adoption of the FAIR principles of findability, accessibility, interoperability and reusability (https://www.force11.org/group/fairgroup/fairprinciples).

This report is a detailed account of the outcomes produced by this WP to contribute to data FAIRness in general, and data interoperability in particular. It is structured in the following way:

- ***Introduction and background***. Before presenting the results of ORION, we give a quick overview of some concepts of linked data, data interoperability and data FAIRness.
- ***Section 1: OHS interoperability tools***. Data interoperability tools produced in this WP, which are publicly available, are presented.
- ***Section 2: using interoperability tools to publish FAIR data.*** This WP has also tested data workflows in practice, using tools developed in ORION or other existing tools. Examples of surveillance data production, reporting and sharing are described.
- ***Section 3: opportunities to connect data in OHS.*** Based on the "proof of concept" workflows presented in section 2, we reflect on the opportunities to connect OHS data.
- ***Conclusions and lessons learned: considerations for a FAIR-ER OHS future.*** We present reflections on the lessons learned from a general OHS perspective in the European Union.

[1] Definition of Interoperability. In: HIMSS Dictionary of Healthcare Information Technology Terms, Acronyms and Organizations. 2nd edition. 2010. p. 190
[2] Cardoso L, Marins F, Portela F, Santos M, Abelha A, Machado J. The next generation of interoperability agents in healthcare. Int J Environ Res Public Health. 2014

LINKED DATA, DATA INTEROPERABILITY AND DATA FAIRNESS

Connected data is smarter data –
Even in the face of intelligent applications, disconnected data result in dumb behaviour.
Dean Allemang and Jim Hendler[3]

The construction of smarter applications does not depend on smarter data, but on making sure that "the right data can get to the right place, so that smart applications can do their work" (Allemang and Hendler, 2011)[3]. Surveillance is intrinsically an activity that depends on the connection of many disparate sources of data. In OHS, this complexity is amplified by the need for data across many different health sectors and knowledge domains. Keeping data connected, and most importantly, ensuring that the context of data is preserved across this network of users is a great challenge. To add complexity still, knowledge is constantly evolving. If we are able to connect the right applications, we can develop an "ecosystem of solutions" for OHS, but the ecosystem needs to stay stable and synchronized along time.

The ***linked data model*** proposes that the development of applications that are able to connect to each other and produce consistent, future-proof results, relies on the ***separation of data from knowledge***. In this model, data should describe entities in the world, rather than store information (which is an interpretation of the data based on some previous knowledge). Information is generated by applications that query data based on a layer of knowledge. This "knowledge layer" models the connections between the entities represented in the data, and it is this layer that mediates the connections between applications. Any translations or assumptions needed to connect disparate data can be modelled in this layer, and as knowledge evolves, the model can be evolved, without losing the connection between the data sources. This can be better exemplified by an example.

Consider a diagnostic test which has a numerical result: for instance an ELISA reading. An absorbance value can only be interpreted if we know the cut-off value above which the result is to be interpreted as positive. If we test an animal serum sample with a given ELISA test, and record the result as "positive", we are recording *information*. If we provide data to our applications only as "positive", we won't know if this value is compatible with other test protocols used in different laboratories and other countries, and changes in the protocol may mean that future results cannot be compared to old results. The context was not preserved (false conclusions can be taken when comparing these data to other sources), and the information is not future-proof. In this example, the actual entities in the process are the sample used, the diagnostic assay used, and the absorbance value in the result. If we record all of these data explicitly, then the translation of this information into a positive or negative result based on a cut-off value can be made by a human – but can also be delegated to applications. And if we are connecting various applications, they can "communicate" and translate values among them, by stating explicitly the diagnostic protocol and cut-off values used. "Applications" in this case can be automated systems (machines), people across different institutions, or it can even be ourselves looking at the data some time in the future.

Every time we record information based on an assumed knowledge, those data are only useful if the user interprets them with the same knowledge, and working under the same assumptions. *Within* health sectors, this can be sometimes enforced through data standardization. *Across domains*, as in the case of OHS, this is not possible or even desirable. When we reuse data collected within animal health, public health and food safety surveillance in order to generate OHS information, we want data to have been structured based on and preserving their original context of production. But reusing those data into a

[3] Allemang, D., Hendler, J., 2011. Semantic Web for the Working Ontologist: effective modeling in RDFS and OWL. Morgan Kaufmann. 384p. Paperback ISBN: 9780123859655. eBook ISBN: 9780123859662

DRAFT

OHS context requires some knowledge about how these sectors connect. Some of this knowledge is straightforward to humans, but requires proper knowledge modeling to empower smart data applications.

Allemang and Hendler (2011) state that modeling, explicitly, the knowledge needed to convert data into information enables:

1) Communication among people, as it makes the knowledge and assumptions used to process the data explicit;
2) Discovery of patterns in the data and predictions;
3) Mediation among multiple viewpoints; and
4) Development of a common collection of knowledge, which can be developed and evolved as a common effort.

Points 1) and 3) are particular important across domains – contexts are intrinsically different between the health sectors involved, and this impacts both how they record and how they use their data. Knowledge models allow explicit exploration of the differences, while creating a representation for the commonalities or the possible links and translations that can be applied to use data across sectors.

Semantic modeling – communicate, explain/predict, mediate

The separation of data and knowledge under the linked data model relies on three main necessary pillars:

1) Creating a knowledge model for the desired application
2) Linking data to the knowledge model
3) Implementing applications which use the knowledge model to query the data.

There are a number of modeling languages available for creating a knowledge model, with different levels of expressivity:

- ***RDF – the Resource Description Framework***, is the basic framework used for the construction of interlinked applications in the "Semantic Web". RDF provides a mechanism for allowing anyone to make a basic statement about anything, and for layering statements into a single model. It has been a recommended language of the W3C (the World Wide Web Consortium, an international community that develops open standards to ensure the long-term growth of the Web) since 1999.
- Since 2004 the W3C recommends the use of ***RDFS, the RDF Schema language***, which adds to RDF by allowing expressivity of the basic notions of commonality and variability familiar from object languages, namely classes, subclasses and properties.
- ***OWL - Web Ontology Language***, adds *logic* to semantic modeling, allowing expression of detailed constraints between classes, entities and properties.

RDF allows us to declare concepts – or classes – and the properties connecting them. We can for instance declare the concept, or *class* of a "person", and model one of the *relationships* between persons to be "is biological father of". RDF can be used to declare that individual people are instances of the class "person" and to connect two people by this relationship. Figure 1 shows how we would normally see this is an Excel spreadsheet (A) and how this would be declared in RDF (B). If we also want to use logic to model constraints about this relationship, for instance that a person can only have one biological father, we can model this logic in OWL (Figure 1-C). The logic modelled in OWL can be used to check data for error, infer relationships between instances based on logic (if A *is father* of B, then B *has father* A), or discover knowledge based on the logic, for instance to discover who are the grandparents of a person, using the logic that grandparents are parents of their parents.

Resources described using RDF can have "human friendly" labels such as shown in Figure 1, but they are made unique by a ***Uniform Resource Identifier (URI)***. In the example shown, all concepts, relationships and individual persons listed would have their own URI. The concept "person" will have a unique URI, and the specific instances of a person - "Ada Lovelace", "Lord Byron" – will also have their

DRAFT

own unique URIs. All things "said" about "Ada Lovelace" anywhere in the web can be mapped to that unique person using this URI. Similarly, any model can be built to say additional things about a "person". As long as the properties in these new models point to the same URI used for "person" in the model shown in Figure 1, these models will complement each other. Reasoners can be used to check if the logic between these models is consistent, or discover knowledge in data based on these models combined logic.

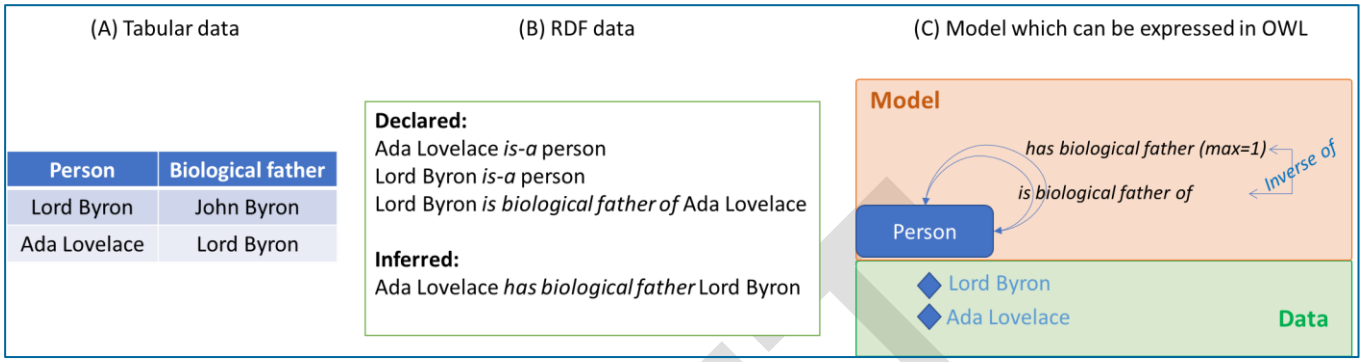

Figure 1. Example data in tabular format (A), RDF format (B), and a simple model of how the concepts in the data are connected, which can be modelled in OWL (C).

In the simple ELISA example given before, if we store the cut-off information in the model, then we use the model to retrieve positive *results as a query*: "all observations which absorbance reading was equal or greater than a given value. This instead of declaring, in the data, observations to be positive or negative. The data user is able to choose the constraints applied to the data at the time of retrieval.

This has an obvious application for mediation between contexts of data creation and data usage.

In section 3, we discuss the view of OHS as a knowledge model that connects data from different domains, without imposing any changes to the data structure within individual health sectors.

DRAFT

SECTION 1: OHS INTEROPERABILITY TOOLS

As stated in Section 1, the promotion of semantic interoperability through the use of ontologies requires three main steps:

1) the development of a knowledge model covering the knowledge areas among which interoperability is to be achieved
2) annotation of existing data using this knowledge model
3) tools to consume these semantically annotated data

Figure 2 depicts this process in a One-Health surveillance (OHS) scenario.

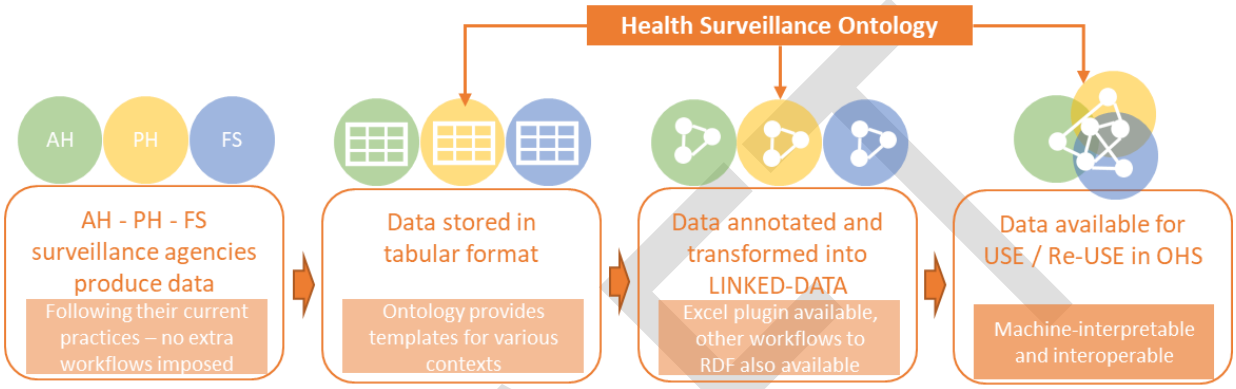

Figure 2. Vision of semantic interoperability in One Health supported by the Health Surveillance Ontology.

In the pursuit of supporting this vision, ORION has developed the ***Health Surveillance Ontology*** and provided proof-of concept workflows for the other steps of the semantic interoperability continuum. These are detailed in this section.

Health Surveillance Ontology (HSO)

> *"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].

In order to attend the need for a human- and machine-readable knowledge model for surveillance, ORION has developed a Health Surveillance Ontology reusing knowledge from existing ontologies, as well as reusing terminologies already commonly used in practice, such as those adopted by EFSA and ECDC. Identification of concepts and their specialization was informed by data examples from the various "OH pilots" carried out in ORION. Figure 3 depicts some of the main concepts in HSO, and how they can be used to annotate specific surveillance results. The goal is for surveillance methodology to always be associated with its results, and for numbers (such as number of positive samples) to be fully annotated with the context in which they were produced (such as surveillance objective, sampling design, etc). Figure 3 also shows some of the EFSA catalogues that were reused and made fully compatible with HSO.

[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

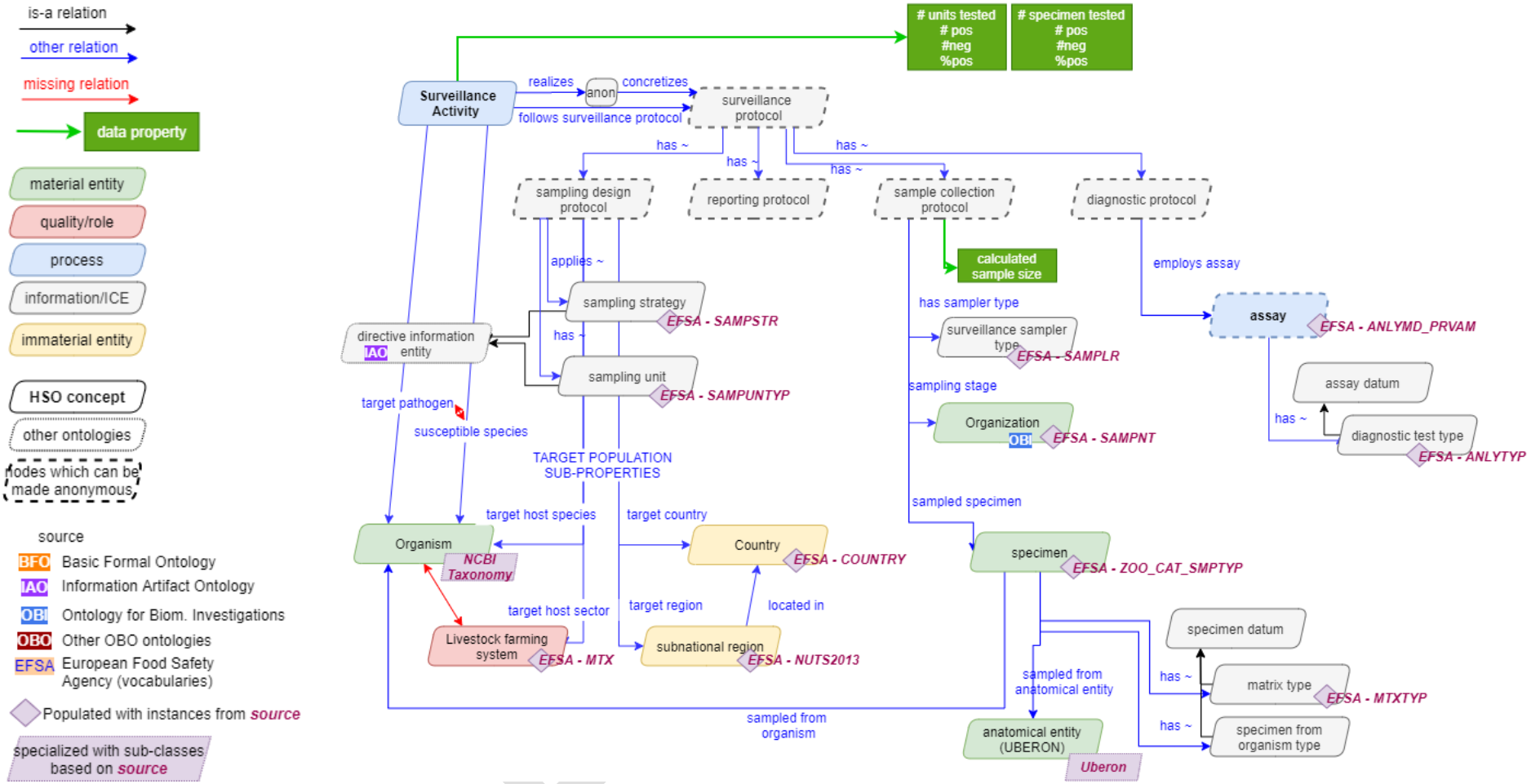

Figure 3: Some of the key classes in the Health Surveillance Ontology, and a schematic view of how they can be used to annotate surveillance results (strong-green rectangular boxes).

Full and up to date documentation of the ontology can always be found in HSO's globally unique and eternally persistent identifier: **https://w3id.org/hso**
This link is a permanent address provided by the W3C Permanent Identifier Community Group[5]. It ensures that the ontology will always have the same address, even if its hosting location is changed (for instance if in the future the ontology is maintained by other groups or projects).

This link is also subjected to content negotiation: humans accessing this link via browser will be referred to a page listing all ontology documentation and additional resources, such training materials. Software agents pointed to the same address will find the machine-readable codes for the knowledge model (written using the Web Ontology Language - OWL).

The ontology is also browsable on Bioportal: https://bioportal.bioontology.org/ontologies/HSO

The development of the ontology was described in deliverable JIP1-3.2 (available here). Since then, HSO has been accepted as a member of the Open Biological and Biomedical Ontology (OBO) Foundry (http://www.obofoundry.org/).

> *The Open Biological and Biomedical Ontologies Foundry is a group of people dedicated to build and maintain ontologies related to the life sciences. The OBO Foundry establishes a set of principles for ontology development for creating a suite of interoperable reference ontologies in the biomedical domain[6].*

HSO's persistent uniform resource locator (PURL) under OBO is:

http://purl.obolibrary.org/obo/hso.owl

Acceptance into OBO attests that HSO fulfils the community's principles[7]. We highlight the compliance to openness, unique identifiers for concepts, and interoperability (while orthogonal, that is, avoiding duplication) to other ontologies in the foundry (Figure 4).

More information about ontologies and their use and other supporting materials are available at http://datadrivensurveillance.org/ontology/.

Figure 4: HSO is interoperable with other ontologies in the Open Biological and Biomedical Ontology (OBO) Foundry, such as the Genetic Epidemiology Ontology and the Food Ontology.

[5] https://www.w3.org/community/perma-id/
[6] https://en.wikipedia.org/wiki/OBO_Foundry
[7] http://www.obofoundry.org/principles/fp-000-summary.html

DRAFT

Tools to annotate data using HSO

As HSO is, on itself, FAIR, it provides the required data annotation model for any data source to attend the FAIR principle of interoperability I2 ("To be interoperable: I2 (meta)data use vocabularies that follow FAIR principles").

The data annotation process is highly dependent on the data management tools used at each institution. In ORION we have identified that epidemiologists most frequently manipulate and exchange data in flat formats, such as ".xls", ".xlsx" or ".csv" formats. For that reason, we have developed tools for semantic annotation of data in Excel, and subsequent exportation of the data in Resource Description Framework (RDF) format (see Introduction for information about RDF).

1.1. The ExcelRDF plug-in

The Excel plug-in is free and open source, and it was developed in conjunction with the RealEstateCore project in Sweden. Codes for developers, as well as a guide to install the plug-in for users are available at https://github.com/RealEstateCore/ExcelRDF. ExcelRDF is a Visual Studio Tools for Office (VSTO) plugin.

The plugin reproduces, semantically, the way people understand tabular data (Figure 5). It allows a user to associate each column of a tabular dataset to a specific concept in the ontology, and all content of that column is then understood as subclasses or instances of the class assigned to the column.

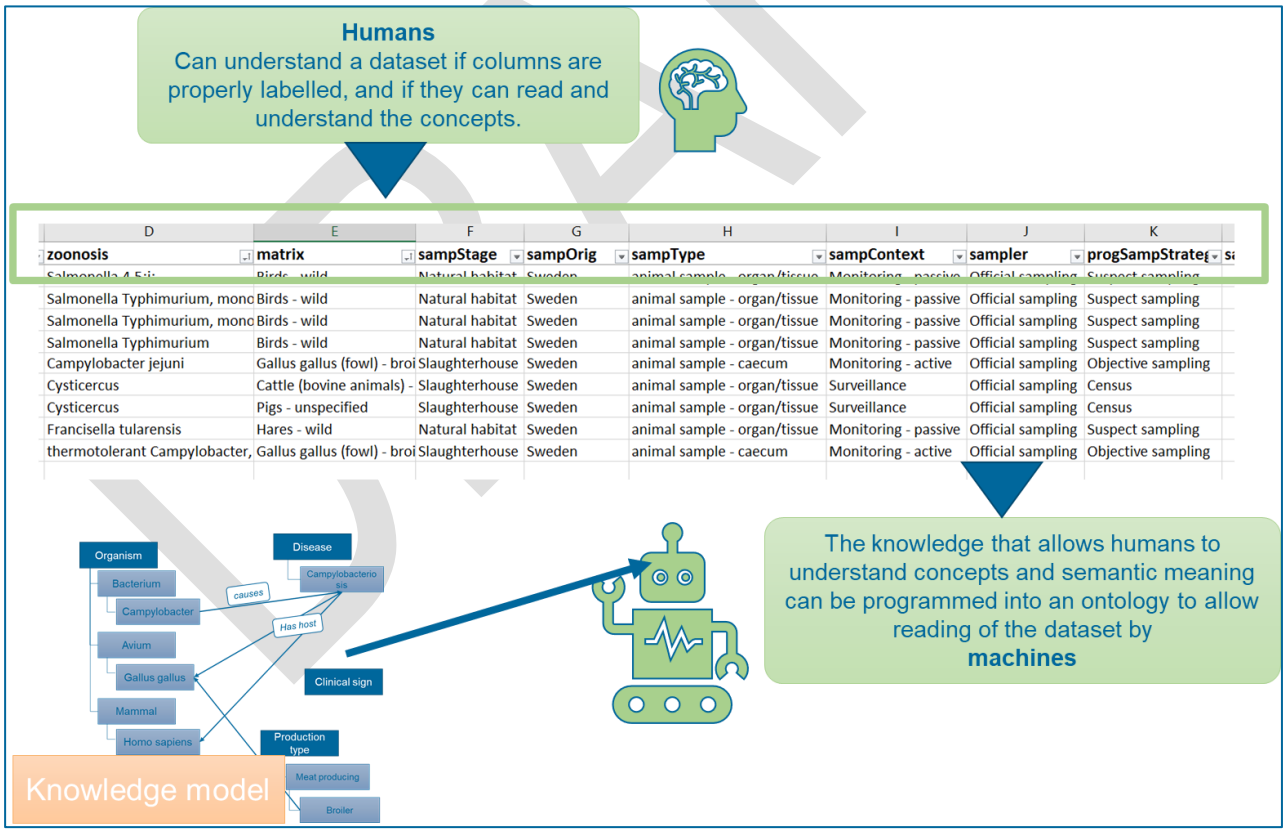

Figure 5. Processing of tabular data by humans and by machines, and the role of a knowledge model.

Once installed, the plugin appears in the "DATA" menu of Microsoft Excel, as shown in Figure 6. Users can use "Load ontology" to capture concepts directly from the ontology, and populate a blank Excel spreadsheet with columns which represent concepts form the ontology. As the ontology is quite extensive, we have created an empty template which users can download directly from the ontology homepage (http://datadrivensurveillance.org/health-surveillance-ontology-hso/).

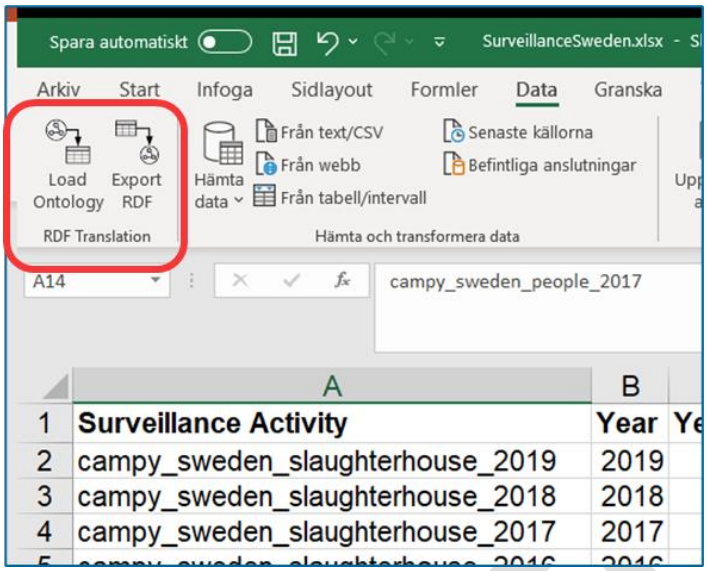

Figure 6. ExcelRDF plugin controls under the "Data" menu in Microsoft Excel.

The plugin assigns classes to columns using annotations. Users can also edit those annotations directly, rather than using "Load Ontology". In the example shown in Figure 7, the annotations are declaring that the first column contains unique instances of "Surveillance Activity" (Unique Resource Identifier (URI): http://purl.obolibrary.org/obo/HSO_0000001); and that the second column expects data in the format of integer, which will then be assigned to the Surveillance Activity through the data property "surveillance activity year" (URI http://purl.obolibrary.org/obo/hso#HSO_0000213).

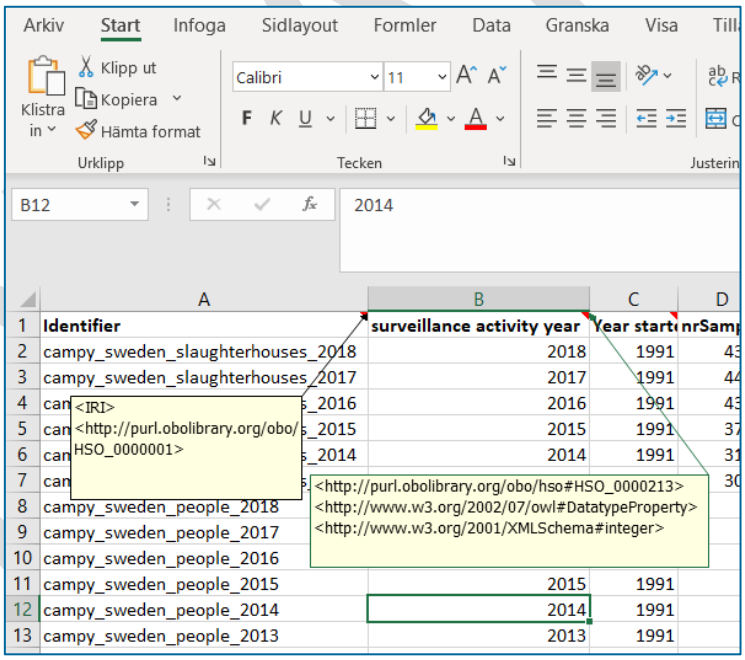

Figure 7. Annotations used by the ExcelRDF plugin to associate columns to ontology concepts.

Data is filled and manipulated in the Excel spreadsheet as usual. When a user clicks on the "Export RDF" button (Figure 6), an RDF version of the data is generated.

1.2. RDF and Excel converter Web applications

Web based tools were also developed- Their purpose is similar to the ExcelRDF plugin, but the conversion process is fully automated and needs less user interactions. The tool are developed with the open source Software KNIME (www.knime.com). This workflow is hosted at the BfR's KNIME Web Server infrastructure, and therefore does not require any installation process by the user. They can be accessed via this link:
https://foodrisklabs.bfr.bund.de/one-health-linked-data-toolbox/.

1.2.1. Excel to RDF converter web application

The Excel to RDF workflow is separated into three steps: (i) input step, (ii) control step, and (iii) output step.

In the input step the user can provide an Excel file with the same structure described for the ExcelRDF plugin via an upload button. Further the user can decide which sheet within the uploaded Excel file is to be converted (Figure 8). An example file is provided for download under the upload button.

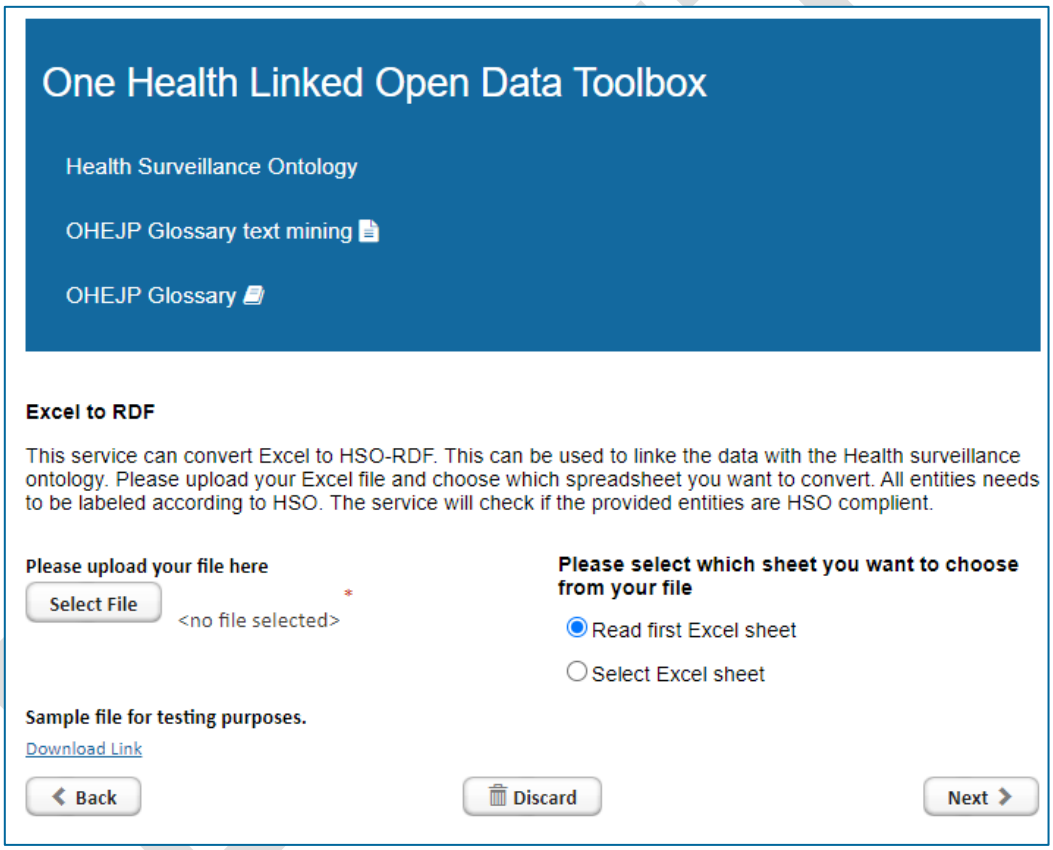

Figure 8. Input step of the Excel to RDF Web application. In this section the user can provide an Excel file to be converted into the RDF format.

After uploading the Excel file the user can move to the next step. The workflow then checks all the column names in the selected spreadsheet against concepts found in the ontology. In the control step (Figure 9) the user can check the coverage of the column names with the concepts available in HSO. Matches are indicated with a check mark and non matches as red crosses. In case of concepts not found in the ontology, the user can review the file, or leave as it is, which will cause columns not identified as specific ontology concepts to be imported as annotations (textual descriptions). No information is lost in the conversion. It is also possible to suggest new concepts or labels for HSO via the HSO GitHub Page (https://github.com/SVA-SE/HSO/issues).

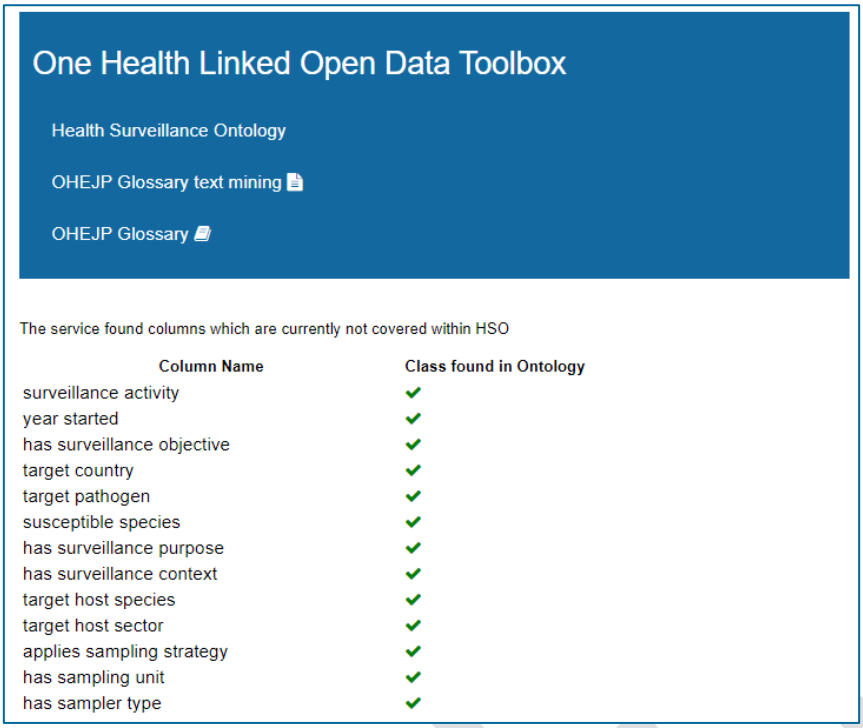

Figure 9. Control step of the Excel to RDF Web application. In this section the user can control the coverage of the columns within the HSO concepts.

Upon clicking next, the data table will be converted into RDF by matching the content with HSO. Column names will be automatically transformed into the corresponding HSO concept and the content of each column into the corresponding sub classes and instances of HSO.

In the output step (Figure 10) of the workflow the user gets the link to download the converted RDF and can see the content of the file. Instances in the dataset are assigned a uniquely generated URI which can be used to query the data.

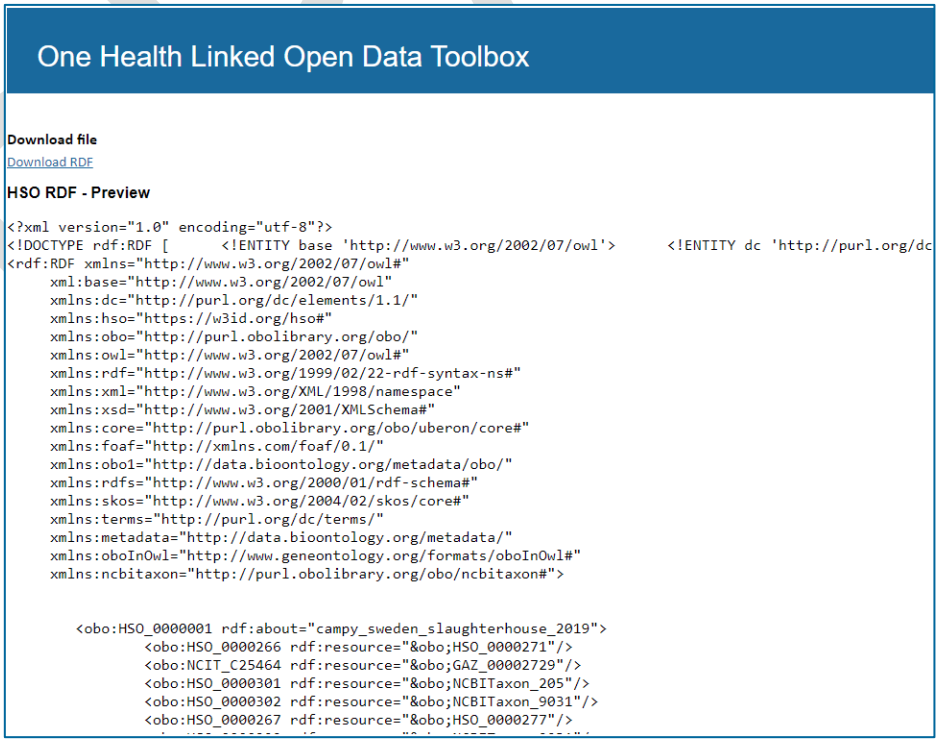

Figure 10. Output step of the Excel to RDF Web application – an RDF version of the uploaded data.

.

1.2.2. RDF to Excel converter web application

The workflow to convert RDF back to an Excel file works in a similar way and has the same user interface, but has only two sections.

In the input step (Figure 11) the user can upload an RDF file. An example RDF file is available.

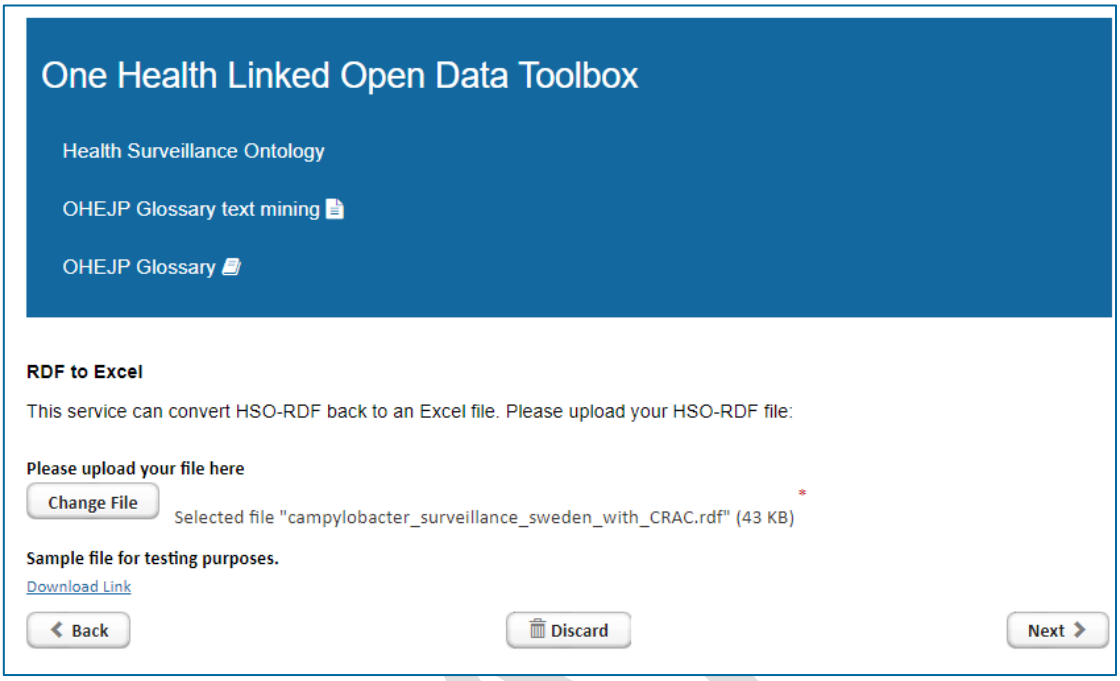

Figure 11. Input step of the RDF to Excel Web application.

Once the file is uploaded and the user presses next, the content of the RDF file will be automatically matched with the HSO. The user will be directed to the second section, the output section (Figure 12), where the content is shown in tabular format. This application is built with KNIME and deployed in the BfR's KNIME Web Server as well. It can be accessed at: https://knime.bfr.berlin/knime/#/EJP_ORION/OH-LOD%20Toolbox/LOD_Converter/RDF_to_EXCEL.

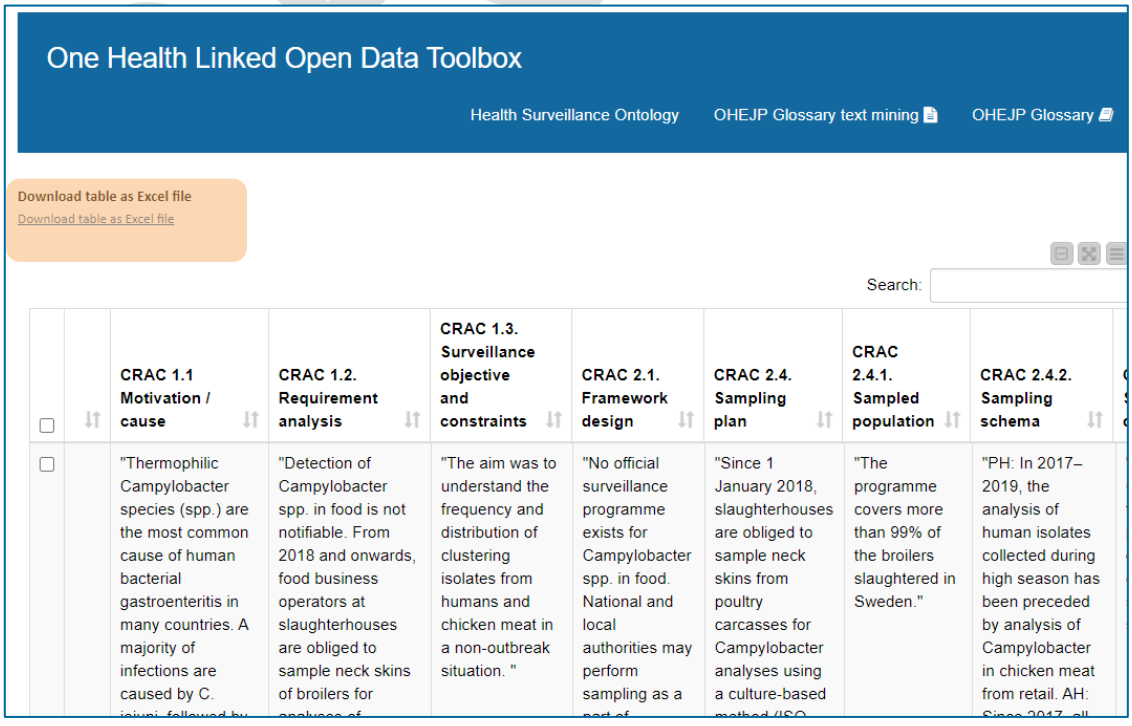

Figure 12. Output step of the RDF to Excel Web application.

This meeting is part of the European Joint Programme One Health EJP.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

SECTION 2: USING INTEROPERABILITY TOOLS TO PUBLISH FAIR DATA

The tools presented in Section 1 support semantic interoperability and the linked data model described in the Introduction.

Interoperability, in turn, is only one of the main challenges in computer-mediated, data-driven knowledge discovery. In order to "assist humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflow", the FORCE11[8] (The Future of Research Communications and e-Scholarship) community has described a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable – now widely known as the FAIR principles (https://www.force11.org/group/fairgroup/fairprinciples).

Although primarily used to drive FAIRness in the publishing of research data, the FAIR principles are applicable to any data of scientific value, and where maximization of the potential of generating knowledge from data is desirable. In the case of surveillance, the FAIR principles can guide the production of data that maximizes its utility for generating *evidence*.

The FAIR principles are listed in Table 1.

Table 1. Guiding principles to make data Findable, Accessible, Interoperable, and Reusable (FAIR)[9].

To be FINDABLE	F1. (meta)data are assigned a globally unique and eternally persistent identifier. F2. data are described with rich metadata. F3. (meta)data are registered or indexed in a searchable resource. F4. metadata specify the data identifier.
To be ACCESSIBLE	A1 (meta)data are retrievable by their identifier using a standardized communications protocol. A1.1 the protocol is open, free, and universally implementable. A1.2 the protocol allows for an authentication and authorization procedure, where necessary. A2 metadata are accessible, even when the data are no longer available.
To be INTEROPERABLE	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles. I3. (meta)data include qualified references to other (meta)data.
To be RE-USABLE	R1. meta(data) have a plurality of accurate and relevant attributes. R1.1. (meta)data are released with a clear and accessible data usage license. R1.2. (meta)data are associated with their provenance. R1.3. (meta)data meet domain-relevant community standards.

[8] https://www.force11.org. FORCE11 is a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing. Individually and collectively, we aim to bring about a change in modern scholarly communications through the effective use of information technology.
[9] https://www.force11.org/group/fairgroup/fairprinciples

As seen in Table 1, the production of FAIR data relies on adressing not only datasets contents (*data*), but also their *metadata*. As it can be found in the OHEJP OH Glossary[10], ***metadata*** is:

> *"literally, "data about data"; data that defines and describes the characteristics of other data, used to improve understanding of data and data-related processes."*

While metadata is intended to be used to document the context associated with specific data, the definition above implies that what constitutes metadata is, on itself, context dependent. For an animal sample being subjected to a laboratory test, for instance, information about how that sample was collected – the animal species, who collected the sample, where and when – can be considered by the laboratory personnel as metadata. Epidemiologists analysing results of a large number of collected samples may consider all of those details to be data, and their interpretation of the context of these data would require further "metadata" about the process that generated that sample collection, such as whether it was a result of a specific surveillance program, and what was the goal of that surveillance activity.

OHS depends on the collation of data collected at many different points of the farm-to-fork continuum, by different sectors and many different agents, each with different interests and objectives for the process of data collection/generation. All information collected and stored, which contributes to preserving the context of data collection is useful to promote data interoperability and reusability. *The discussion of whether people perceive the collected information as data or metadata, is not relevant. From a technological perspective, both can be expressed in the same formats and read by machines interchangeably.*

When sharing data however, for instance at the occasion of (cross-sectoral) collaborations, when publishing data, and in particular when publishing results of data analyses, it is more common to "package" defined sets of data – *datasets*. Datasets of common occurrence are spreadsheets (in formats such as Excel – *.XLS, *.XLSX, or plain comma-separated-values – *.CSV); editable documents (such as Microsoft Word files - *.DOC, *.DOCX); or Portable Document Format (*.PDF). In this report, we will use ***data*** to refer to any information contained ***within*** these datasets. No matter how many levels of information can be recorded – for instance data about samples in one sheet, and contextual information about the collection of these samples in another sheet in the same file – everything in these shared files will be called data. We use the term ***metadata*** to refer to ***information about these datasets***, which is not contained inside the dataset itself. If a dataset is made publicly available in a link, for instance, the access link is metadata of the dataset. The title of the file, a description of its content, the owner of the dataset, its license, etc, are the things we will refer to as metadata.

Data and metadata play complementary roles in achieving FAIRness.

Findability and Reusability – principles associated with publishing data with appropriate metadata (*F__R data*)

Within OHS, not all data we produce can be shared in formats that are accessible and interoperable. In particular, in the case of information meant for dissemination of results, or communication and collaboration among people, the priority is to publish the information in formats that are "human-friendly", not "machine-readable".

It is however important that those information are findable by their potential target users, and re-usable. We argue that datasets can be packaged in Findable and Re-usable ways writing FAIR metadata, even when data are, themselves, not FAIR. We will describe an example based on the Swedish One Health pilot.

[10] https://aginfra.d4science.org/web/orionknowledgehub/catalogue

All countries involved in the ORION project have carried at least one "One Health pilot" to test, in practice, the implementation of the tools and principles developed during the course of the project. All pilot reports are available publicly as ORION deliverables.

We will focus on the Swedish pilot in this section.

Every year since 2009 a national report on the outcome of surveillance activities of infectious diseases in animals and humans is produced in Sweden. The report is produced with contributions from the animal health, public health and food safety sectors. The Swedish National Veterinary Institute (SVA) coordinates the production of the report. The purpose of the pilot was to strengthen the "One-Healthness" of the process of collaboration across de three sectors.

Here, we focus on the final product – a PDF report – and the workflow developed to publish this report in a findable and reusable (F__R) format. In this case, the entire content of the report – text, figures, charts, etc – is considered the data, which is available in the format most suitable for its target audience of humans (mostly the general public), and not meant for publishing in machine accessible or interoperable format. Before the ORION OH-pilot, this PDF was published yearly online by SVA, but the lack of metadata information and, in particular, the lack of an explicit license, meant that this report was not findable nor re-usable.

Knowledge models exist for the annotation of "data about data", metadata. In the European Union, public sector information is made available through the EU Open Data Portal (https://data.europa.eu/). To publish datasets in this portal, data owners must specify metadata using the DCAT Application Profile for data portals in Europe (DCAT-AP[11]), which is a specification based on the Data Catalogue Vocabulary (DCAT) developed by the World Wide Web Consortium (W3C, w3.org). Sweden, as many countries, has its own chapter of the open data portal for public sector information (PSI), hosted at dataportal.se. A Swedish adaptation (profile) of DCAT-AP, the DCAT-AP-SE[12] is specified for publishing datasets in this portal.

We used the ExcelRDF plugin introduced in Section 1 to write the metadata for the Swedish surveillance reports published between 2006 and 2019 using the DCAT-AP-SE specification. We followed this workflow to publish it in the dataportal:

1) The metadata about the surveillance report was collected in an Excel file. Table 2 shows all the information collected. Title, description and publisher are mandatory fields.
2) These metadata were converted to RDF using the ExcelRDF plugin described in Section 1, generating the RDF file shown in Figure 13. Inspecting the RDF result, note that:
    a. Some fields have textual entries, such as title and description, and therefore the text provided by data owners was directly used.
    b. Some fields have a set of values available for choice, similar to a "Pick list". This is the case for instance for themes, accrual periodicity and language. In these cases the DCAT-AP(-SE) specification provides the URI for the choices available.
    c. In some cases, the metadata needs to point to another metadata set, describing specific related resources. For instance the publisher and the contact, in this case, must be declared separately with their own set of metadata. All information about SVA as an organization, and as the publisher of the dataset were stored in their own RDF file. It is the URI for this file that is then provided as metadata to the dataset. The links provided in Table 2 for publisher and contact point can be used to access these RDF files directly.
3) Metadata specifications for a dataset assume that the same data can be provided in multiple formats. These are called distributions. We will describe distributions in better details with the fully FAIR example presented further below. In the case of the surveillance report there is only

[11] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe
[12] https://docs.dataportal.se/dcat/en/

one data format, one distribution – the PDF file. A second set of metadata needs to be written for that distribution specifically. The only mandatory field for the metadata of a distribution in the access URL. For the example shown here, the metadata for this distribution was written in a separate RDF, then given a unique identifier (https://data.sva.se/dcat/surveillancereport/sr2019pdf.rdf), and this identifier is provided to the dataset in the field "distribution" (see Figure 13).

4) These RDF files were uploaded to an SVA catalogue to create their unique identifiers. This is actually a recursive process with steps 2) and 3). For instance, it is saving the RDF file generated in step 2 that creates it is unique place in the web: https://data.sva.se/dcat/surveillancereport/sr2019.rdf. But this address for the file is then used within the RDF file as its unique identifier.
5) As SVA is a public organization in Sweden, it has the right to its own catalogue in dataportal.se. Once the metadata files are uploaded to SVA's catalogue, they are automatically uploaded to the dataportal overnight. For this example, the result can be found at https://www.dataportal.se/en/datasets/59_1643/surveillance-of-infectious-diseases-in-animals-and-humans-in-sweden-2019, and a screenshot is provided in Figure 14.

Table 2. Metadata collected for the report "Surveillance of infectious diseases in animals and humans in Sweden, 2019". Fields shaded in green are mandatory.

Metadata field	Data collected
Title	Surveillance of infectious diseases in animals and humans in Sweden, 2019
Description	Surveillance of infectious diseases in animals and humans is the annual report describing the surveillance activities carried out in Sweden during the year. The report covers surveillance for important animal diseases and zoonotic agents in humans, food, feed and animals, carried out and compiled by experts from several Swedish governmental agencies, university and private industry with surveillance mandates along the entire food chain, from farm to fork.
Publisher	SVA
Distribution	PDF distribution, see text and Figure 2
Theme	Agriculture, fisheries, forestry and food
	Government and public sector
	Health
Keywords in English	surveillance; disease
Keywords in Swedish	övervakning, sjukdom
Contact Point	SVA
Periodicity	Annual
Date issued	2020-06-16
Language	English
License	CC-4
Related	Related resources: the dataset of Campylobacter surveillance (presented further below in this report) and a page with a list of all surveillance reports published previously

DRAFT

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
        xmlns:adms="http://www.w3.org/ns/adms#"
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:odrs="http://schema.theodi.org/odrs#"
        xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
        xmlns:dcterms="http://purl.org/dc/terms/"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
        xmlns:dcat="http://www.w3.org/ns/dcat#"
        xmlns:foaf="http://xmlns.com/foaf/0.1/"
        xmlns:dc="http://purl.org/dc/elements/1.1/">
<dcat:Dataset rdf:about="https://data.sva.se/dcat/surveillancereport/sr2019.rdf">
        <dcterms:title xml:lang="en">Surveillance of infectious diseases in animals and humans in Sweden,
2019</dcterms:title>
        <dcterms:description xml:lang="en">Surveillance of infectious diseases in animals and humans is the
annual report describing the surveillance activities carried out in Sweden during the year. The report covers
surveillance for important animal diseases and zoonotic agents in humans, food, feed and animals, carried out
and compiled by experts from several Swedish governmental agencies, university and private industry with
surveillance mandates along the entire food chain, from farm to fork.</dcterms:description>
        <dcterms:publisher rdf:resource="https://data.sva.se/dcat/main/svaAgent.rdf"/>
        <dcat:distribution rdf:resource="https://data.sva.se/dcat/surveillancereport/sr2019pdf.rdf"/>
        <dcat:theme rdf:resource="http://publications.europa.eu/resource/authority/data-theme/AGRI"/>
        <dcat:theme rdf:resource="http://publications.europa.eu/resource/authority/data-theme/GOVE"/>
        <dcat:theme rdf:resource="http://publications.europa.eu/resource/authority/data-theme/HEAL"/>
        <dcat:keyword xml:lang="en">surveillance</dcat:keyword>
        <dcat:keyword xml:lang="en">disease</dcat:keyword>
        <dcat:keyword xml:lang="sv">övervakning</dcat:keyword>
        <dcat:keyword xml:lang="sv">sjukdom</dcat:keyword>
        <dcat:contactPoint rdf:resource="https://data.sva.se/dcat/main/svaOrg.rdf"/>
        <dcterms:accrualPeriodicity
rdf:resource="http://publications.europa.eu/resource/authority/frequency/ANNUAL"/>
        <dcterms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2020-06-16</dcterms:issued>
        <dcterms:language rdf:resource="http://publications.europa.eu/resource/authority/language/ENG"/>
        <dcterms:license rdf:resource="http://creativecommons.org/licenses/by/4.0/"/>
        <dcterms:relation rdf:resource="https://old.sva.se/en/reports-and-publications-in-english/disease-
surveillance/disease-surveillance-reports"/>
        <dcterms:relation rdf:resource="https://www.dataportal.se/en/datasets/59_1684/campylobacter-
surveillance-in-sweden"/>
</dcat:Dataset>
</rdf:RDF>
```

Figure 13. RDF version of the metadata for the Surveillance of infectious diseases in animals and humans in Sweden, 2019, presented in Table 2.

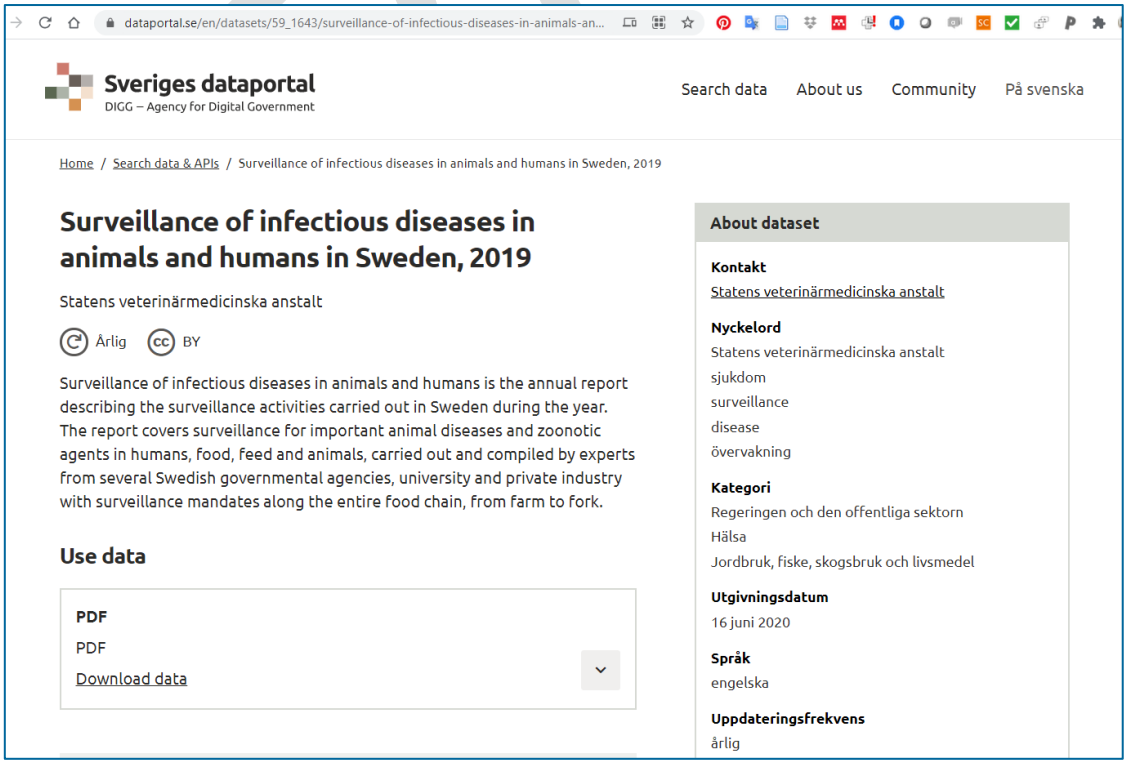

Figure 14. Screenshot of the Swedish data portal for Public Sector Information showing the published surveillance report.

Note that the actual PDF report was never uploaded to the dataportal or to any other location than its regular publication address at SVA's website. The writing of proper metadata, and publishing of these metadata using an open data portal allowed us to make this dataset reusable by linking the license information directly to the file, and by making it findable to indexing engines capable of searching datasets based on the specification used. Here we used DCAT as a global specification used for the description of datasets. In section 3 we discuss how the availability of OHS specifications could improve findability in the specific cross-sector health surveillance context.

Accessibility and Interoperability – principles associated with producing smarter data (_AI_)

The tools and workflows presented in Section 1 were used to create an accessible and interoperable OHS dataset from the Swedish OH pilot.

The "Surveillance of infectious diseases in animals and humans in Sweden" described in the OH-pilot is published yearly, containing results for all diseases subjected to surveillance in the country, for one specific year. We created a dataset that contains results presented for only one specific hazard – Campylobacter – but covering multiple years. The dataset, from here on referred to as the "Campylobacter surveillance dataset" is available at https://data.sva.se/opendata/surveillance/campy/campylobacter_surveillance_sweden.csv, and a screenshot is presented in Figure 15.

All data in the dataset is already available publicly through the annual surveillance reports (years 2010 to 2019), but in the RDF version the specific concepts are mapped to the Health Surveillance Ontology.

ORION has also developed a One Health Consensus Report Annotation Checklist (OH-CRAC[13]) to promote OHS report harmonization. The specific steps in OH-CRAC are available in HSO as annotation properties, that is, any surveillance activity declared using HSO can be annotated with textual information for each of the surveillance steps recommended in the OH-CRAC. To create the "Campylobacter surveillance dataset", we have extracted all surveillance methods and results from the surveillance reports published from 2010 to 2019 which could be translated into HSO concepts or properties. Moreover, for the main year of the OH-pilot (2019), we have pasted all text from the Campylobacter surveillance chapter into the relevant OH-CRAC fields (columns not shown in Figure 15, but accessible in the public file). The resulting file is therefore a tabular representation of information, but also preserves all text published in the original report, and indexed by the specific step of the surveillance process using the One Health Consensus Report Annotation Checklist.

To create a proof-of-concept of how surveillance data could be made available in accessible and interoperable formats, this dataset was converted to RDF using both available workflows presented in Section 1: the ExcelRDF plugin, and the KNIME workflow. In both cases, the result is an RDF file which provides a machine readable version of the human-friendly data presented in Figure 15.

DRAFT

[13] available in the OHS Codex: https://oh-surveillance-codex.readthedocs.io/en/latest/5-the-dissemination-principle.html

	A	B	D	F	G	H	W	X	Z	AA	AB
1	Surveillance Activity	Year	has Surveillance Objective	Target Pathogen	Susceptible species	has Surveillance purpos	Number tested	number positive	Total cases	Domestic cases	Imported cases
2	campy_sweden_slaughterhouse_2019	2019	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	4363	230			
3	campy_sweden_slaughterhouse_2018	2018	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	4331	377			
4	campy_sweden_slaughterhouse_2017	2017	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	4419	474			
5	campy_sweden_slaughterhouse_2016	2016	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	4389	678			
6	campy_sweden_slaughterhouse_2015	2015	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	3759	437			
7	campy_sweden_slaughterhouse_2014	2014	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	3162	363			
8	campy_sweden_slaughterhouse_2013	2013	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	3046	267			
9	campy_sweden_slaughterhouse_2012	2012	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	2346	217			
10	campy_sweden_slaughterhouse_2011	2011	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	2788	357			
11	campy_sweden_slaughterhouse_2010	2010	prevalence estimation	Campylobacter spp	Gallus gallus	detect changes	3357	444			
12	campy_sweden_people_2019	2019	case detection	Campylobacter spp	Homo sapiens	control			6693	2865	3828
13	campy_sweden_people_2018	2018	case detection	Campylobacter spp	Homo sapiens	control			8132	3645	4487
14	campy_sweden_people_2017	2017	case detection	Campylobacter spp	Homo sapiens	control			10608	6023	4585
15	campy_sweden_people_2016	2016	case detection	Campylobacter spp	Homo sapiens	control			11021	6893	4128
16	campy_sweden_people_2015	2015	case detection	Campylobacter spp	Homo sapiens	control			9180	4709	4471
17	campy_sweden_people_2014	2014	case detection	Campylobacter spp	Homo sapiens	control			8288	3709	4579
18	campy_sweden_people_2013	2013	case detection	Campylobacter spp	Homo sapiens	control			8114	3305	4809
19	campy_sweden_people_2012	2012	case detection	Campylobacter spp	Homo sapiens	control			7902	3155	4747
20	campy_sweden_people_2011	2011	case detection	Campylobacter spp	Homo sapiens	control			8214	3275	4939
21	campy_sweden_people_2010	2010	case detection	Campylobacter spp	Homo sapiens	control			8001	3143	4858
22	campy_sweden_food_2019	2019		Campylobacter spp							
23	campy_sweden_onehealth_2019	2019		Campylobacter spp							

Figure 15. Excel version of the "Campylobacter surveillance dataset".

All the way FAIR publishing

The workflow to annotate metadata presented earlier was also applied to the Campylobacter surveillance dataset to publish it in the Swedish PSI data portal, making it a FAIR resource which can be accessed here: https://www.dataportal.se/en/datasets/59_1684/campylobacter-surveillance-in-sweden. A screenshot is provided in Figure 16.

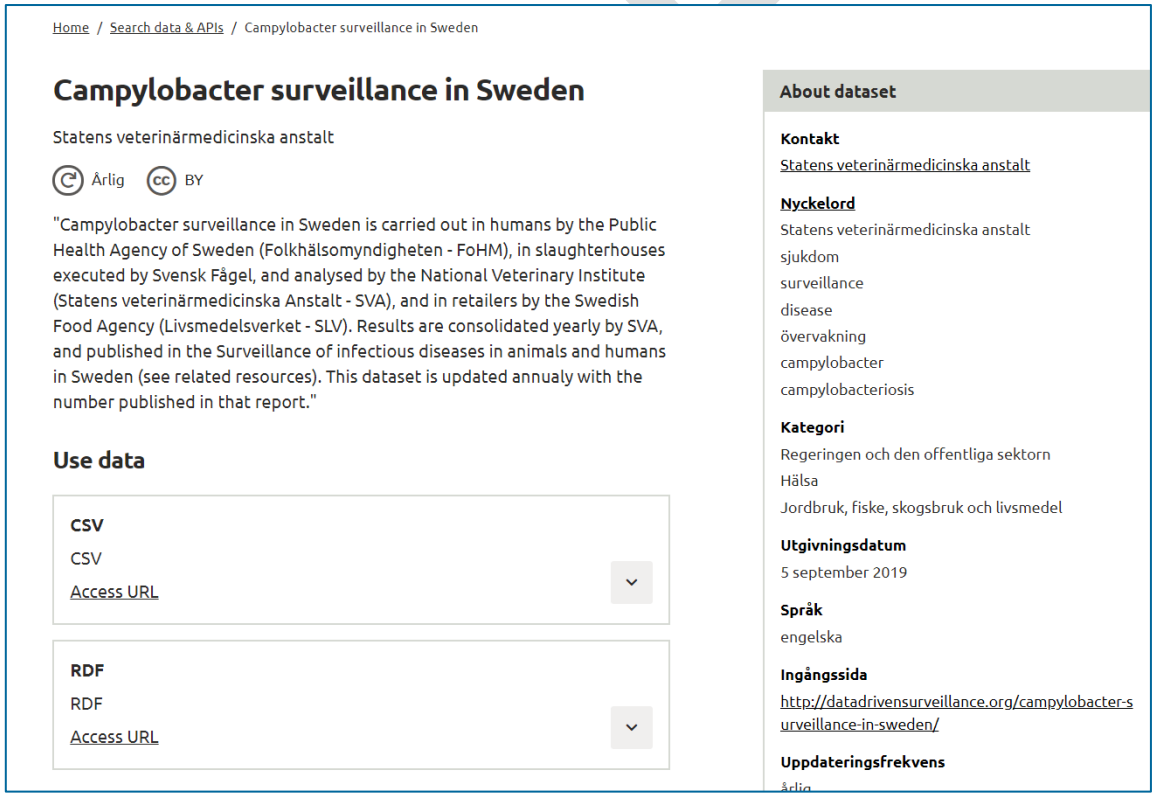

Figure 16. Page at the Swedish PSI data portal for the Campylobacter surveillance dataset.

In this case, we have an example where a dataset is made available with two different distributions: a CSV distribution, which a user can access through the direct link, and open in any software that handles spreadsheets, or any statistical analysis software; and the RDF distribution, meant for use by smart applications capable of handling semantically explicit, linked data.

All the metadata for this dataset, as well as its specific distributions, were also written using the workflows presented in this report. They can be downloaded through the dataportal.se, or directly

through the dataset URI: https://data.sva.se/dcat/surveillancereport/campylobacter_surveillance_sweden.rdf.

This FAIR dataset allows the results of campylobacter surveillance in Sweden, across all sectors, not only to be findable and reusable, but also available for both human and software agents. Having a machine-readable version of this dataset allows smart software applications to be built to consume these data. In this case, it is particularly important to have one permanent link to the dataset (https://data.sva.se/opendata/surveillance/campy/campylobacter_surveillance_sweden.rdf), and as new surveillance activities are carried out yearly, this dataset grows, allowing application which consume its data to be automatically updated. This is in contrast with yearly reports, for which a new link is produced every year.

DRAFT

This meeting is part of the European Joint Programme One Health EJP.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

SECTION 3: OPPORTUNITIES TO CONNECT DATA IN OHS

So far, we have discussed what semantic interoperability is, and how it can be operationalized using specific tools. We exemplified the use of these tools in practice with specific examples implemented in the ORION project.

But why is semantic interoperability relevant for One-Health Surveillance (OHS)? In this Section, we discuss the data interoperability needs in the OHS context specifically, as identified during the ORION project, and describe a vision for the future of OHS supported by semantic interoperability tools.

Early in the ORION project we have drawn the *surveillance pathway*, structuring our understanding of the steps involved in collecting, processing, and transforming data into information to support decision in disease prevention, control and eradication. This is depicted in Figure 17 as a linear sequence of steps, but we see this pathway as one iteration of a cycle that repeats itself in a continuous feedback loop, with results from one cycle informing design, adjustment and optimisation in the next cycle.

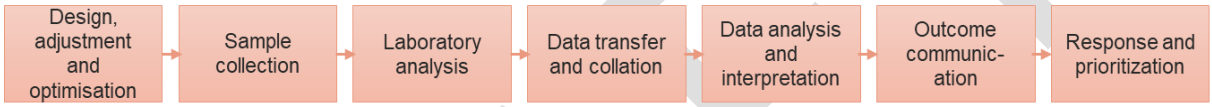

Figure 17. Surveillance pathway depiction used in ORION to anchor discussions throughout all work packages.

In a OHS context, if we consider the specific example of surveillance against foodborne zoonoses (FBZ), we can generalize this as represented by three parallel surveillance pathways, carried out within the animal health, public health and food safety sectors. This is schematically depicted in Figure 18.

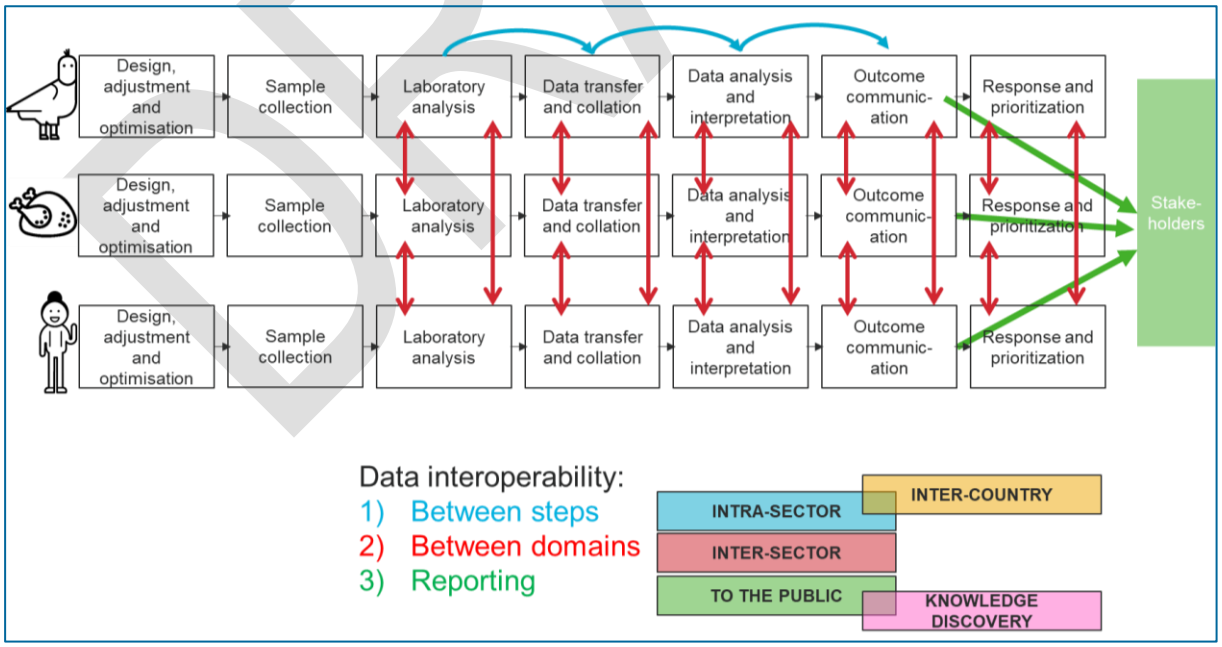

Figure 18. Data interoperability needs in One-Health Surveillance (OHS).

Figure 18 represents the data interoperability needs in OHS highlighting three main flows of data and information: 1) between steps in the surveillance pathway; 2) across health surveillance domains; and 3) when reporting results to stakeholders, including the general public. We will generally speak of "data interoperability", since we focus here on the development of computational tools. But we can also see

this as a issue of ***transparency***, where there is a need to explicitly annotate data with its context and inherent assumptions for reuse or for communication purposes.

Two additional levels of interoperability are identified in the bottom part of Figure 18: inter-country interoperability; and knowledge discovery. Addressing interoperability in these levels is not always a requirement to be able to carry out OHS activities within a region. However, they are an added benefit, allowing even more value to be generated from data. If the primary three levels of interoperability are addressed through semantic interoperability, then the knowledge model based on which interoperability is achieved can be extended (and even made available in several languages) to enable inter-country interoperability.

Knowledge discovery is the process of extracting useful knowledge from data[14]. The process requires a lot of data, and typically a knowledge model documenting assumptions and logic rules over which a machine can reason when analysing and validating the data. It is highlighted here in two contexts where interoperability would enable linking of large amounts of data to improve the power of knowledge discovery. This first is *accumulating evidence* on disease situation awareness to provide actionable information to decision makers in OHS. The second is enabling surveillance data to join the corpus of *research data*. Research data is increasingly made public and semantically explicit due to increased awareness among researchers of the power of data-driven discovery, and more and more even as requirement of funding agencies.

In the Introduction we have described the process of semantic interoperability as three steps: 1) creating a knowledge model; 2) annotating data with this model; and 3) consuming annotated data in smart (semantically aware) applications. The tools developed in ORION and introduced in Section 2 address specifically the two first steps. They are tools to enable the third step, where the full power of the linked data model is materialized. Tools for data consumption were not explicitly developed in ORION because they, ultimately, depend on the data workflows in use within institutions. It is the behaviour, structure and needs of the data consumers that will determine whether they can interact with RDF data.

In year 1 of ORION we carried out a requirement analysis reported in the deliverable JIP1-3.1[15]. The lessons learned from studying data workflows within institutions allowed us to conclude that earlier in the surveillance pathway, where data is mainly flowing within an institution or between institutions tasked with a specific surveillance activity, the data governance challenges are greater than technological challenges. While the application of the linked data model earlier in the pathway would potentialize its benefits, there would be little gain in demonstrating the benefits of the linked data model in steps where adoption is not likely to happen.

It was beyond the possibilities of ORION to impact data workflows and tools within institutions. These workflows were developed to maximize utility within the institution, and can understandably not be tailored to serve the purpose of interoperability with external actors. It was not our goal, for instance, to impose changes from the currently widely adopted relational database management systems (RDBMS), queried through Structured Query Language (SQL), into graphical databases or triple stores where data can be stored as RDF and queried through knowledge formulated queries expressed in SPARQL (SPARQL Protocol and RDF Query Language) or user-friendly graphically constructed queries.

Instead, we chose to focus on areas where data can already be shared, and therefore focused our pilot and tools on data for dissemination and reporting. That is, we focused our "linked-data" model in the specific case of "linked-open-data" (LOD). This is shown schematically in Figure 19.

[14] https://www.sciencedirect.com/topics/computer-science/knowledge-discovery
[15] https://zenodo.org/record/3754615#.YGwFH-gzY2w

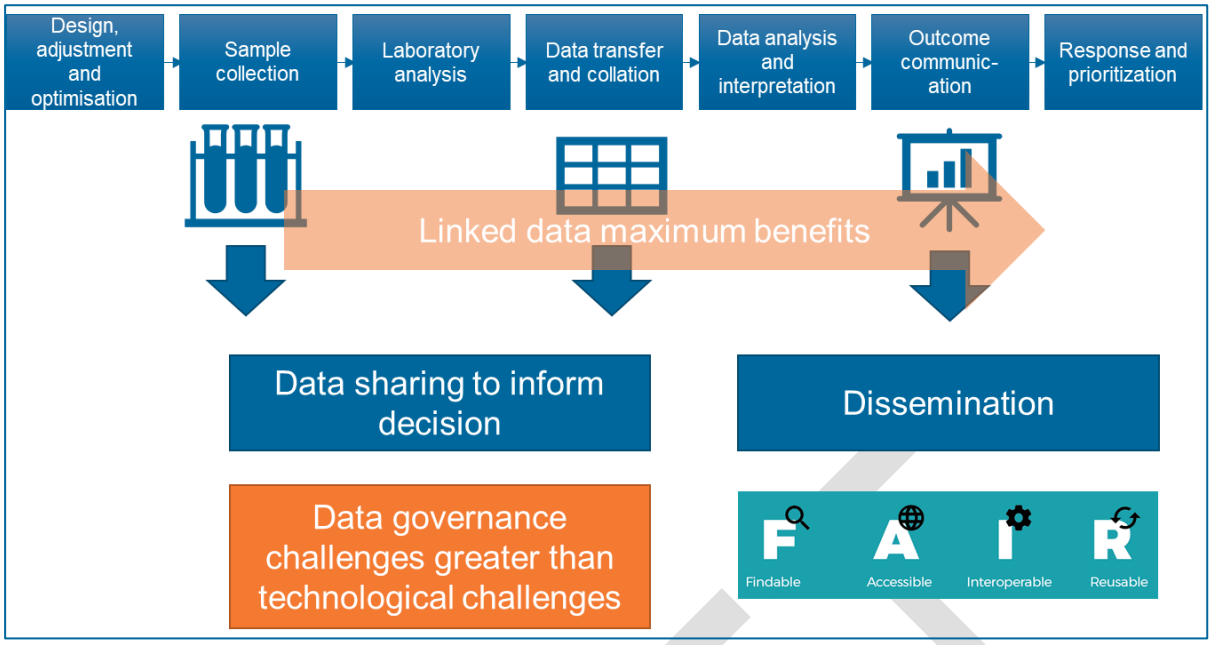

Figure 19. ORION focused on data dissemination, and the application of the FAIR framework to data already made public by institutions.

By choosing to focus on data that are already open, we hope to demonstrate the benefits of the linked-data model, and encourage its adoption trickling down from the end stages of the surveillance pathway, into earlier and earlier stages, enabling the OHS systems of the future.

In the OHS systems of the future, we do not foresee centralization of data. Instead, data owners are part of an "ecosystem of data". The ecosystem recognizes the data sources available and their content. When a surveillance actor needs information or evidence to support decisions, they pose queries to the ecosystem, and data that can contribute to answering the questions are shared on a "need-to-basis". *The ecosystem respects, rather than overcomes, existing governance restrictions,* and various models of data sharing ("code to data" versus "data to code", for instance) are in place to allow the query to be answered without centralization of data.

The goal of ORION was to support this vision of the OHS systems of the future as an ecosystem of interconnected data and tools in two main ways:

1) By allowing OHS to become a knowledge model that connects existing data, without imposing any (re)coding. Data are preserved in the context where they are created, and OHS is explicitly represented as a context of data usage, which can evolve in time.
2) By connecting data through a layer of semantic interoperability that does not require any centralization of data. Smart applications can be built to operate in this layer, creating the OHS ecosystem.

One-Health Surveillance as a knowledge model

Semantic interoperability is, as explained in earlier sections, based on the principle of separating data from knowledge. Interoperability achieved with the adoption of common terminologies or data standards results in the creation of a silo of coded data (Figure 20). This model has been used within health sectors, where the context of data generation and data usage is not expected to differ greatly, and data (re)coding is meant to solve syntactic (structural) differences, including language differences.

In OHS surveillance, however, the main barrier to overcome is preserving the context of data collection within each health domain when data are being (secondarily) used for inference in conjunction with data from other sectors, and differences exist not only in the data structure, but in the meaning (semantics).

DRAFT

The creation of the Health Surveillance Ontology (HSO), in particular its creation as a member of a large family of biomedical ontologies (The Open Biological and Biomedical Ontology (OBO) Foundry[16]), aimed to demonstrate how the context of data reuse can be captured in an explicit model separated from the data. Data is preserved in their original format, which is designed to best address the goals of a specific sector or institution, while all the transformation and links needed to connect those data to OHS questions are stored in a layer of interoperability. This is shown in Figure 20.

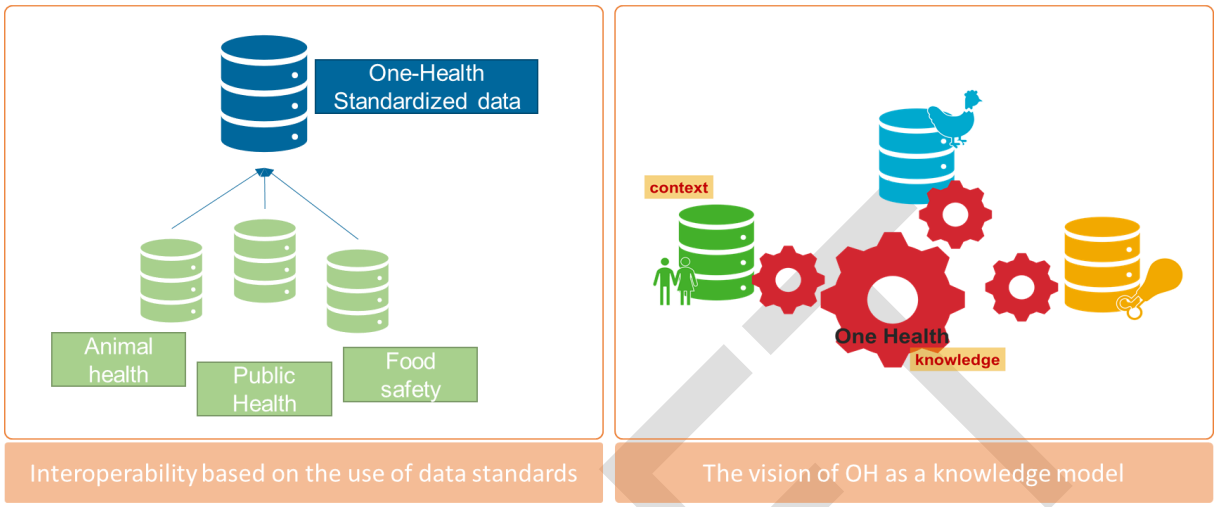

Figure 20. Two visions of OHS – based on structural interoperability, versus semantic interoperability.

In the knowledge model scenario, users – humans or smart applications – can query the data based on "knowledge questions", rather than data questions. Moreover, the model can evolve in time, without losing compatibility with past or future data.

HSO has the potential to be the building block for the "ecosystem of OHS solutions" depicted as cogwheels in Figure 20. This vision allows institutions to continue collecting, producing, storing and sharing data according to their current practices, as long as the knowledge model that connects all data sources and tools in the ecosystem is *maintained*. While ORION provided a proof of concept for the construction of such model, reaching the vision depicted in Figure 30 will depend on HSO being adopted by a community of users committed to maintaining the ontology, and curating ever evolving versions. Figure 21 details the responsibilities and timelines associated with this vision.

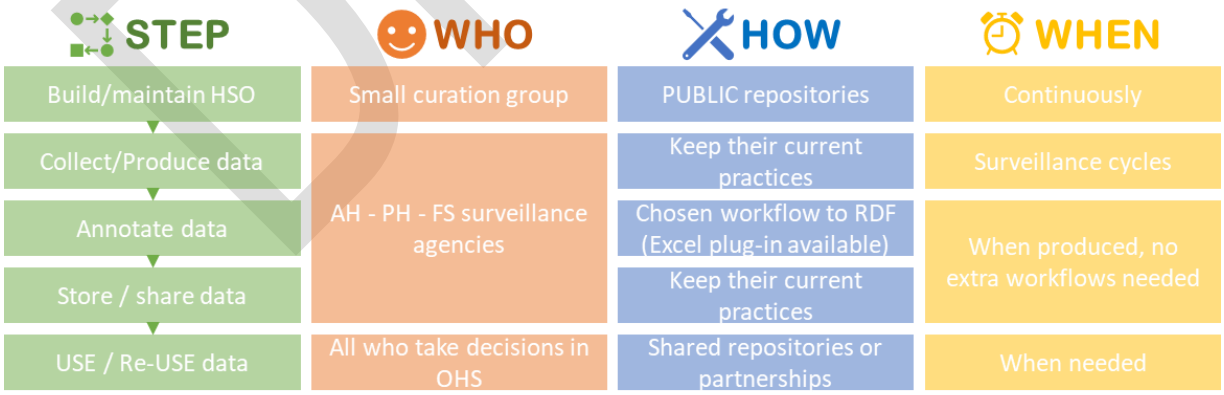

Figure 21. Responsibilities and timelines in a vision of semantic interoperability across health sectors supported by the Health Surveillance Ontology (HSO).

[16] http://www.obofoundry.org/

An ecosystem of connected data

The second fundamental building block of this vision of the OHS systems of the future is the ability of the ecosystem to connect data sources. Again, data would not be centralized, or even integrated, but the ecosystem would be aware of all existing data sources and their contents. If a user posts a query to the ecosystem, the ecosystem would be able to pose queries to the data source on a need-to basis, or, depending on the data governance rules in force, simply inform the user of the existence of a data source that can contribute to their question. The user would have access to information about the data owners and could contact them directly, but data sharing would not need to be guaranteed through the ecosystem.

Let's consider again the specific example of FBZ surveillance. The surveillance pathways involved were depicted in Figure 19 as three parallel pathways in order to highlight the connections between steps across sectors. In reality, these activities are not carried out at the same time in parallel. Surveillance activities would be carried out at different times and at different steps of the food production chain, generating sparse data as shown in Figure 22.

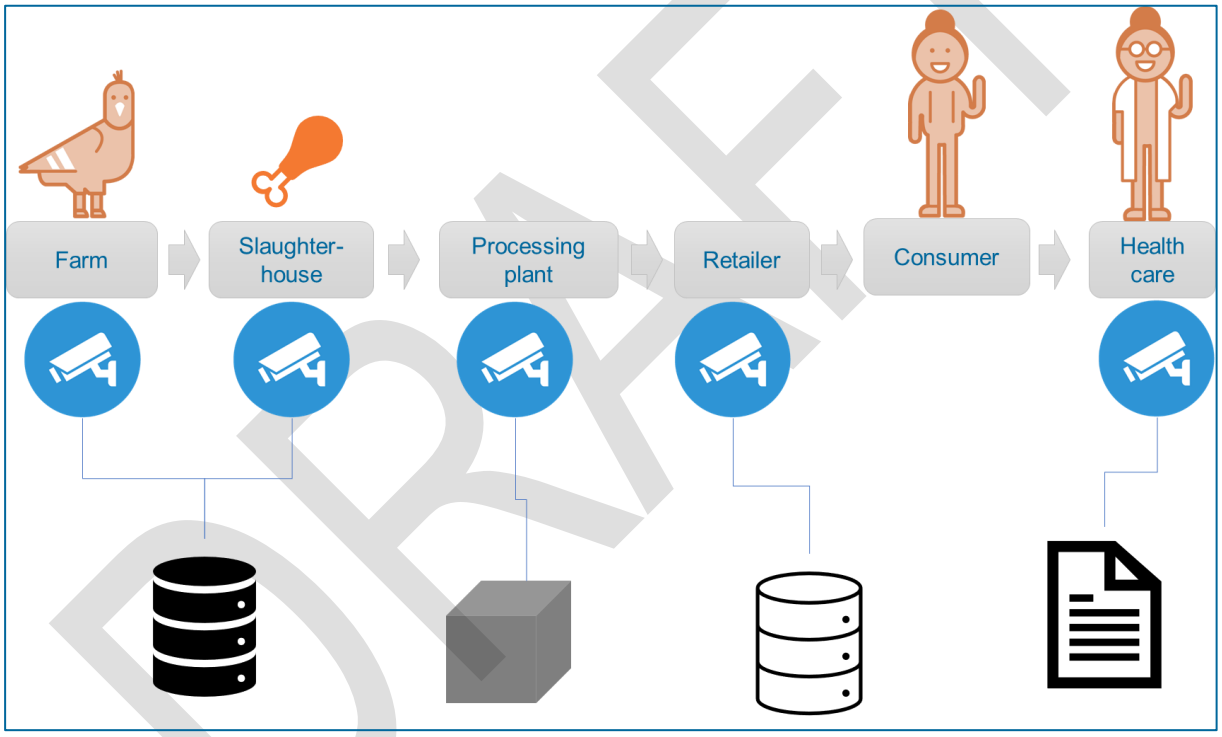

Figure 22. Surveillance in the different points of the food chain.

Currently, these data sources are stored under different formats, and from a technical integration perspective, they are "black boxes" which all other data sources cannot access. In Section 2, we exemplified how the tools developed in ORION can be used to annotate data and metadata, and publish FAIR data, or simply to annotate metadata, making data findable and reusable[17]. If all data owners published data in "FAIR" or "F__R" formats, the OHS ecosystem could easily be built as a "hub" that connects these data, as shown in Figure 23.

[17] Reusable as described in the FAIR principles presented in Section 2 does not mean public or open. Different desired levels of restrictions and permissions of usability are possible, but they should be expressed through specific licensing options explicitly stated in the metadata

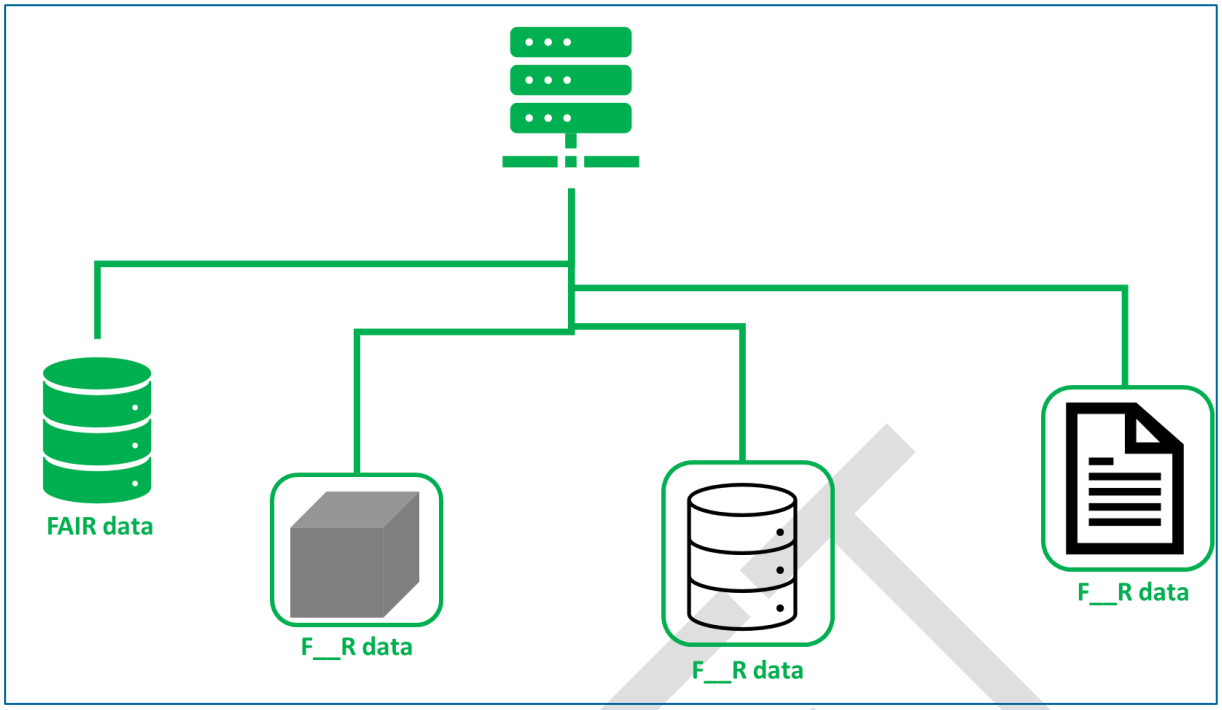

Figure 23. An ecosystem of connected data where not all data are FAIR, but all data are F__R.

In this report we have focused on HSO as a tool for data annotation, and demonstrated the use of tools that can attach HSO tags to both data and metadata. In Section 2 we have mentioned that ORION has also developed a One Health Consensus Report Annotation Checklist (OH-CRAC[18]) to promote OHS report harmonization. As we highlighted then, the specific steps in OH-CRAC are available in HSO as annotation properties. The properties can be used not only to annotate data, but also metadata.

In Section 2 we described the publication of the campylobacter surveillance in Sweden dataset as a fully FAIR resource, which URI is: https://data.sva.se/dcat/surveillancereport/campylobacter_surveillance_sweden.rdf

Applications pointed to this URI access the metadata of the dataset as RDF. In this RDF, CRAC annotations have been used to annotate the dataset. That is, CRAC has been used as a particular set of information to make this dataset findable by OHS applications.

DRAFT

[18] available in the OHS Codex
https://oh-surveillance-codex.readthedocs.io/en/latest/5-the-dissemination-principle.html

CONCLUSIONS AND LESSONS LEARNED: CONSIDERATIONS FOR A FAIR-ER OHS FUTURE

One current definition of One-Health surveillance (OHS) is "the systematic collection, validation, analysis, interpretation of data and dissemination of information collected on humans, animals and the environment to inform decision for more effective, evidence-and system-based health interventions[19]". Bordier et al, 2018[20] pointed out that other concurrent definitions of OHS all emphasize the role of cross-sectoral collaboration in the improvement of health management.

Barriers to data sharing are often listed when evaluating challenges to the establishment of such cross-sectoral frameworks. The challenge of extracting information from data coming from such heterogeneous contexts goes far beyond the simple access and aggregation of data. Ammon and Makela (2010)[21] described in detail the integrated collection and analysis of data on zoonoses in the European Union (EU), first established in 1992, and currently a joint task of the European Centre for Disease Prevention and Control (ECDC) and the European Food Safety Agency (EFSA). Despite this opportunity for joint data analysis, the authors pointed out several challenges to data comparability, from methodological differences between countries, and challenges of data quality and validation, to differences of population structure and population reporting level among sectors.

The experiences reported highlighted the complex data and meta-data structure needed to capture and take into account all the contextual information about the data collected and the data collection processes. As Beauté and colleagues (2020)[22] more recently pointed out, an accurate description of structural elements of surveillance systems is essential for interpretation and evaluation, but even when these descriptions exist, confusion can remain on their interpretation.

Within individual countries, activities of surveillance in public health (PH), animal health (AH) and food safety (FS) all generate data which can contribute to OHS. Converting those data into valuable information for decision requires not necessarily that those data are aggregated, but that they are interoperable, so that sharing can be performed on demand, for specific problems, respecting various models of data disclosure. Interoperability focuses on cooperation among systems, referring to their ability to continuously communicate and exchange information, and use the information that has been exchanged[23].

In this deliverable, we have described ORION's contributions to data interoperability in One Health Surveillance (OHS). We have focused on the specific case of sharing surveillance outputs across sectors to enable joint surveillance evaluation, prioritisation and design. Our developed tools are therefore most applicable to the case of publishing reports and data at the end of specific surveillance cycles, such as for instance yearly reports published individually by countries and, in the specific case of the European Union, jointly by EFSA and ECDC.

As we have centred tool development around the development of a knowledge model for surveillance, the achievements in terms of data and meta-data structuring, and harmonisation and disambiguation through semantic expression are achievements on themselves, which are preserved for future use in the form of a publicly available ontology – the Health Surveillance Ontology (HSO). As shown in Figure

[19] Stärk KDC, Arroyo Kuribreña M, Dauphin G, Vokaty S, Ward MP, Wieland B, et al. One Health surveillance - More than a buzz word? Prev Vet Med. 2015;120(1):124–30.
[20] Bordier M, Uea-Anuwong T, Binot A, Hendrikx P, Goutard FL. Characteristics of One Health surveillance systems: A systematic literature review. Prev Vet Med [Internet]. 2018;(October):0–1. Available from: https://doi.org/10.1016/j.prevetmed.2018.10.005
[21] Ammon A, Makela P. Integrated data collection on zoonoses in the European Union, from animals to humans, and the analyses of the data. Int J Food Microbiol [Internet]. 2010;139(SUPPL. 1):S43–7. Available from: http://dx.doi.org/10.1016/j.ijfoodmicro.2010.03.002
[22] Beauté, J., Ciancio, B.C., Panagiotopoulos, T., 2020. Infectious disease surveillance system descriptors: proposal for a comprehensive set. Euro Surveill. 25. https://doi.org/10.2807/1560-7917.ES.2020.25.27.1900708
[23] Definition of Interoperability. In: HIMSS Dictionary of Healthcare Information Technology Terms, Acronyms and Organizations. 2nd editio. 2010. p. 190

DRAFT

21 (Section 3), the development of an ontology is not a "one-time" process, and ontology maintenance will require a community of users committed to keeping the knowledge model evolving. One of the main advantages of a knowledge model is its ability to adapt to new knowledge, but this of course requires the work of dedicated curators. We have, during the lifetime of ORION, succeed in making HSO a member of the Open Biological and Biomedical Ontology (OBO) Foundry to ensure that it can be part of a larger community of ontology developers and users.

A FAIR OHS future

The FAIR principles have been discussed in Section 2. The FAIR principles were also highlighted in EFSA's technical report "Publication of scientific data from EU-coordinated monitoring programmes and surveys" in 2019[24].

The Health Surveillance Ontology (HSO) is a FAIR model, which therefore enables annotation of data – and meta-data – in fulfilment of the interoperability principle. To implement data annotation, the existence and public availability of HSO is of course not enough. Data owners must have access to data *annotation* workflows. While these workflows can only be implemented by individual institutions, and a "one workflow fits all" does not exist, the ORION project has provided a number of proof-of-concept workflows using country specific and EFSA and ECDC publicly available data, as extensively described in Section 1 and Section 2.

While we have focused on publicly available data, it is important to note that being "open or publicly available" is not a requirement for data to be FAIR. FAIR data is findable by those who must find it, and accessible to the software agents who will process it. In turn, "publicly" available data which does not have an explicitly declared license, is not reusable, and therefore not FAIR.

The European Food Safety Authority Advisory Forum Task Force on Data Collection and Data Modelling delivered their conclusions[25] on September 2020 focused on four key priority areas: data collection and reporting processes, data models, IT infrastructure, and data analysis. The implementation of data interoperability through semantic annotation enables (and in fact provides the foundation for) many of the recommendations made by the task force for a future ideal EU food safety system, including a specific recommendation to initiate and promote the development of ontologies.

In the introduction, we have presented the linked data model as enabling knowledge to be stored in applications, so that data in consumed by the right applications on demand. This preserves data context while at the same time making it evolvable and sharable without the need for data coding or centralization. This is in agreement with the task force's recommendation 4.5 to develop IT architectures centred in *building services, not websites*. The key element of a future food-safety system specifically, and OHS generally, is to avoid data centralization, and indeed even data analysis centralization, moving towards an "***ecosystem of solutions***". In this ecosystem, data providers do not transfer their data, rather choose to make the ecosystem aware of their existence, content and structure. Data analysis tools are also developed independently by many actors who choose to add their tools to the ecosystem. Data analysis tools and actual data meet on demand, and data access is negotiated for purpose. This view of an ecosystem meets yet another important recommendation from the task force – (2.3) "Good practice aimed at data interoperability should be sought through collaborative data governance".

A FAIR-ER future

Surveillance is a cyclical process, that happens in a continuous feedback loop. Surveillance execution happens in cycles, usually annually, and the surveillance pathway introduced in Section 3 is a good representation of the activities within one cycle. However, we should not lose sight of the fact that the main reason to share surveillance outputs across sectors is for surveillance evaluation and action, in particular for the redesign and prioritisation of activities in the next cycle.

[24] doi:10.2903/sp.efsa.2019.EN-1544
[25] doi:10.2903/sp.efsa.2020.EN-1901

In Section 2 we have demonstrated the creation – and FAIR publishing – of a dataset of surveillance outputs for one specific system (Campylobacter surveillance in Sweden), but across several years. As opposed to the annual surveillance report system, which contains results for all surveillance systems in the country but only for one year, the Campylobacter dataset is "***Extendable***".

In this view of an ecosystem of solutions, the Campylobacter dataset would be a resource in the ecosystem with a fixed resource identification. As more results become available – and indeed even as data may be updated or corrected from previous years – applications that use those data can always be sure to be accessing the most up to date data available. Surveillance methods themselves can be changed, but these changes are also documented in the extended data. Extendable datasets are how we see data within institutions, but the idea is not yet built into the cycles of surveillance data publishing.

A last idea we would like to add to the FAIR principles is the idea of "***Reproducibility***", which is also directly related to "***Transparency***". In an ecosystem of solutions, data analysis is performed within the ecosystem using codes that are themselves part of it. Any user is free to reproduce the analyses, and more importantly, check the inherent assumptions made within.

When we publish data in yearly cycles as independent resources, in aggregated formats, and without enough surveillance descriptors to track the surveillance methods that generated those results, we lose reusability even within the system, as compatibility with historical and future values becomes fraught.

Developers of data analytical methods are used to the concepts of reproducibility and versioning, but we miss these same concepts in the way we store surveillance methods and results. ORION has provided two key resources to advance solutions that make surveillance data inherently associated with its methodological context: the One Health Consensus Report Annotation Checklist (OH-CRAC[26]), and the Health Surveillance Ontology.

Moving towards a FAIR-ER future we suggest that OHS actors in general, and data providers in particular, investigate workflows that allow them to capture all methodological details in force in a given surveillance system and year, as established in OH-CRAC. If these data are to be shared, we then recommend their semantic annotation with HSO and following the FAIR principles, in particular making sure that data have enough meta-data to be findable, and proper licensing to be reusable. If FAIR data publishing is achievable, then we urge data providers to think about whether unique resource identifiers can be given to specific datasets individually but perpetually, so that these datasets are Extendable through the years, rather than continuously replaced by a new yearly batch. Lastly, we suggest versioning any changes to both the dataset and its metadata, and ensuring that assumptions and analyses performed to aggregate raw data into the published data are reproducible.

ORION's contribution to this future have been documented in this deliverable finalized in June 2021, but the ORION tools will live through its community in the links provided throughout this deliverable.

We highlight in particular the One-Health Surveillance Codex for a list of all ORION resources.

DRAFT

26 https://oh-surveillance-codex.readthedocs.io/en/latest/5-the-dissemination-principle.html#one-health-consensus-report-annotation-checklist-oh-crac