SummerFAIR schema proposal (V0.3)

AUTHOR
Stefano Rapisarda (RDM)

PUBLISHED
May 29, 2024

1 Implementation list

From Egil's feedback and discussion:

- [x] Propose an alternative database version organised in three files concerning animals, housing/environment, and measurements, respectively;
- [x] Considers general animals (or hosts) instead of just broilers;
- [x] Take into account housing information (or, in general, location);
- [x] Improve readability of column information;
- [x] Separate schema-related metadata and general metadata. Researchers are not supposed to edit schema-related metadata;
- [x] Update name and properties according to the Infection Transmission Ontology;
- [x] Move the description of the column names from the schemas to the general metadata;
- [ ] Provide practical examples of data stored in the proposed schema (not yet for this specific version).

2 Updates

Most of the updates follow convention used in the Infection Transmission Ontology.

- When possible, column names have been codified in the form class_property;
- Fixed format acronyms are removed. Those fields have now general "string" type and not specific format;
- The interventions table has now be changed to events table. Events can either be measurements, inoculations, or treatment. For each event, researchers can specify type, relative quantity (measure, inoculated, or administrated for treatment) and corresponding unit;
- Dates (host age, host death, etc) are now measured in days from the beginning of the experiment;
- Even if not mentioned as data property in the ontology, I used "day" instead of "date" when the time is measured as day after the beginning of the experiment. This is to avoid any possible confusion;
- The column describing the cause of death has been removed. I will keep as reference the work done in the past, if more specific records are needed for the experiment, researchers can follow instructions about how to integrate new measurements in this specific data schema;
- The term "animal" has been changed to "host" according to the "Infection Transmission Ontology";
- Modified "birth_date" into "host_age", where age is recorded at the experiment start date;

- Operator column removed;
- Inoculation method column removed. I think the information is already specified in incoulation type;
- Housing has been changed to environment;
- Environment levels have been introduced (instead of environment types);

3 .csv file columns

3.1 Hosts

name	description	type	format		
name	**description**	**type**	**format**	**values**	**unit**
host_id	Host unique identifier	string	AA0_00000	NaN	NaN
host_groupNumber	Integer indicating the group of the host	integer	NaN	NaN	NaN
host_sex	Sex of the host	string	A	[M, F]	NaN
host_age	Age of the host at the beginning of the experiment.	integer	NaN	NaN	day
host_death	Date of death of the host from the beginning of the experiment.	integer	NaN	NaN	day
host_species	Species of the host	string	NaN	NaN	NaN
host_breed	Breed of the host within the species	string	NaN	NaN	NaN

3.2 Events

name	description	type	format		
host_id	Host unique identifier	string	AA0_00000	NaN	NaN
event_day	Date of the event in days after the beginning of the experiment	integer	NaN	day	NaN
event_time	Time of the event_time (local time)	string	HH:MM	NaN	NaN
event_type	Type of event	string	NaN	NaN	[measurement, inoculation, treatment]
measurement_type	Measurement type	string	NaN	NaN	NaN
measurement_quantity	Measured quantity	float	NaN	NaN	NaN
measurement_unit	Measured quantity unit	string	NaN	NaN	NaN
inoculation_type	Inoculation type	string	NaN	NaN	NaN
inoculation_pathogen	Inoculation pathogen	string	NaN	NaN	NaN
inoculation_dose	Amount of inoculated pathogen	float	NaN	NaN	NaN
inoculation_unit	Inoculation dose unit	string	NaN	NaN	NaN
treatment_type	Treatment type	string	NaN	NaN	NaN

name	description	type	format		
treatment_dose	Amount of a substance administrated in the treatment	float	NaN	NaN	NaN
treatment_unit	Treatment dose unit	string	NaN	NaN	NaN

3.3 Environment

name	description	type	format		
host_id	Host unique identifier	string	AA0_00000	NaN	NaN
allocation_day	Date of allocation of the host in the environment measured from the beginningthe experiment	integer	NaN	day	NaN
allocation_time	Time of the allocation of the animal in the housing (local time).	string	HH:MM	NaN	NaN
environment_level	Environment level	integer	NaN	NaN	[1, 2, 3]
environment_id	Environment unique identifier	string	A0_A0	NaN	NaN

4 File description and usage:

When starting an experiment, researchers are supposed to create a new directory with the name of the experiment with the following structure:

```
experiment1
├── metadata.toml
├── _validation_schema.toml
├── data
│   ├── environment.csv
│   ├── events.csv
│   └── hosts.csv
```

4.1 Metadata description

- The **metadata.toml** file contains keywords describing general information about the experiment and, for each file, column name, description, and unit;
- The **_validation_schema.toml** file contains column names, formats, values, etc. This file is not supposed to be edited by researchers (unless they need to modify the data schema)

4.2 File description

- The **hosts.csv** file is supposed to be edited when host information is collected and when hosts die;
- The **events.csv** file is supposed to be edited every time an event occur. An event can be either a measurement, an inoculation, or a treatment. For each event, the operator has to specify host id, event day and time, and all the information related to that specific event, leaving empty the fields related to the other two;

- The **environment.csv** file is supposed to be edited every time a host is located into a housing or re-located into another housing;

5 Questions:

- Now host birth and death are recorded in day units. Let me know if more specific time tags (hour and minutes) are needed;
- Possible information about the host include health status, vaccination status, geographic location, genetic factors, nutritional status, behavioral factors, exposure history, socioeconomic status, occupation, and immune status. I did not include them. I suggest to provide these as options that can eventually be introduced in the data schema for specific experiments;