A domain-ontology and algorithms for the reuse of transmission data

-making reanalyses & federated analysis easy and transparent -

Slavco E.[1], De Vos M [1] , Hostens M [1] , Top J [2] , Bootsma M.C.J. [3] , Hobbelen P. [4] , Pacholewicz E[4] , Fischer E.A.J.[1*]

1 Universiteit Utrecht, The Netherlands *e.a.j.fischer@uu.nl 2 WFBR, Wageningen, The Netherlands 3 UMCU, Utrecht, The Netherlands 4 WBVR, Lelystad, The Netherlands

Problem statement Transmission parameters, such as the effective contact and R_0, give valuable information to understand and mitigate infectious diseases. Data to estimate these parameters is often undisclosed and non-accessible for reanalysis with new method or for meta-analysis.

Federated data analysis allows for reuse of data without transferring data. Federated data analysis as well as efficient reanalysis, however, requires standardization of data formats.

house_pen	animalnr_col	weight_d0	weight_d21	BS1	BS1	swab1	swab1
H1_S1	3_Ge	35.72	732	-*		++	Acinetabacter pitti

isolator	animalnr	weightd0	weightd21	d1_count_ESBL	d2_count_ESBL
138	138	45	883.93	10	1000

Isolator	ID	D5 9:00	D5 16:00	D6 8:00	D6 16:00	D7 8:00
1	101	0	0	1	1	1

Figure 1 "Familiar situation": Three data sets with the same type of data but different headings (e.g. house_pen or isolator to indicate the location) and values (e.g. ++ or 1 to indicate positive animal). Some of the headers include information on the values (e.g. *d0 to indicate the day of sampling is day 0).

Solution A domain ontology allows for accessing data by a standard algorithm producing aggregate measure easy and safely shared and combined with other of such measures.

Key aspects:

Sampling

- Measurement is taken from a Host or an Environment.
- Concepts from Ontology of units of Measurement are re-used
- Sample has a categorical result (positive / negative) or a value (bacterial count)
- Time of measurement can be date-time specified or time since inoculation

Location of hosts

- Host is located in an Environment
- Environment allows 3 levels of mixing, as well as the location in space

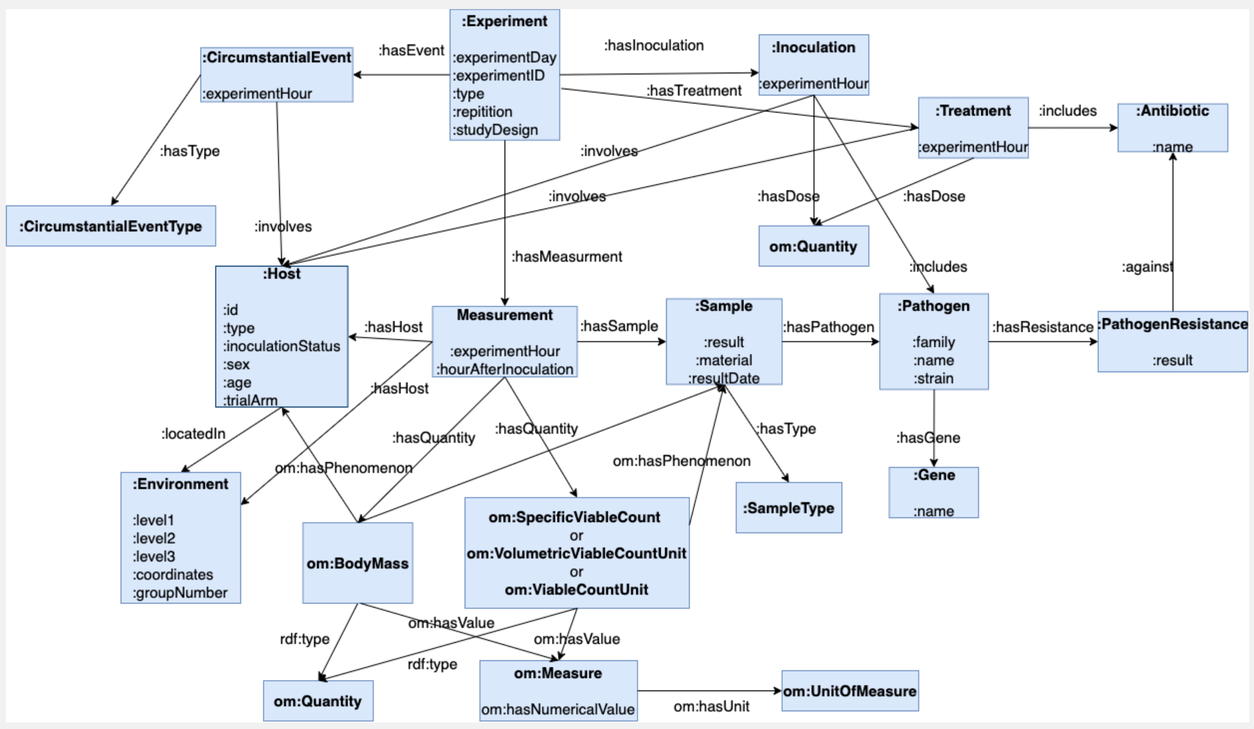

Figure 2 Schematic overview of the domain ontology. An ontology is a set of concepts with properties and the relationships between concepts. Prefix om: correspond to the Ontology of units of Measure with URI http://www.ontology-of-units-of-measure.org/resource/om-2/.

Implementation A workflow including pre-processing of data, mapping data to the ontology, local estimation and combining of aggregated output.

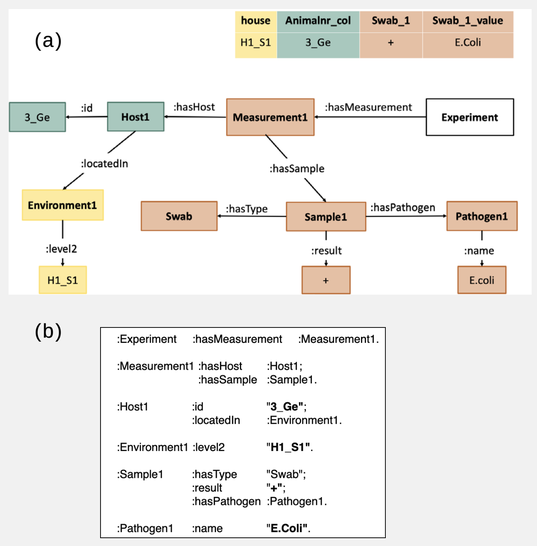

Figure 3 Example of mapped dataset
(a) Data owner pre-processes data to make variables names unambiguous and without hidden information (such as time of sample). Data owner maps data by defining the variable names in data set corresponding to ontology classes. Data owner can leave sensitive information out of the mapping such that these can never be accessed.
(b) Annotated dataset consists of multiple triples (subject ->predicate->object). This figure represents the triples that correspond to the example in Figure 3(a).

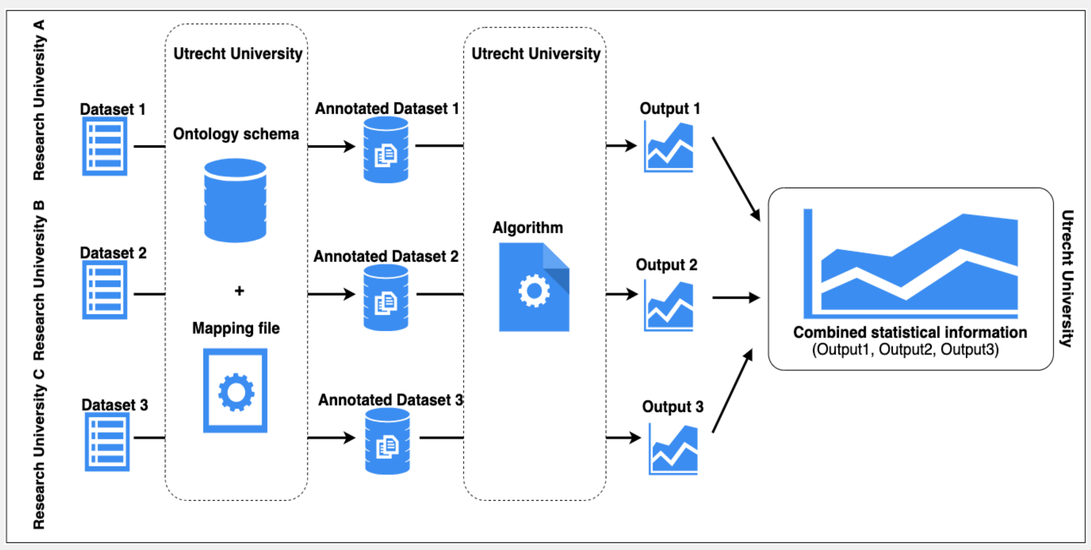

Figure 4 Workflow of the SUMMERFAIR project. Data at different locations (Research Universities) is mapped to the ontology using the mapping file, generic algorithm is run at each location on the ontology and only aggregated output is sent to the central location (Utrecht University) at which aggregated outputs are merged (e.g. by a meta-analysis of aggregated data).

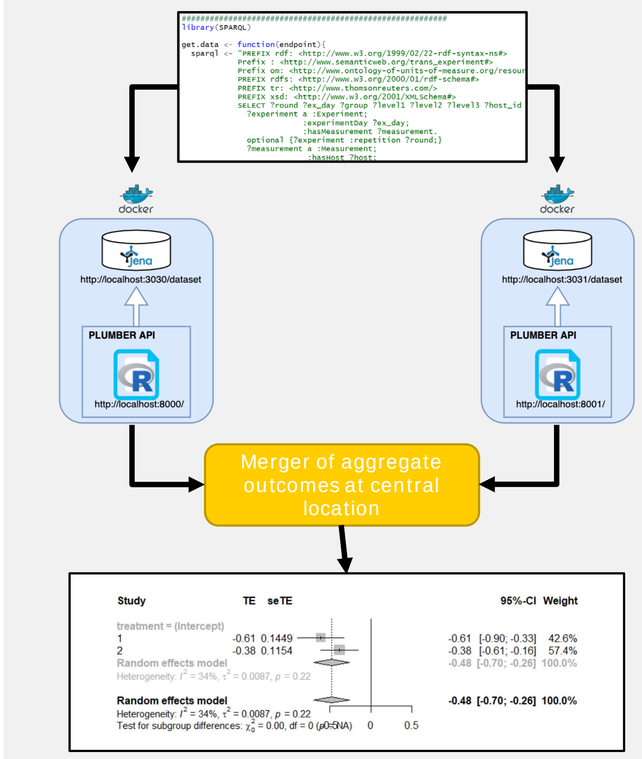

Figure 5 One generic R-script sent to two locations. Aggregated results sent to central location and merged (here by meta-analysis).

More information and contact details:

SCAN ME

The SUMMERFAIR project is financially supported by

ZonMw

uu.nl/veterinarymedicine

One Ontology, One Algorithm to rule them all.