Health Surveillance Ontology: supporting semantic interoperability in One-Health

Fernanda Dórea[a], Estibaliz Lopez de Abechuco Garrido[b], Nazareno Scaccia[b], Matthias Filter[b]
[a] *Swedish National Veterinary Institute, Department of Disease Control and Epidemiology, Sweden*
[b] *German Federal Institute for Risk Assessment (BfR), Department 4—Biological Safety, Germany*

Abstract. The Health Surveillance Ontology (HSO) aims to enable semantic interoperability among health surveillance sectors – public health, animal health and food safety. HSO is being developed under the OBO Foundry guidelines in order to reuse knowledge from these ontologies, while letting HSO represent a high-level knowledge graph that connects the concepts needed to model and semantically integrate high level, aggregated OHS data. This will also promote interoperability between surveillance data and the corpus of biomedical data already made available as linked-open-data (LOD).

Keywords. One-Health, surveillance, public health, animal health, food safety

1. Introduction

Health surveillance is a multidisciplinary task that relies on extensive re-use of data originally collected for different primary purposes [1]. The challenge of data reuse increases as we evolve surveillance towards a One-Health perspective, where health risk management decisions rely on information collected from animals (and food), humans and the environment [2]. Animal health and food safety surveillance, in particular, rely on data collection along many different steps of the "farm to fork" continuum. One-Health surveillance (OHS) therefore has high demands for information sharing architectures that are "dynamic, distributed, and context-aware" [1].

In this context, semantic technologies become the obvious choice to allow preservation of data context and repurpose of information, while enabling automation and knowledge discovery by computers. Semantic interoperability is particularly important in OHS in order to allow data reuse across sectors, and reuse of data for research and knowledge discovery, while preserving the original context of the data.

The Health Surveillance Ontology (HSO) aims to support semantic interoperability betweem the One-Health surveillance sectors "public health", "animal health" and "food safety". HSO - in its current status - represents a knowledge graph connecting the concepts needed to model and semantically integrate high level, aggregated OHS data.

2. Modeling surveillance processes

A surveillance *system* is "a range of surveillance components (and the associated organizational structures) used to investigate a single hazard in a specified population" [3]. A surveillance *component* is, in turn, "a single surveillance activity used to investigate one or more hazards in a specified population" [3]. Following these definitions, we identified a *surveillance activity* as the smallest unit of a surveillance

system, and made it the central concept in HSO. Our modeling focuses on defining how the methodological details of each individual surveillance activity can be captured, so that the data linked to a surveillance activity or system are as context explicit as possible, making them transparent and interoperable across One-Health agencies, even from different sectors.

Surveillance systems are a public good – they are usually carried out by governmental agencies and the results are generally made public in yearly cycles. In the European Union (EU), integrated collection and analysis of data on zoonoses among member states (MS) was first established in 1992 [4], and is currently a joint task of the European Centre for Disease Prevention and Control (ECDC) and the European Food Safety Agency (EFSA). Besides data reported from MS[1], EFSA is also responsible for publishing and maintaining a harmonized terminology[2] used as the standard for data reporting by the MS. The Standard Sample Description (SSD1) provides standards for the description of samples and analytical results. Its extension (SSD2) covers also zoonotic agents in food and animals, antimicrobial resistance and food additives. Similar harmonized terminologies exists from ECDC on human disease surveillance systems. Data reported to the ECDC is made publicly available in the ECDC Atlas of Infectious Diseases[3].

The large body of historical, publicly available data provided by currently established systems is not available in semantically aware machine readable formats. Nevertheless, data and reports constitute a rich source of domain knowledge that can be used to extract concepts, and build the semantic models that will enable capturing, in the future, surveillance contextual information in machine readable and interoperable formats. Surveillance data is also generally reported in very high levels of aggregation, such as total number of cases in humans, total number of chicken slaughter batches tested, and total number of those that were positive for *Campylobacter spp*. Currently, this level of aggregation is frequently the only option to share surveillance data across sectors due to legal or data protection constraints. The primary focus for HSO development was therefore the modeling and annotation of these types of aggregated data.

For example, in the "European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks" [5], in 2017 a total of 10,608 cases of campylobacteriosis were reported in humans in Sweden, translated to an estimated rate of 106.1 cases per 100,000 habitants. The total number of reported cases in Italy was much lower, 1,060. This latter number was not converted to a population rate due to the lack of national coverage. In the methods section of the report, authors caution against comparing data between countries, due to limitation of data quality and differences in surveillance systems. These differences, however, are not captured or described. Capturing all the contextual information needed about the data collected and the data collection process requires complex structure of data and meta-data, which cannot be achieved in current tabular reporting formats. An initiative to capture such high-level meta-information is the Consensus Report Annotation Checklist (CRAC), that is currently developed under the ORION project[4].

HSO aims to provide an overarching knowledge model that is capable of describing methodological and contextual details ranging from a single surveillance activity to whole surveillance systems. This also means that HSO does not focus on modelling the "raw data" collected as part of a surveillance activity. For such kind of data ontologies already exist, e.g. to support the annotation of specimen and disease cases data. To some extent HSO can be interpreted as an Ontology for meta-data. For instance, for the 10,608 cases of campylobacteriosis reported in Sweden in 2017 – HSO would support the provisioning of information on the reporting practices in Sweden, the case definition

[1] https://zenodo.org/record/3527706#.Xu8eomgzYkX
[2] https://zenodo.org/record/344473#.Xu8enWgzYkX
[3] https://www.ecdc.europa.eu/en/surveillance-atlas-infectious-diseases
[4] https://oh-surveillance-codex.readthedocs.io/en/latest/

used that year, the population coverage and the expected proportion of cases that get reported, etc.

2.1. Ontological structure

Seeking maximum interoperability with OBO Foundry compliant ontologies (http://www.obofoundry.org/)[6], HSO uses the Basic Formal Ontology (BFO)[7] as upper level ontology. Following BFO's main classification of concepts, a surveillance activity is a process. In keeping with the OBO Foundry goal of orthogonality, we reviewed existing ontologies to identify processes that are closely related to surveillance. It was concluded that a surveillance activity is just one specific type of biomedical investigation, and therefore we sought, as much as possible, to align with the Ontology for Biomedical Investigations (OBI)[8]. A Surveillance activity was therefore defined as a sub-class of a "planned process" (http://purl.obolibrary.org/obo/OBI_0000011).

2.2. Data-driven, bottom up collection of concepts

The identification of concepts that need to be modelled in HSO has been data-driven. We have identified, through the ORION project[5], a number of reports and datasets that are often shared across health surveillance sectors. In particular, the reports for mandatory surveillance activities of MS and EFSA/ECDC. These reports and data examples, whether currently collected as spreadsheets or simply as textual descriptions, were scrutinized in an iterative process to create an inventory of concepts that needed to exist in the HSO knowledge model. Every new report or data source generated a new list of concepts, and the following steps were taken iteratively to place them in HSO.

2.3. Top-down modelling

Once new concepts are available, other ontologies in the OBO Foundry are reviewed to identify whether the concepts already exists. If a concept exists, it is reused. Where concepts do not yet exist in other ontologies, we review BFO and related ontologies to identify where these concepts should be placed so that they are aligned with the general modeling choices of other biomedical ontologies.

In OBI, for instance, all planned processes are expected to have a *plan specification*, composed of *action* and *objective* specifications. We have created a *surveillance protocol* as a sub-class of a plan specification. A surveillance protocol is expected to have one ore more objective specifications, and the following specific types of objectives specifications were created: *surveillance objective*, *surveillance purpose*, and *surveillance context*.

OBI and the Genetic Epidemiology Ontology (GenEpiO)[9] have been the main sources of content reuse and modeling advice in this process.

2.4. Mapping terminologies

As mentioned above, the reports and data available from European MS are a great source of OHS concepts. We envision that all data reported by MS on foodborne disease occurrence, could become interoperable through HSO in the future. With this goal, we are progressively reviewing also data reporting models, and including their concepts in HSO. For all concepts incorporated into HSO for which an SSD2 catalogue is available, the full catalogue is imported into HSO as instances of the relevant classes. This will allow automated annotation of all data already coded with these terminology catalogues. Moreover, the ontology is being used to "detangle" terminology entries that represent a mix of multiple concepts. For instance, in the SSD2 catalogue MTX, term AOC78 refers

[5] https://onehealthejp.eu/jip-orion/

to a "*Gallus gallus* broiler". *Gallus gallus* is the animal species, while broiler is the production type (a chicken raised for meat production)..

The choice to keep entries of these terminologies as instances, not classes, was made to make it explicit that these terminologies represent only "one version of the truth". The instance "outbreak investigation" with URI explicitly marking its origin (SSD2 catalogue PRGTYP, term code K032A) represents the SSD2 definition of an outbreak investigation. Should domain experts deem that outbreak investigation is a relevant sub-class to have under "surveillance context", this class can be used to gather instances of other terminologies, and named using terminology that is considered easier to understand by all OHS sectors.

3. HSO documentation and resources

OWL codes for HSO are available in BioPortal (https://bioportal.bioontology.org/ontologies/HSO) and at the perma-address https://w3id.org/hso. This same address can be used in a browser to access a page with all ontology documentation. Figure 1 gives an overview of the contextual data that can be currently associated with some simple surveillance results.

4. Conclusions and next steps

We continue reviewing OHS data examples, and expanding the knowledge model. We are also implementing workflows to allow both datasets and metadata to be annotated using simple, Excel based workflows. Excel to RDF conversion is enabled by a plugin developed in connection to our research (https://karlhammar.com/ExcelRDF/).

In our effort to capture as much contextual information about surveillance activities as possible, we integrate the metadata concepts from CRAC into HSO. CRAC is based on the Generic Statistical Business Process Model (GSBPM) that was developed by official statistical institutions to support their data and metadata harmonization. As the reporting of surveillance metadata can vary greatly, in particular in historical reports, CRAC annotation properties will allow capturing of any meta-information as free-text.

Together with other resources in the One-Health Codex, HSO and CRAC aim to maximize the One-Health potential of available surveillance data, increasing usability and interoperability across health sectors.

References

[1] Mirhaji, P. Public Health Surveillance Meets Translational Informatics: A Desiderata. J. Lab. Autom. 2009 14, 157–170.

[2] Stärk, KDC, et al. One Health surveillance - More than a buzz word? Prev. Vet. Med. 2015 120, 124–130.

[3] Hoinville LJJ, Alban L, Drewe JAA, et al. Proposed terms and concepts for describing and evaluating animal-health surveillance systems. Prev Vet Med. 2013 112: 1–12.

[4] 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. 2010 139, S43–S47.

[5] EFSA and ECDC (European Food Safety Authority and European Centre for Disease Prevention and Control). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food‐borne outbreaks in 2017. EFSA Journal. 2018 16(12):5500

[6] Smith, B, Ashburner, M, Rosse, C et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 25, 1251–1255

[7] Smith B, Kumar A, Bittner T. Basic Formal Ontology for Bioinformatics. IFOMIS Reports. 2005.

[8] Bandrowski A, Brinkman R, Brochhausen M, et al. The Ontology for Biomedical Investigations, PLoS One. 2016 Apr 29;11(4):e0154556.

[9] Griffiths E, Dooley D, Graham M, et al. Context Is Everything: Harmonization of Critical Food Microbiology Descriptors and Metadata for Improved Food Safety and Surveillance. Front. Microbiol., 2017.