**Title:**

The ShareTrait Entity-Relationship Model: a step toward building the ShareTrait database

**Date:** last modified on 2024-04-08

**Version:** 0.9.5

**Description:**

Document for decision making for the Adaptation of the ShareTrait Dataset, the data portal for making trait data interoperable and reusable.

**Source:**

The ShareTrait Dataset Version 1.0.0 (DOI: <https://doi.org/10.5281/zenodo.8138904>) is the source used for the design and development for the conceptual model, the logical and physical model for the ShareTrait Database.

**Contributors:**

Brett Olivier, Irene Martorelli, Jacintha Ellers, Matty Berg, Wilco Verberk, FelixLeiva

**References:**

* ShareTrait Github repository: <https://github.com/ShareTraitProject/ShareTrait/tree/main/ShareTrait_DB>
* ShareTrait Zenodo repository: <https://zenodo.org/records/8138904>

**Diagram:**

* [ER Diagram](#_yijew6p3278l) Model ShareTrait\_DOMEL LM, Version 0.1.6, date 2024-04-08

**Table Definitions:**

The ShareTrait dataset contains 112 attributes annotated, defined in <https://github.com/ShareTraitProject/ShareTrait/blob/main/ShareTrait_DB/v1.0.0/3_release/ShareTrait_MetaData_v1.0.0.csv> are obtained from 44 record datasets.

The 44 records of version 1.0, belong to 3 main traits defined as "Development", "Fecundity" and "Metabolic Rate".

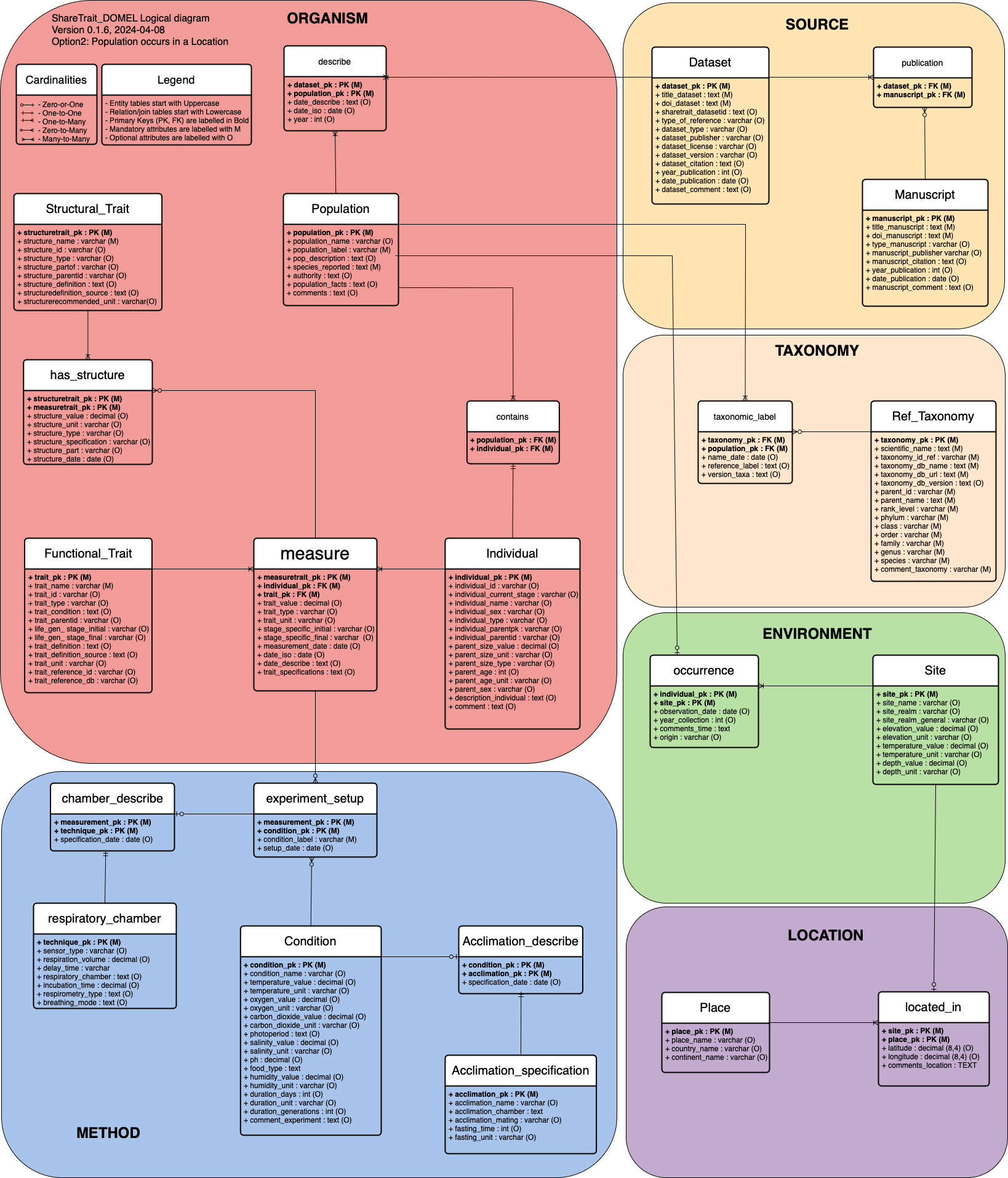
The ShareTrait dataset can be divided into blocks defined as Category blocks. Each Category contains a common collection of entities. For example, a Category defined as Publication may contain information about manuscript, journal, researcher, institute and so on.

Categories are not physical aspects of the database; the aim of these categories is to guide us to know which region/s of the database we are focusing on and to which relevant collection of entities we are dealing with. The colouring labels are used to distinguish these regions of interest.

### ER diagram

# 

Logical model, name diagram: ShareTrait\_AllCat\_DOMEL\_v0.1.6, 2024-04-08



| **Category** | **Definition** |
| --- | --- |
| [Source](#_1mvzqxx2x7tw) | The research main source of information obtained from the scientific collection of investigations dealing with animal traits |
| [Organism](#_p99uq0iry1iu) | Contains information regarding the individual animal groups of interest |
| [Method](#_5ii8pw4d26pc) | Contains relevant information of the investigation applied to the organisms collected for the study |
| [Taxonomy](#_1y8lvfod7qgc) | Contains information regarding the classification of the organisms |
| [Environment](#_cje5e8o5ulg8) | Information belonging to the geographical site of where the data is collected |
| [Location](#_qbzxmp6zs23) | Geographical information regarding the spaces and physical items |

### Category Source

| **Category** | Source | | |
| --- | --- | --- | --- |
| **Definition** | Primary Relevant source of information for ShareTrait. The information is made available by means of published datasets, relevant manuscripts and/or direct contact with researchers | | |
|  | | | |
| **Table Name** | Dataset | | |
| **Table Type** | Entity | | |
| **Definition** | Primary relevant information of the dataset study regarding a trait | | |
| **attribute\_name** | **type** | **definition** | **example** |
| dataset\_pk | primary key | Primary Key of dataset entity, **MANDATORY** | TRADAT000001 |
| title\_dataset | text | title of the dataset | Overwintering strategies and life-history traits of different populations of Aphidius platensis along a latitudinal gradient in Chile |
| doi\_dataset | varchar(256) | dataset DOI, provided in URL (doi\_dataset), **OPTIONAL** | <https://doi.org/10.5281/zenodo.7774767> |
| sharetrait\_datasetid | id | reference id to sharetrait dataset (dataset\_id) | Alfaro\_Tapia\_et\_al\_2022 |
| type\_of\_reference | varchar(128) | Indicate whether the reference corresponds to a primary or secondary reference (type\_of\_reference) | primary |
| dataset\_type | varchar(128) | type of dataset | dataset |
| dataset\_publisher | varchar(256) | publisher of the dataset | zenodo |
| dataset\_license | varchar(256) | license of the dataset (license) | Creative Commons Attribution 4.0 International |
| dataset\_version | varchar(64) | version of the dataset | 1.0 |
| dataset\_citation | text | complete citation of the dataset (full\_citation) | Alfaro-Tapia, A., Alvarez-Baca, J. K., Tougeron, K., Lavandero, B., Le Lann, C., & Van Baaren, J. (2022). Overwintering strategies and life-history traits of different populations of Aphidius platensis along a latitudinal gradient in Chile (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7774767 |
| year\_publication | date | year of publication (year\_publication) | 2022 |
| date\_publication | date | ISO date of the publication | 2022-04-29 |
| dataset\_comment | text | free text remarks regarding the dataset, relevant to population and location (comments\_reference) | five Chilean populations of the aphid parasitoid Aphidius platensis. |
|  | | | |
| **Table Name** | Manuscript | | |
| **Table Type** | Entity | | |
| **Definition** | Information regarding the published article | | |
| **attribute\_name** |  | **definition** | **example** |
| manuscript\_pk | Primary key | PK of manuscript, **MANDATORY** | TRAMAN000001 |
| title\_manuscript | text | title of the manuscript | Overwintering strategies and life-history traits of different populations of Aphidius platensis along a latitudinal gradient in Chile |
| doi\_manuscript | varchar(256) | manuscript DOI, provided in URL (doi\_publication), **MANDATORY** | <https://doi.org/10.1127/entomologia/2021/1186> |
| type\_manuscript | varchar(64) | type of manuscript, can be a journal article, published data method, software paper | article |
| manuscript\_publisher | varchar(256) | publisher of the manuscript | Entomologia Generalis |
| manuscript\_citation | text | full citation of manuscript used for download | Hermaniuk, Adam et al. “Low Temperature and Polyploidy Result in Larger Cell and Body Size in an Ectothermic Vertebrate.” *Physiological and biochemical zoology : PBZ* vol. 89,2 (2016): 118-29. doi:10.1086/684974 |
| year\_publication | date | year of publication of the manuscript | 2022 |
| date\_publication | date | ISO date of publication | 2022-02-11 |
| manuscript\_comment | text | free text remark for manuscript |  |
|  | | | |
| **Table Name** | publication | | |
| **Table Type** | Relation | | |
| **Definition** | Association of a dataset to a published article | | |
| **Notes** | This association does not currently exist in sharetrait, however it is relevant as there can be more than 1 dataset associated to a manuscript, or there can be more than 1 manuscript associated to a dataset (i.e. ShareTrait dataset\_id) | | |
| **attribute\_name** | **type** | **definition** | **example** |
| dataset\_pk | PK | Reference to Dataset entity (foreign key) | TRADAT000001 |
| manuscript\_pk | PK | Reference to Manuscript entity (foreign key) | TRAMAN000001 |

**Association Rules for the Category Source**

**Relationship Dataset - Manuscript (through the Relational Table publication)**

* A dataset is a mandatory entity for ShareTrait that contains information about study of trait(s).
* A manuscript is an entity that contains information about a published article describing a study and it is mandatory to provide a DOI.
* A dataset may be published, in this case it is OPTIONAL to provide a DOI of the dataset.
* A dataset can be part of a *publication* but not necessarily. A dataset can be part of multiple publications (0,N). (This is why the original 1,1 relationship is not possible and manuscript cannot be an attribute of dataset)
* A manuscript to be in ShareTrait must at least describe a dataset, and a manuscript may reference and describe multiple datasets (1,N)

### Category Taxonomy

| **Category** | Taxonomy | | |
| --- | --- | --- | --- |
| **Definition** | Resource containing information about the systematic classification of organisms | | |
|  | | | |
| **Table Name** | Reference\_Taxonomy | | |
| **Table Type** | Entity | | |
| **Definition** | The reference taxonomy classification | | |
| **Notes** | For this version we define the taxonomic ranking complete list for facilitating the aggregation of data during the query (for example: (such as group by based on specific taxonomic levels. For future dev we will provide the ID url of the reference taxonomy. | | |
| **attribute\_name** | **type** | **definition** | **example** |
| taxonomy\_pk | PK | PK of taxonomic name, **MANDATORY** | TRATAX001 |
| scientific\_name | varchar(256) | complete full scientific name of the taxonomic name | Aphidius platensis Brethes, 1913 |
| taxonomy\_id\_ref | ID | taxonomic id of the reference taxonomy. The taxonomy\_id is then generic as the new attribute tax\_db will contain ott db type. (ott = open tree of life) | FD9T |
| taxonomy\_db\_ref | varchar(256) | taxonomic db full name | Catalogue of Life |
| taxonomy\_id\_url | text | taxonomic db url | http://www.ichneumonoidea.name/local.php?taxonidLC=90216058 |
| taxonomy\_db\_version | varchar(64) | taxonomy reference db version | [2024-03-26](https://www.catalogueoflife.org/data/metadata) |
| parent\_id | ID | parent id of taxonomic name if exists | 62D92 |
| parent\_name | text | Complete name of the parent of the scientific name | Aphidius |
| rank\_level | varchar(128) | taxonomic level type of the scientific name | species |
| phylum | varchar(128) | phylum of the scientific taxonomic name (phylum) | [Arthropoda](https://www.catalogueoflife.org/data/taxon/RT) |
| class | varchar(128) | class level of the scientific taxonomic name (class) | Insecta |
| order | varchar(128) | order level of the scientific taxonomic name (order) | [Hymenoptera](https://www.catalogueoflife.org/data/taxon/HYM) |
| family | varchar(128) | family level of the scientific taxonomic name (family) | [Braconidae](https://www.catalogueoflife.org/data/taxon/7D4) |
| genus | varchar(128) | genus level of the scientific taxonomic name (genus) | Aphidius |
| species | varchar(128) | Species level name (species\_ott) | Aphidius platensis |
| comments\_taxonomy | text | Any specific comments on the taxonomy naming (comments\_taxonomy) | Manually parsed the names from url |
|  | | | |
| **Table Name** | taxonomic\_label | | |
| **Table Type** | Relation | | |
| **Definition** | The reference taxonomic labelling mapped to the population observed | | |
| **attribute\_name** | **type** | **definition** | **example** |
| taxonomy\_pk | PK | reference to Ref\_Taxonomy (FK), **MANDATORY** | TRATAX001 |
| population\_pk | PK | reference to Population (FK), **MANDATORY** | TRAPOP001 |
| naming\_date | date | ISO date used for taxonomic labelling | 01-01-2024 |
| reference\_label | varchar(64) | Reference label given to the taxa db | COL |
| version\_name | varchar(128) | version of the taxonomic mapping | 26-03-2024 |

**Association Rules for the Category Taxonomy**

**Relationship Population - Taxonomy (through the Relation taxonomic\_label)**

* A population (group of individuals) can be identified with a species name by the researchers contributing to the study
* A population must be labelled with a scientific name and have an assigned official reference taxonomic name (1,1)
* A population may have multiple approved scientific names based on different reference resources (1,N)
* A taxonomic scientific name can be assigned to an population of individuals, but the same reference name can also be used to label many different individuals (0,N).

### Category Environment

| **Category** | Environment | | |
| --- | --- | --- | --- |
| **Definition** | Contextual information regarding the geographical site where a population occurs and has been observed | | |
|  | | | |
| **Table Name** | Site | | |
| **Table Type** | Entity | | |
| **Definition** | The surface area condition where the population is observed in its natural habitat | | |
| **attribute\_name** | **type** | **definition** | **example** |
| site\_pk | PK | PK of the site, **MANDATORY** | TRASIT001 |
| site\_name | PK | Name of the site | strawberrryfields |
| site\_realm\_general | varchar(256) | indicates where species inhabit (real\_general) | terrestrial |
| site\_realm\_specific | int | Specific details of the habitat (realm\_specific) | terrestrial |
| elevation\_value | decimal(8,2) | Elevation value of the position where the population has being observed (elevation\_value) | 268 |
| elevation\_unit | varchar(30) | Standard unit of elevation value | meter |
| temperature\_value | decimal(8,2) | (avg) air temperature of the site | 17 |
| temparature\_unit | varchar(30) | Standard unit for the temperature value | celcius |
| depth\_value | decimal(8,2) | Depth value where the organism is being observed (depth\_of\_collection) | 0.3 |
| depth\_unit | varchar(30) | Standard unit of the depth value | meter |
|  | | | |
| **Table Name** | occurrence | | |
| **Table Type** | Relation | | |
| **Definition** | The association to identify where a population occurs | | |
| **Notes** | This table may not have the relation with Place directly as currently hardly any natural observation occurrence provides spatial information | | |
| **attribute\_name** | **type** | **definition** | **example** |
| population\_pk | PK | PK of the population, **MANDATORY** | TRAIND001 |
| site\_pk | PK | PK of the geographical area, **MANDATORY** | TRASIT001 |
| observation\_date | date | The ISO date where a population is being observed in its natural habitat (yyyymmdd\_of\_collection) | 06-06-2023 |
| year\_collection | int | Year where the population is observed and collected for further investigation (year\_of\_collection) | 2023 |
| comments\_time | text | Some contextual free-text based on the point that the population is collected (comments\_timing) | Population observed under a pine tree in a warm evening |
| origin | varchar (128) | Defined term, Indicates the source where individuals were collected (origin) | field |

### Category Location

| **Category** | Location | | |
| --- | --- | --- | --- |
| **Definition** | Contextual information regarding the geographical site where a population occurs and has been observed | | |
|  | | | |
| **Table Name** | Place | | |
| **Table Type** | Entity | | |
| **Definition** | Rancagua Chile South America | | |
| **attribute\_name** | **type** | **definition** | **example** |
| place\_pk | PK | PK of taxonomic name, **MANDATORY** | TRAPLA001 |
| place\_name | varchar(256) | A place name of a location. For now there is no specification of what to provide for place\_name. There is no controlled way yet to define this | Rancagua |
| country\_name | varchar(256) | The country where the place is located | Chile |
| continent\_name | varchar(256) | A geographical collection of countries | South America |
|  | | | |
| **Table Name** | located\_in | | |
| **Table Type** | Relation | | |
| **Definition** | The association of a an area of interest with the reference geographical political naming system | | |
| **attribute\_name** | **type** | **definition** | **example** |
| site\_pk | PK | PK of the site, **MANDATORY** | TRASIT001 |
| place\_pk | PK | PK of the geographical area, **MANDATORY** | TRAPLA001 |
| latitude | DECIMAL(8,2) | Latitude, in decimal degrees (lat\_decimal) | 51.15530556 |
| longitude | DECIMAL(8,2) | Longitude value of the position of the site, in decimal degrees (long\_decimal) | 4.398777778 |
| comment\_location | text | Relevant notes about the location (comments\_location) | Population observed under a pine tree in a warm evening |

//add comments here and cardinalities

We have decided to define a site on its own such that it can be extended to provide more contextual information about the physical habitat, as for now with a few metadata fields.

Location gps positions is a choice to provide only between Site and Place as this depends on the relationship between population and site. A population can have different site observations, thus different

### 

### Category Organism

| **Category** | Organism | | |
| --- | --- | --- | --- |
| **Definition** | Primary Relevant source of information regarding the individual species that is being collected and described based on a condition and/or during a study of interest. Characteristics based on individuals are measured and these consist of morphological features and functionalities. | | |
|  | | | |
| **Table Name** | Population | | |
| **Table Type** | Entity | | |
| **Definition** | Information about the group of individuals sharing the same characteristics and located in the same place | | |
| **Notes** | The ShareTrait dataset does not currently contain a defined set of population, it refers to species reported record | | |
| **attribute\_name** | **type** | **definition** | **example** |
| population\_pk | PK | Defines if population reference provides measurements that are derived from a collection rather than individuals | TRAPOP001 |
| population\_name | varchar(256) | Reference to dataset naming of the collection/population | popid1 |
| population\_label | varchar(30) | Defines if population reference provides measurements that are derived from a collection rather than individuals | yes/no |
| pop\_description | text | Free text based characteristics worth to mention | A very big population, only decided to take a few |
| species\_reported | varchar(256) | species name given based on the study (species\_reported) | aphidius\_platensis |
| authority | text | comment containing year and name noted to the species (authority) | rare |
| population\_facts | text | Any relevant information specific for population if mentioned by author/article | Different populations of Aphidius platensis |
| comments | text | any relevant remarks regarding the species described | Species verified by multiple collectors |
|  | | | |
| **Table Name** | describe | | |
| **Table Type** | Relation | | |
| **Definition** | The reference of the dataset to a population observed and described for a study | | |
| **attribute\_name** | **type** | **definition** | **example** |
| dataset\_pk | PK | reference to resource | TRADAT001 |
| population\_id | PK | reference to specimen. This ID is generated | TRAPOP001 |
| date\_describe | text | Date when a population is described | 19 July 2018 |
| date\_iso | date | Standard ISO date | 2018/07/18 |
| year | int | Standard ISO year | 2018 |
|  | | | |
| **Table Name** | Individual | | |
| **Table Type** | Entity | | |
| **Definition** | Information about the physical individual observed and collected for the conduction of the study and the description of the trait. | | |
| **Notes** | The information provided in this entity belongs to the condition of the individual at the moment that has been collected. For example, age of individual is based on the moment of collection, but age can be also an attribute of experiment condition, in which an individual has a certain age when it was mated. At the moment, certain characteristics which we are not aware of the evolution are fixed descriptors for an individual. These are: age, sex, parent info. All parent info will be stored under individual as we are not able to trace which are the parents of the individuals | | |
| **attribute\_name** | **type** | **definition** | **example** |
| individual\_pk | PK | PK of the individual | TRAIN001 |
| individual\_id | varchar(64) | An individual id for identifying subgroups | INDDAT1 |
| individual\_current\_stage | varchar(256) | The life stage if identified during the measurement stage. Specify if it is a egg, larvae, adult stage when individual is collected (life\_stage\_general\_initial) | adult |
| individual\_name | varchar(256) | a name provided to the specimen, this can be associated to a dataset reference | IND454 |
| individual\_sex | varchar(64) | Sex type of individual as indicated in the study. (sex) | female |
| individual\_type | varchar(64) | indicates if it is part of offspring generation or parent type. Used to controlled the parenthood relationships | parent |
| individual\_parentpk | ID | specification if the specimen has a related specimen in the collection, such as parent is known. If parent is unknown, then leave blank | TRAIND001 |
| individual\_parentid | ID | ID of parent as specified in a study | mom54 |
| parent\_size\_value | DECIMAL(8,2) | Current size of the individual when collected (parental\_size\_value) | 9.1 |
| parent\_size\_unit | varchar(64) | Standard unit of measurement for the size metric of the parent individual collected (parental\_size\_units) | millimeter |
| parent\_size\_type | varchar(64) | Type of size metric of the parent (parental\_size\_type) | Fresh body mass |
| parent\_age | INT | age of the parent (parental\_age) | 1 |
| parent\_unit\_age | varchar(64) | units used to define the age of the parent (parental\_age\_units) | year |
| parent\_sex | varchar(64) | Indicates the sex of the parent | female |
| description\_individual | TEXT | free description and remarks relevant to the individual as provided in the dataset |  |
| comment | TEXT | Any extra comments that are not mentioned in the study |  |
|  | | | |
| **Table Name** | contains | | |
| **Table Type** | Relation | | |
| **Definition** | The reference of the population to its individuals | | |
| **attribute\_name** | **type** | **definition** | **example** |
| population\_pk | PK | reference to population, **MANDATORY** | TRAPOP001 |
| individual\_pk | PK | reference to specimen. This ID is generated, **MANDATORY** | TRAIND001 |
|  | | | |
| **Table Name** | Functional\_trait | | |
| **Table Type** | Entity | | |
| **Definition** |  | | |
| **attribute\_name** | **type** | **definition** | **example** |
| trait\_pk | PK | PK of the sharetrait trait, **MANDATORY** | TRATRA001 |
| trait\_name | text | This will be a defined trait (trait\_name) | amount if energy expended per time unit |
| trait\_id | ID | name given to the trait | METRATE |
| trait\_type | varchar(64) | this will include Development, Fecundity, Respiration. There can be more | metabolic\_rate |
| trait\_condition | text | Functional expression used for defining the trait | Frequently measured as rate of oxygen uptake or rate of CO2 production. |
| trait\_parentid | ID | Tree id of functional trait | TRATRA002 |
| life\_gen\_stage\_initial | varchar(64) | Life stage term used for a state of phase that can be specific for trait (life\_general\_stage\_initial) | egg |
| life\_gen\_stage\_final | varchar(64) | End stage term if defined for a phase (life\_general\_stage\_final) | adult |
| trait\_definition | text | Full trait definition to explain the term of the trait | amount if energy expended per time unit |
| trait\_definition\_source | varchar(256) | Source (provide full name) of expert or citation for definition | wilco |
| trait\_unit | varchar(64) | Preferred standard unit for the trait | mLO2/h/ind |
| trait\_reference\_id | text | Reference id (possibly url) of the trait defined | author |
| trait\_reference\_db | varchar(64) | Name of the resource | Can be a book |
|  | | | |
| **Table Name** | Measure | | |
| **Table Type** | Relation | | |
| **Definition** | Quantitative sharetrait characteristic that is measured for an individual / population | | |
| **attribute\_name** | **type** | **definition** | **example** |
| measurement\_pk | PK | PK of the morphological trait, **MANDATORY** | TRAMEA001 |
| individual\_pk | PK | name given to the trait | TRAIND001 |
| trait\_pk | PK | this will be size, developmental stage, life stage | TRATRA001 |
| trait\_value | DECIMAL(5,2) | Measured value of the sharetrait trait (trait\_value) | 96 |
| trait\_type | varchar(64) | Generic measure check for trait name | fecundity |
| trait\_unit | varchar(64) | Unit of measurement used for the trait (trait\_unit) | offspring number |
| stage\_specific\_initial | varchar(256) | Specific initial stage (life\_stage\_specific\_initial) | Gosner stage 25 |
| stage\_specific\_final | varchar(256) | Specific final stage (life\_stage\_specific\_final) | Gosner stage 26 |
| measurement\_date | varchar(256) | Date of the measurement | 21 july 2023 |
| date\_iso | date | ISO date | 21-07-2023 |
| date\_describe | text | Description of the date |  |
| trait\_specifications | text | A comment regarding the trait specifications for the measurement condition |  |
|  | | | |
| **Table Name** | Structural\_trait | | |
| **Table Type** | Entity | | |
| **Definition** | A morphological characteristic of a physical entity | | |
| **attribute\_name** | **type** | **definition** | **example** |
| structuretrait\_pk | PK | PK of the morphological trait | TRASTR002 |
| structuretrait\_name | varchar(256) | this will be size, developmental stage, life stage | body size |
| structuretrait\_id | ID | name given to the trait | size |
| structure\_type | varchar(256) | this will be size, developmental stage, life stage (size\_type) | quantitative |
| structure\_partof | varchar(256) | Body part of measured morphological trait | cortex |
| structure\_parentid | varchar(64) | Label if trait is derived | TRASTR001 |
| structure\_definition | text | this will include Development, Fecundity, Respiration. There can be more | The height and mass of a body. It is often expressed as the ponderal index: body height divided by the cube root of body weight. |
| structure\_definitionsource | varchar(256) | Formula reference to calculate the trait | Oxford dict |
| trait\_recommended\_unit | varchar(64) | Recommended unit if exists | millimeter |
|  |  |  |  |
| **Table Name** | has\_structure | | |
| **Table Type** | join | | |
| **Definition** | Relation that joins a measurement to a structural specification | | |
| **Notes** |  | | |
| **attribute\_name** | **type** | **definition** | **example** |
| structuretrait\_pk | PK | Primary key of the association Trait Individual | TRASTR001 |
| measure\_pk | PK | FK reference to individual entity | TRAMEA001 |
| structure\_value | decimal(5,2) | Numerical value given to the trait measurement (size\_value) | 0.1886 |
| structure\_unit | varchar(64) | Standard unit used for the type of measurement (size\_unit) | milligram |
| structure\_type | varchar(64) | Specification for the structural measurement (size\_type) | dry body mass |
| structure\_specification | varchar(64) | Specification to check to which type of individual is the structure measurement | tadpole |
| structure\_part | varchar(256) | Part of morphological structure if defined | body |
| structure\_date | DATE | Date if known of when the measurement was taken | 21-02-2023 |

### Category Method

| **Category** | Method | | |
| --- | --- | --- | --- |
| **Definition** | The controlled conditions applied to individuals in order to measure specific trait | | |
|  | | | |
| **Table Name** | Condition | | |
| **Table Type** | Entity | | |
| **Definition** | Primary relevant information of the laboratory controlled condition prior to the trait measurement taken | | |
| **attribute\_name** | **type** | **definition** | **example** |
| condition\_pk | primary key | Primary Key of condition entity, **MANDATORY** | TRACON001 |
| condition\_name | varchar(256) | Name of the condition | Name of the condition |
| temperature\_value | decimal (8,2) | Temperature value of the condition (either test, maintenance and acclimation) | 10 |
| temperature\_unit | varchar(30) | Standard unit for the temperature measurement | celsius |
| … |  |  |  |
|  | | | |
| **Table Name** | experiment\_setup | | |
| **Table Type** | join | | |
| **Definition** | Relation of the condition to a measured trait | | |
| **attribute\_name** |  | **definition** | **example** |
| measurement\_pk | Primary key | PK of the measure, **MANDATORY** | TRAMEA001 |
| condition\_pk | PK | PK of the condition entity, **MANDATORY** | TRACON001 |
| contion\_label | varchar(256) | Condition is either test, maintenance, acclimation | test |
| setup\_date | date | ISO date when an experiment condition is setup | 06-06-2023 |
|  | | | |
| **Table Name** | Acclimation\_describe | | |
| **Table Type** | Relation | | |
| **Definition** | Association of a condition to acclimation specifications which are only relevant for condition=acclimation. | | |
| **Notes** | Extension of attributes for condition is acclimation | | |
| **attribute\_name** | **type** | **definition** | **example** |
| condition\_pk | PK | Reference to Condition entity (foreign key) | TRACON001 |
| acclimation\_pk | PK | Reference to Acclimation\_specification (foreign key) | TRAACC001 |
| specification\_date | date | ISO date linked to acclimation specification | 07-07-2023 |
|  | | | |
| **Table Name** | Acclimation\_specification | | |
| **Table Type** | Entity | | |
| **Definition** | Extension specifications for acclimation | | |
| **Notes** |  | | |
| **attribute\_name** | **type** | **definition** | **example** |
| acclimation\_pk | PK | Primary key of acclimation specifications | TRAACC001 |
| acclimation\_name | varchar(256) | Name of acclimation |  |
| acclimation\_chamber | varchar(256) | Provide acclimation time after transfer to respirometry chamber (chamber) | 1 |
| acclimation\_mating |  | Indicates whether mating was allowed or not during the acclimation period (mating) |  |
| fasting\_time |  | Provide duration of animal fasting before placement in respirometry chamber (fasting\_time) | 4 |
| fasting\_unit | varchar(30) | Standard unit of fasting time | hours |
|  | | | |
| **Table Name** | chamber\_describe | | |
| **Table Type** | join | | |
| **Definition** | Specifications regarding the chamber used for specific trait measurements | | |
| **Notes** |  | | |
| **attribute\_name** | **type** | **definition** | **example** |
| measurement\_pk | PK | Primary key of acclimation specifications | TRAMEA001 |
| technique\_pk | PK | Reference to technique\_pk (FK) | TRATEC001 |
| specification\_date | date | Specific date of chamber setup | 04-04-2024 |
|  | | | |
| **Table Name** | respiratory\_chamber | | |
| **Table Type** | entity | | |
| **Definition** | Chamber characteristics and setup used for certain trait measurements | | |
| **Notes** |  | | |
| **attribute\_name** | **type** | **definition** | **example** |
| technique\_pk | PK | Primary key of respiratory\_chamber | TRATEC001 |
| sensor\_type | varchar(30) | Sensor type used to measure the oxygen during the experiment (sensor\_type) | fiber optic-based oxygen analyzer |
| respiration\_volume | decimal | Volume of the empty respirometric chamber (respiration\_volume) | 5000 |
| delay\_time | varchar(256) | Provide wait (delay) time excluded from closed measurement cycles. Also referred as equilibration period. (delay\_time) | NA |
| respiratory\_chamber | text | Material of respirometer, (respiratory\_chamber\_material) | plastic |
| incubation\_time | decimal | Time spend in the respirometry chamber (incubation\_time) | 6 |
| respirometry\_type | text | Type of respirometry (respirometry\_type) | closed |
| breathing\_mode | text | Breathing mode based on the respiratory medium used in the experiment (breathing\_mode) | aquatic |

**Decision making updates:**

[update 09-04-2024]

* Population and not individual is now associated with Site
* Provided a set of individual characteristics as entity "Structural trait" in order to distinguish with the "Functional traits". This will help the query structure for which Functional traits are the ones defined by the authors of sharetrait and it is a requirement to provide a functional trait measurement for this. While structural trait are only specifications of individuals that are measured when the functional trait is measured and exists. For this current state we only have size for which is defined in many different ways. We can define all these different types of sizes for now. This will help to point out the diversity of the size measurements.
* Population decision diff with sharetrait: While ShareTrait defines the attribute species reported as part of the Taxonomy Entity, it is an author contribution of the report, thus is obtained from the reference dataset. Additionally, a dataset may contain more than 1 species reported, thus, this attribute cannot be part of the entities (either Dataset, or Taxonomy), but is part of a relationship, that builds the association with the population ID (Describe3: Tuple id works for dataset\_pk, populatio\_id as foreign key)

[update 08-04-2024]

* location is currently rarely provided so for escience first meeting we keep it as it is (thus association with individual occurs in a specific site located somewhere.

[update 04-04-2024]

**Trait**

* [*current*] Trait entity definition is defined as two sub-entities: Functional and Structural. Functional will contain definitions based on ShareTrait dataset Version 1 (e.g. respiration productivity) and structural is based on the structure of the individual (e.g. size, shape..)
* [*future/proposed*] The trait \_\_Entity)) will be one whole entity with types: Functional, structural, maybe even others. These definitions will the direct association to the measurement value. Then Life stage will be defined as Entity, that will contain the states such as embryo stages. These definitions are build a join association to measurement in which provide specifications based over time.

**Individual-Parent measurements**

* [*current*] : parent measurements (used in particular for fecundity trait annotation are now stored in Individual Entity. This is because it is currently unknown which individuals in the database are the parents.
* [*future/proposed*] : all parent measurements will be stored in the Measurement relation table. This will be possible if the attribute parent\_pk is known

*[update 22-03-2024]*

*Trait extensions*

* future implementation: life cycle with stages specifications!!
* change trait\_name to full trait types and replace current trait\_name to trait\_definition
* Size should be a trait. An individual can have different sizes depending on the stage (point in time) and specific time (such as timestamp, if ever provided). Then size can have multiple types, such as size of body parts (e.g. dry body mass, fresh body mass, tibia length, cephalothorax length..).
* duration value as defined as period of time from one point to another is an attribute of stage having a measurement (e.g. expressed in day) and/or as defined term for stage (for example larvae, embryo..).
* Does it make sense to define Development as a type of trait?

*Reference Taxonomy*

* currently: we will have the Taxonomy entity with: scientific name, and attributes: genus\_name, species\_name, reference\_db and doi of name ?. This is just a list of names, no association yet with the taxonomy tree, but we can show the potentials by using the parent\_id as a genus group
* future implementation: list of reference names (1,N). 1 as they want a mandatory name. Who decides this? This name may differ from species\_reported attribute provided by the author(s) of the dataset

*[update 21-03-2024]*

*Experiment decisions*

* as we do not know to which condition type the chamber and technique specifications refer to, this annotation cannot be related to Condition. Also, unfortunately, it cannot be in addition to the Acclimation specification as we do not know if the chamber condition is relevant during acclimation. For this reason, the respiratory chamber is a specification of the experiment setup if it exists.
* For this prototype we will not split the specifications of the chamber and technique (i.e. following specifications and standard protocols) to the respiratory\_chamber information. This means that all the metadata for now if provided will be part of the same chunk. Not very happy with this decision, but we only model with the data we have
* Chamber stuff can and IS only part of MR, so this relationship makes more sense to add it as a specification of Condition, but we do not know to which condition unfortunately. Because we cannot obtain chamber stuff from acclimation, we need to place this as specifications of measurement

*Environmental decisions & Location*

* realm\_general attribute of sharetrait will be defined as the parent biogeography attribute of the Site Entity.
* For now realm description(s) are attributes of Site. For future implementation of this division, such as for the main realms {terrestrial, freshwater, seawater...}, there must be some sort of classification reference that could be used. For now we stick to the terms used by the authors.
* check realm classification: <https://en.wikipedia.org/wiki/Biogeographic_realm>
* Loc parent\_name of Country is for now Country. We place continent, but for future implementation use GeoNames
* Elevation attribute is now part of Site and not the located\_in relation.
* origin attribute (def: Indicates the source where individuals were collected), thus it is part of occurrence relation (linking Individual - Site)

*[update 19-03-2024]*

Experiment condition decisions

* strategy\_of\_protection is attribute of individual
* the controlled conditions are the inferences to the individual. There are part of the lab experiment setup.
* The experiment\_setup is a relation linking Measurement and Condition. In the relation we will provide the type of condition.
* The possible condition types (for now) available are: Test, Maintenance, Acclimation.
* Acclimation type Condition may have specific properties that will be part of a standard set of Technique machinery setup (for example incubator specifications).
* Behavioural conditions are the conditions which there are is no human interference with the organism. This is the case of observing the animal behaviour in its natural habitat and annotating some information about the context. This information is associated to Individual and NOT measurement (see diagram add version). This will solve the problem that certain traits can be described/measured in a natural environment in a specific time/space.
* At the moment, ShareTrait contains Measurements that contain Condition information. This means that all the measurements are measured in a controlled environment (e.g. lab).

*[update 18-03-2024]*

* Strategy\_of\_protection is attribute of individual
* Sex (immutable for now) is characteristic of individual
* All parent features are part of individual
* Attributes such as Maintained, acclimated can be deleted. It is not necessary to provide them as they exist or not exist based on the association Measurement-Condition.
* For now all measurements contain Condition information. This means that all the trait measured are taken under controlled conditions (such as the lab).
* For all condition measurements, we added units (for example: for temperature, oxygen value, co2 value..)
* Maintenance can have humidity values. And experiment comment can change into a generic comment. OK solved. Maintenance does have the attribute humidity
* The only specification for maintenance is : duration\_days and duration\_generations. Let us see who provides them and how to deal with them
* Maybe comments\_experimental condition falls into experiment\_setup.
* Change maintenance and acclimation duration in value and then define duration\_unit in days
* So the only extra attribute is duration and generation, but that is fine to add them and they are all the same
* Mating is part of acclimation. <https://github.com/ShareTraitProject/ShareTrait/blob/main/ShareTrait_DB/v1.0.0/3_release/ShareTrait_MetaData_v1.0.0.csv>
* <https://research.wayne.edu/iacuc/acclimationofanimals>
* If measurement is a MR type, then it will provide the chamber respiratory standards. Where do we put these fields?

{

Mating: acclimation specific

Method\_type: acclimation

Acclimation\_chamber: acclimation

Fasting\_time: before acclimation

Sensor\_type

Respiration\_volume

Delay\_time

Respiratory\_chamber\_material

Incubation\_time

Respirometry\_type

Breathing\_mode}

Acclimation specific and acclimation type can be part of acclimation? I will do that for now

* Chamber setup is only a property for the Metabolic Rate specification. This means that Chamber setup is associated to experiment setup. Experiment setup contains the measurement\_pk, thus we can obtain the type.
* As chamber is only a specification of MR, we can test if all experiment setup condition is part of a MR trait type.

*[update 12-03-2024]*

* Currently no relationships individual-parent. Parent meta are now properties of Individual.
* no life stage Entity, for now they are part of Trait. (A STAGE SHOULD BE PART OF LIFE CYCLE AND NOT OF TRAIT!!)
* Size is an attribute of Trait\_measure and is not defined as Trait
* Traits are now only treated as quantities (i.e. having measurement values).
* Immutable characteristics are for now treated as attributes of Individual. (i.e. age, sex, adult.. because there are no time stamps for these variables).
* ShareTrait dataset x may contain the same dataset DOI. Data DOI not a mandatory field for now as we have unpublished data that is provided directly from the author.

*[update 08-03-2024]*

* Create 5 records for each dataset. A dataset can have multiple traits
* Current status of sharetrait does not provide relationships with individuals. Therefore we will for now just add parent\_id and measurements linked to parents as features of individuals. So far this type of relationship is 0,1 (an individual may have a parent, or at most one parent).
* We will provide a few records with the link Individual - Parent. This can be a test trial to show. But we won't implement this now as we have no data.
* Ideally we would define traits as either a quantity or description. For quantities we refer to measurements characteristics of the organisms. Such as growth, size, length. The descriptions are related to physical features describing the organism which belong to the morphology and anatomy of individuals. These can be organism group specific, such as wings, fur. An individual will have associations of these characteristics based on condition. For example, a specific size of the individual is measured based on a phase/stage of its lifecycle. These two values will be linked based on specific condition that is measured. We link these based on timestamp or identifier, or a condition that can be experiment/enviromental.
* We only have specific measurements for individuals and no data over time. So this is a 1,1 relationship. Does for this case, a measurement/description can be an attribute of the individual entity, or attribute value of the mapping to trait. We will use this in our model for now and provide a test case for showing that individual has 1,N relationship that can be associated with a collection/condition id.
* The prototype that we will show will: replicate the records of the sharetrait db and provide 2-3 scenarios based on individual-parent relationships and individual-trait multiple relationships.
* A measurement is associated to a single trait\_type only

*[update 05-03-2024]*

* Specific stage annotation are attributes of trait\_measures
* Size, length are not defined as traits but are specifications of the measurements (we have the case of body size, with size\_type). Size for now is an attribute
* Size will also have a stage specification so we define at what stage step the size is being measured. This is the way to distinguish a time point with measurement taken as we do not have any timestamp linked to our dataset

*[update 24-02-2024]*

* To do: structure repo
* Open repo : I will try osf
* Maintenance and acclimation types can only exist if Test exists, thus Mainenance and Acclimation (and in future other types) are parameters of the condition Entity (check version 0.3.3)
* Trait value is development time to change stage is for development trait type
* Trait value is number of offspring for fecundity trait type
* Parents and offspring can have different types of size measured.
* An experimental condition contains a collection of traits measured
* For each main trait, you always have values
* Size of body and type of size of body is not always provided and it is always associated to the main trait type.
* Trait can be defined with a type

*[update 15-02-2024]*

Diagram: category Organism v0.3.1 - Experiment v0.1.1

Comment version : sorting out the relationship between measurement and controlled condition.

* A condition in sharetrait is part of experiment
* A condition has (currently) 3 types : test (T), maintenance (M) and acclimation (A)
* A measurement can depend on any combination of the condition types but can also exist with no condition type. For example, a measurement can be obtained with a test condition after a specific maintenance.
* Type M usually exists if T already exists.
* Condition could exist between measurement and individual (a measurement is obtained in a test condition). If this is a mandatory requirement to provide the test metadata, then it makes sense, however there are measurements with undefined conditions. Or there are traits that exist independently of a controlled condition. For example if sex is defined as a trait
* The 3 types of conditions share the same attribute definitions.(for now)
* A condition can not be directly associated with individual only as then you lose track of which measurements are associated with a condition.
* However, an individual can be associated to condition M and A and T is the only one associated with measurement relation.
* To simplify we will say: sex is a characteristic of individual and can have one of these values: M, F, H and undefined.
* A study can describe multiple traits (D, F, MR)
* We will define the two options: test as attribute of measurement or condition in between measurement and individual.

*[update 12-02-2024]*

Diagram: category Organism v0.3.0

Comment version: this version takes care of the following:

* this version does not distinguish between previous naming and current naming of attributes. (for example, attributes originated from sharetrait dataset)
* This version also provides the definition of Mandatory and Optional fields.
* It is mandatory to provide a scientific name to a given name by author for a population. (Changed 0,N -> 1,N), part of the association Taxonomy - Population.
* For individual: remove all traits related (such as sex, size, these will become relations. For example sex can change over time, size can change over time, age can change..) So for all characteristics that change become relationships with traits. **All characteristics that are immutable are attributes of individuals** (for example, a parent)
* ~~Can provide the test of population measurements as it can only be 1,1.~~ Need to check this as what happens when a population has multiple measurements? Or can have both? It can have both thanks to the association 'contains' and (1,N). But like this you can't test it.
* A population/individual must have at least one trait in order to be part of sharetrait db. (Changed 0,N -> 1,N)
* Still need to find out if traits can be part of a natural environmental state or a trait is measured only in a controlled condition? (for example: size of eggs are measured in the nest area or size of egg measured in the lab?)

Diagram: category Source v0.3.0

* Only manuscripts that are mentioned are provided in sharetrait db (1,N)
* Sex for now will be a fixed character of an individual. We will have 3 types: Male, female, hermaphrodite
* As sharetrait defines, number of offspring is a type of trait

*[update 23-01-2024]*

**Category ORGANISM v0.2.3 [ref:** [**https://app.diagrams.net/#G1zGuzbWagqCQ-wtMIytazmfPVhmfxVx-g%23%7B%22pageId%22%3A%22Iuc56\_GUX5dgKyquYNnz%22%7D**](https://app.diagrams.net/#G1zGuzbWagqCQ-wtMIytazmfPVhmfxVx-g%23%7B%22pageId%22%3A%22Iuc56_GUX5dgKyquYNnz%22%7D)

**Tab - v0.2.3 and not v0.2.3a**

[**https://app.diagrams.net/#G1zGuzbWagqCQ-wtMIytazmfPVhmfxVx-g%23%7B%22pageId%22%3A%22QBO40cwAilZ490CBVbLD%22%7D**](https://app.diagrams.net/#G1zGuzbWagqCQ-wtMIytazmfPVhmfxVx-g%23%7B%22pageId%22%3A%22QBO40cwAilZ490CBVbLD%22%7D)

**VERSION 0.2.3**

\*This version solves the following:

Version: <https://app.diagrams.net/#G1zGuzbWagqCQ-wtMIytazmfPVhmfxVx-g%23%7B%22pageId%22%3A%2232UN9ffHRvUSznTqJovJ%22%7D>

* Population is the species labelled and is associated to the assigned reference name
* Identifies the population(s) that provide information that is an aggregate measurement of a number of individuals. Currently sharetrait excludes studies that contain these values and only relies on individual measurements.
* Solves the possibility of having the foreign key constraint for individual - dataset for population\_id.
* Includes the possibility of having more than one reference name provided for a population
* Includes the possibility of searching data only based on distinct measurements. This supports as well it increased the integration strategy for including more trait based datasets.
* Gives a good outline of a proper classification of traits.

*[update 09-01-2024]*

*\*The relationship between Dataset and Taxonomy can only exist through the relationship describe if the population name reported only belongs to* ***one and only one*** *scientific name defined by a PK. Is this always the case? No. This is because a population name can be assigned from many different scientific reference sources. FOr example, the name reported name1 can be associated to the tree of life reference scientific name id name001, but can be also assigned to another name id belonging to another reference taxonomic db. For this reason I would not include this version.*

*The only way to include this version is only include the tree of life reference and then check via the tree of life reference if it contains cross references to other ids. This could be ideal, but we now define 2 possible solutions to avoid having a species\_reported having multiple reference names assigned.*

***Solution1:***

*define the entity Population to be placed between Taxonomy and Dataset, then contain two relationships assigned\_taxa for associating Taxonomy entity to Population and describe relationship for associating Population to Dataset.*

***Solution2:***

*Keep it the same, in which population\_id is an attribute of an individual and is associated with taxonomy. This is the most realistic approach, because there can be a study (aka dataset) in which only one individual is measured and this one you are not able to provide a name. We cannot drop Individual entity and replace it with a population entity. This is because a population as PK contains multiple individuals, thus is not unique.*

*Solution2 is dropped as it does not describe the actual ShareTrait dataset.*

*Solution1 can be extended. Population entity then is linked to Individual entity by the relationship contains. This version is provided: ShareTrait DB ER diagram, Version 0.6.2, 2024/01/09 (*[*https://app.diagrams.net/#G1zGuzbWagqCQ-wtMIytazmfPVhmfxVx-g%23%7B%22pageId%22%3A%224G0yprBwcniEKtkjoRCv%22%7D*](https://app.diagrams.net/#G1zGuzbWagqCQ-wtMIytazmfPVhmfxVx-g%23%7B%22pageId%22%3A%224G0yprBwcniEKtkjoRCv%22%7D)*).*

*Solution2 solves the following issues of CER diagram Version 0.6.1*

* *Extension of assigning taxonomic scientific names to species reported in the dataset*
* *Technical issue of foreign key constraint of using population\_id as pk.*
* *A population can have individuals belonging to different species (extreme case)*

*In a way, CER diagram version 0.6.1 is a normalised version of Version 0.6.2, where contains relation is dropped and population id becomes an attribute of Individual*

**In this version then we are going to reference the following:**

* **Conceptual diagram: ShareTrait\_DB\_CER Version 0.6.2, 2024/01/09**
* **Logical diagram: ShareTrait\_DB\_LDM Version 0.2.3, 2024/01/09**

**BELOW IT IS JUST PREVIOUS NOTES**

**Notes for self to organise and incorporate:**

please note that the attribute definition of the dataset provided in Zenodo for the ShareTrait dataset v1.0.0 ([ShareTrait\_metadata\_v1.0.0.csv](https://zenodo.org/records/8138904/files/ShareTrait_MetaData_v1.0.0.csv?download=1)) which maps to the Github repo ([ShareTrait\_metadata\_release.csv](https://github.com/ShareTraitProject/ShareTrait/blob/main/ShareTrait_DB/v1.0.0/3_release/ShareTrait_MetaData_v1.0.0.csv)) is not the same field list provided by Wilco (an extended organised attribute list shared by email on 25-10-2023).

This document will rely on Wilco's extended attribute list for building the db tables. However, for testing the relationships and building query use cases, we will rely to use the attribute values found in the Zenodo ShareTrait dataset v1.0.0 (<https://zenodo.org/records/8138904>)

**Backup reference labels for ER Conceptual Design**

**Version: 0.6.2, 2024/01/09**

**List of categories**

The original ShareTrait dataset can be divided into blocks, each one containing a common collection of entities. For example, publication may contain manuscript info, journal, researcher, institute related.

Categories are not physical aspects of the database, they are just to guide us to know which regions of the database we are focusing on and which relevant collection of entities we are dealing with. The colouring labels are used to distinguish these regions of interest.

| **Category** | **Definition** |
| --- | --- |
| Source | The research information source that links the collection of investigations dealing with animal traits |
| Location | Geographical information regarding the spaces and physical items |
| Organism | Contains information regarding the individuals of animal groups of interest |
| Environment | Information belonging to the site of where the data is collected |
| Method | Contains relevant information of the process applied to the information collected for the study |

**Tables**

For each category we can define a set of entities and relationships.

For this document we used the list provided by Wilco, which is an extension of the available release (ref: <https://github.com/ShareTraitProject/ShareTrait/blob/main/ShareTrait_DB/v1.0.0/3_release/ShareTrait_MetaData_v1.0.0.csv>).

Wilco's dataset (version 2, file name: "*model + entitiesv2.xlsx*", shared on 2023/11/01 by email) contains 6 Tabs, each representing an Entity. The list of Tab Entities are: reference, taxonomy, location, timing, exp conditions, trait.

In this document we will reference wilco's attribute naming by following the rule: *entity\_name.attribute\_name*. For example the reference "*reference.doi\_publication*" refers to tab/entity labelled "reference" having the attribute name *doi\_publication*.

The tables in this document will provide the mapping to wilco's attributes and are provided in column "attribute\_ref\_wilco". For empty values, it means that a new attribute was created for the database

[Update 05-12-2023]

The E-R schema version 0.5, referenced in this document, incorporates the following modifications based on E-R schema version 0.4 as of 2023/11/10.

* The category Dataset has changed to Source
* The Entity table Resource has changed to Dataset
* The Entity table Specimen has changed to Individual
* has changedProperties: attributes labelled as mandatory and optional

[Update 09-01-2024)

**Category SOURCE**

| **Category** | Source (old name Dataset) | | | |
| --- | --- | --- | --- | --- |
| **Definition** | Primary Relevant source of information regarding the study that shares in common the research about morphological traits. The information is made available by means of published information, relevant manuscripts and/or direct contact with researchers | | | |
|  | | | | |
| **Table Name** | Dataset (old name Resource) | | | |
| **Table Type** | Entity | | | |
| **Definition** | Primary relevant information of the dataset study regarding traits | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| dataset\_pk |  | primary key | **MANDATORY**, PK of resource entity | TRADAT000001 |
| *sharetrait\_datasetid* | Reference.id\_reference | id | reference id to sharetrait dataset | Alfaro\_Tapia\_et\_al\_2022 |
| title\_dataset |  | text | title of the dataset | Overwintering strategies and life-history traits of different populations of Aphidius platensis along a latitudinal gradient in Chile |
| *doi\_dataset* | reference.doi\_dataset | varchar(256) | **OPTIONAL**  dataset DOI, provided in URL | https://doi.org/10.5281/zenodo.7774767 |
| *type\_of\_reference* | reference.type\_of\_reference | varchar(128) | Indicate whether the reference corresponds to a primary or secondary reference | primary |
| dataset\_type |  | varchar(128) | type of dataset | dataset |
| dataset\_publisher |  | varchar(256) | publisher of the dataset | zenodo |
| *dataset\_license* | reference.license | varchar(256) | licence of the dataset | Creative Commons Attribution 4.0 International |
| dataset\_version |  | varchar(64) | version of the dataset | 1.0 |
| *dataset\_citation* | reference.full\_citation | text | complete citation of the dataset | Alfaro-Tapia, A., Alvarez-Baca, J. K., Tougeron, K., Lavandero, B., Le Lann, C., & Van Baaren, J. (2022). Overwintering strategies and life-history traits of different populations of Aphidius platensis along a latitudinal gradient in Chile (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7774767 |
| *year\_publication* | reference.year\_publication | date | year of publication | 2022 |
| date\_publication |  | date | ISO date of the publication | 2022-04-29 |
| *dataset\_comment* | reference.comments\_reference | text | free text remarks regarding the dataset, relevant to population and location | five Chilean populations of the aphid parasitoid Aphidius platensis. |
|  | | | | |
| **Table Name** | Manuscript | | | |
| **Table Type** | Entity | | | |
| **Definition** | Information regarding the published article | | | |
| **attribute\_name** | **attribute\_ref\_wilco** |  | **definition** | **example** |
| manuscript\_pk |  | Primary key | **MANDATORY**, PK of manuscript | TRAMAN000001 |
| title\_manuscript |  | text | title of the manuscript | Overwintering strategies and life-history traits of different populations of Aphidius platensis along a latitudinal gradient in Chile |
| *doi\_manuscript* | reference.doi\_publication | varchar(256) | **MANDATORY**,manuscript DOI, provided in URL | https://doi.org/10.1127/entomologia/2021/1186 |
| type\_manuscript |  | varchar(64) | type of manuscript, can be article, method, software | article |
| manuscript\_publisher |  | varchar(256) | publisher of the manuscript | Entomologia Generalis |
| manuscript\_citation |  | text | full citation of authors used for download |  |
| year\_publication |  | date | year of publication of the manuscript | 2022 |
| date\_publication |  | date | ISO date of publication | 2022-02-11 |
| manuscript\_comment |  | text | free text remark for manuscript |  |
|  | | | | |
| **Table Name** | publication | | | |
| **Table Type** | Relation | | | |
| **Definition** | Association of a dataset to a published article | | | |
| **Notes** | This association does not currently exist in sharetrait, however it is relevant as there can be more than 1 dataset associated to a manuscript, or there can be more than 1 manuscript associated to a dataset (i.e. ShareTrait dataset) | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| dataset\_pk |  | PK | Ref to Resource entity (foreign key) | TRADAT000001 |
| manuscript\_pk |  | PK | Ref to manuscript entity (foreign key) | TRAMAN000001 |
|  | | | | |
| **Table Name** | Researcher | | | |
| **Table Type** | Entity | | | |
| **Definition** | The person who is an expert and carries out the research regarding a study | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| researcher\_pk |  | Primary key | PK of researcher | TRARES000001 |
| given\_name |  | varchar(256) | first name of the researcher | Armando |
| family\_name |  | varchar(256) | last name of the researcher | Alfaro-Tapia |
| initial\_researcher |  | varchar(16) | initials of the researcher |  |
| *full\_name\_used* |  | varchar(256) | complete name to use for contribution |  |
| ORCID |  | ID | **RECOMMENDED**  person identifier ORCID id provided in URL | https://orcid.org/0000-0001-9166-9874 |
| personal\_url |  | text | personal website provided in URL |  |
|  | | | | |
| **Table Name** | contribution | | | |
| **Table Type** | Relation | | | |
| **Definition** | The association of a researcher dedicating and providing knowledge for the progress of research and study | | | |
| **Note** | ShareTrait contains only a citation attribute (i.e. a string of authors) for contribution information | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| dataset\_pk |  | PK | ref to Dataset (FK) | TRADAT000001 |
| researcher\_pk |  | PK | ref to Researcher (FK) | TRARES000001 |
| contact\_role |  | varchar(256) | provides the label to inform that this is the contact person for this dataset |  |
| contact\_email |  | varchar(256) | email contact provided independently if the contact role is the contact person responsible, maybe i place this with affiliation, need to check |  |
|  | | | | |
| **Table Name** | organisation | | | |
| **Table Type** | Entity | | | |
| **Definition** | Information of the university | | | |
| **Notes** | Currently there is no information regarding the organisation in ShareTrait, however this type of information can be relevant to know where the study is conducted and for building collaborations | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| organisation\_pk |  | Primary key | PK of the organisation | TRAORG000001 |
| organisation\_name |  | varchar(256) | complete name of the organisation | University of Talca |
| organisation\_abbreviation |  | varchar(64) | abbreviation used for the organisation | U Talca |
| organisation\_type |  | varchar(256) | type of organisation (i.e. institute) | university |
| organisation\_url |  | text | website of the organisation | <https://www.utalca.cl/> |
| organisation\_ror |  | ID | RECOMMENDED  Research organisation registry (ID standard term associated to the organisation) | <https://ror.org/01s4gpq44> |
| org\_domain |  | varchar(256) | type of domain associated to the organisation |  |
| comments |  | text | any relevant information regarding the organisation |  |
|  | | | | |
| **Table Name** | affiliation | | | |
| **Table Type** | Relation | | | |
| **Definition** | Association of a person with an organisation | | | |
| **Notes** | Relationship that currently does not exist in ShareTrait, however can be relevant for building networks in relation to trait studies | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| researcher\_pk |  | PK | ref to researcher | TRARES000001 |
| organisation\_pk |  | PK | ref to organisation | TRAORG000001 |
| affiliation\_contact |  | varchar(256) | email of researcher based on organisation |  |
| initial\_date |  | DATE | ISO date for initial start of affiliation with organisation | 2017-03-21 |

**Comments Source Category**

* We do not associate researcher to Manuscript, sharetrait dataset relies on measurements obtained from available datasets published (such as in zenodo). The only thing we do is provide a citation attribute for manuscript.
* For citation attribute, we do not currently provide the type of citation (e.x. bib)

**Association Rules for the Category Souce (old name dataset)**

**Relationship Dataset - Manuscript (through the Relational Table publication)**

* A dataset is a mandatory entity for ShareTrait that contains information about study of trait(s).
* A manuscript is an entity that contains information about a published article describing a study and it is mandatory to provide a DOI.
* A dataset may be published, in this case it is OPTIONAL to provide a DOI of the dataset.
* A dataset can be part of a *publication* but not necessarily. A dataset can be part of multiple publications (0,N). (This is why the original 1,1 relationship is not possible and manuscript cannot be an attribute of dataset)
* A manuscript to be in ShareTrait must at least describe a dataset, and a manuscript may reference and describe multiple datasets (1,N)

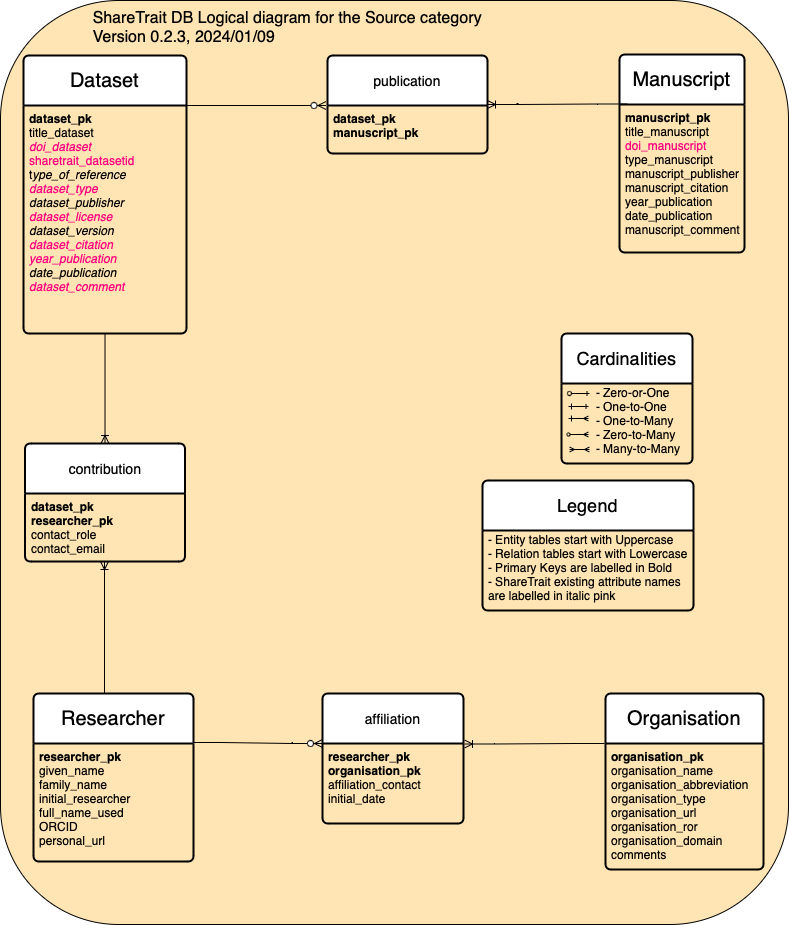
**Relationship Dataset - Researcher (Table contribution)**

* A dataset must contain at least one *contributor* but may contain multiple researchers contributing (1,N)
* A researcher is a mandatory entity that contains information about the person who is involved in a investigation about trait studies
* A researcher in ShareTrait must *contribute* to at least one dataset, but may also contribute to multiple datasets. (1,N)

**Relationship Researcher - Organization (via the Relation Table affiliation)**

* A researcher may be affiliated with an organisation and a researcher may be affiliated with multiple institutions (0,N)
* An organisation must have at least one researcher affiliated with and an institute may have multiple researchers (1,N)

**Logical Schema for Source Category**

****

**Category ORGANISM v0.2.2**

This version is the one that connects individuals directly to a dataset. The individual contains the attribute of population ID that relates the dataset to individual information. This version works, however I felt that this version does not match the logic with the sharetrait experts. The issue is that datasets usually define populations and the naming is directly linked to the study. We keep track of the current definition of these attributes, but below I define a possible clean solution (v.0.2.3) that would make the querylike easier and more logical to the actual sharetrait dataset.

The problem with this version is that population\_id is carried on to Individual from Dataset. This means that both describe and Individual contain the attributes.

| **Category** | Organism | | | |
| --- | --- | --- | --- | --- |
| **Definition** | Primary Relevant source of information regarding the individual species that is being collected and described based on a condition and/or during a study of interest. | | | |
|  | | | | |
| **Table Name** | Individual (//old table name *Specimen*) | | | |
| **Table Type** | Entity | | | |
| **Definition** | Information about the physical individual observed and collected for the conduction of the study and the description of the trait. | | | |
| **Notes** | The information provided in this entity belongs to the condition of the individual at the moment that has been collected. For example, age of individual is based on the moment of collection, but age can be also an attribute of experiment condition, in which an individual has a certain age when it was mated | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| individual\_pk |  | PK | PK of the individual | TRAIND000001 |
| individual\_name |  | varchar(256) | a name provided to the specimen, this can be associated to a dataset reference | IND454 |
| individual\_sex | trait.sex | varchar(64) | Sex type if determined | female |
| lifestage\_type |  | varchar(64) | Specify if it is a egg, larvae, adult stage when individual is collected = this is probably the same as individual stage | adult |
| individual\_type |  | varchar(64) | indicates if it is part of offspring generation or parent type | yes |
| individual\_parentid |  | ID | specification if the specimen has a related specimen in the collection, such as parent is known. If parent is unknown, then leave blank | TRAIND0000001 |
| individual\_parentid2 |  | ID | specification another parent is known. Very rare case, for now this attribute is defined | TRAIND0000001 |
| individual\_stage |  | varchar(256) | Current stage of the individual when it is obtained from its environment. \*\*Check if this is the same as lifestage\_type |  |
| individual\_stage\_general | trait.life\_stage\_general\_initial |  |  |  |
| individual\_stage\_specific | trait.life\_stage\_specific\_inifial |  |  |  |
| individual\_age | trait.parental\_age | INT | age of the specimen - if specimen is parent type, then specimen age is parental age |  |
| unit\_age |  | varchar(64) | units used to define the age of the specimen |  |
| size\_value | trait.size\_value | DECIMAL(8,2) | Current size of the individual when collected |  |
| size\_type | trait.size\_type | varchar(64) | Type of size metric |  |
| size\_unit | trait.size\_units | varchar(64) | Standard unit of measurement for the size metric of the individual collected |  |
| mass\_value |  | DECIMAL(8,2) | Mass of the individual when it is being collected |  |
| mass\_unit |  | varchar(64) | Standard unit of measurement of the mass |  |
| date\_observation |  | DATE | Date when individual actual measurements are being observed |  |
| date\_ISO |  | DATE | Standard date |  |
| description\_individual |  | TEXT | free description and remarks relevant to the individual |  |
|  |  |  |  |  |
| **Table Name** | Describe - [25 jan needs to change to new diagram] | | | |
| **Table Type** | Relation | | | |
| **Definition** | The reference of the dataset to an individual observed and described for a study | | | |
| **Notes** | While ShareTrait defines the attribute species reported as part of the Taxonomy Entity, it is an author contribution of the report, thus is obtained from the reference dataset. Additionally, a dataset may contain more than 1 species reported, thus, this attribute cannot be part of the entities (either Dataset, or Taxonomy), but is part of a relationship, that builds the association with the population ID  Describe3: Tuple id works for dataset\_pk, populatio\_id as foreign key.  Describe2: use individual\_pk for association could be the best possibility. For now we select describe3 f | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| dataset\_pk |  | FK | reference to resource | TRADAT001 |
| population\_id |  | FK | reference to specimen. This ID is generated | TRAPOP001 |
| population\_label |  |  | Defines if population reference provides measurements that are derived from a collection rather than individuals | yes/no |
| population\_name |  | ID | Reference to dataset naming of the collection/population | id1 |
| species\_reported | taxonomy.species\_reported | varchar(256) | species name given based on the study | aphidius\_platensis |
| authority | taxonomy.authority | text | comment containing year and name noted to the species |  |
| populations\_facts |  | text | Any relevant information specific for population if mentioned by author/article | rare |
| comments |  | text | any relevant remarks regarding the species described | Different populations of Aphidius platensis |
| date\_describe |  | date | Date when a population is described | 19 July 2018 |
| date\_iso |  | date | Standard ISO date | 2018/07/18 |
| year |  | date | Standard ISO year | 2018 |
|  |  |  |  |  |
| **Table Name** | Taxonomy | | | |
| **Table Type** | Entity | | | |
| **Definition** | The reference taxonomic labelling for the individual observed | | | |
| **Notes** | We define the taxonomic ranking attributes for facilitating the aggregation of data for the query (such as group by based on specific taxonomic levels) | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| taxonomy\_pk |  | PK | PK of taxonomic name |  |
| scientific\_name | taxonomy.species\_ott | varchar(256) | complete full scientific name of the taxonomic name |  |
| sharetrait\_id\_taxa | taxonomy.id\_taxonomy | ID | reference to sharetrait id |  |
| taxonomy\_id\_ref | taxonomy.species\_ott | ID | taxonomic id of the reference taxonomy. The taxonomy\_id is then generic as the new attribute tax\_db will contain ott db type. (ott = open tree of life) |  |
| taxonomy\_db\_ref |  | varchar(256) | taxonomic db full name |  |
| taxonomy\_db\_url |  | text | taxonomic db url |  |
| taxonomy\_db\_version |  | varchar(64) | taxonomy reference db version |  |
| ~~taxonomy\_ott\_id\_ref~~ | ~~taxonomy.species\_ott~~ | ~~ID~~ | ~~taxonomy open tree of life ID reference~~ |  |
| ~~taxonomy\_ott\_id\_version~~ |  | ~~varchar(64)~~ | ~~taxonomy open tree of life name version~~ |  |
| parent\_id |  | ID | parent id of taxonomic name if exists |  |
| rank\_level |  | varchar(128) | taxonomic level type of the scientific name |  |
| phylum | taxonomy.phylum | varchar(128) | phylum of the scientific taxonomic name |  |
| class | taxonomy.class | varchar(128) | class level of the scientific taxonomic name |  |
| order | taxonomy.order | varchar(128) | order level of the scientific taxonomic name |  |
| family | taxonomy.family | varchar(128) | family level of the scientific taxonomic name |  |
| genus | taxonomy.genus | varchar(128) | genus level of the scientific taxonomic name |  |
| species |  | varchar(128) | Species level name |  |
| comments\_taxonomy | taxonomy.comments\_taxonomy | text | Any specific comments on the taxonomy naming |  |
|  |  |  |  |  |
| **Table Name** | taxonomic\_label | | | |
| **Table Type** | Relation | | | |
| **Definition** | The reference taxonomic labeling for the individual observed | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| population\_id |  | PK | reference to individual (FK) |  |
| taxonomy\_pk |  | PK | reference to taxonomy (FK) |  |
| naming\_date |  | date | ISO date used for taxonomic labelling |  |
| reference\_label |  | varchar(64) | Reference label given to the taxa db |  |
| version\_name |  | varchar(128) | version of the taxonomic mapping |  |
|  |  |  |  |  |
| **Table Name** | Trait (previous name: Morphological\_trait) | | | |
| **Table Type** | Entity | | | |
| **Definition** | characteristic for an individual | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| trait\_pk |  | PK | PK of the morphological trait |  |
| trait\_id |  | ID | name given to the trait |  |
| trait\_name |  | varchar(256) | this will be size, developmental stage, life stage |  |
| trait\_label |  | varchar(64) | Label if trait is derived |  |
| *trait\_type\** |  | varchar(64) | this will include Development, Fecundity, Respiration. There can be more | development |
| trait\_condition |  | text | Formula reference to calculate the trait |  |
| trait\_recommended\_unit |  | varchar(64) | Recommended unit if exists |  |
| trait\_parentid |  | ID | reference to any other major part of trait |  |
| trait\_definition |  | text | Definition of trait |  |
| trait\_definition\_source |  | text | person/documentation reference that gave the definition of the trait |  |
| trait\_reference\_id |  | ID | trait reference id |  |
| trait\_reference\_db |  | varchar(256) | name of reference database for trait |  |
|  |  |  |  |  |
| **Table Name** | trait\_measure | | | |
| **Table Type** | relation | | | |
| **Definition** | trait description for an individual | | | |
| **Notes** | Most likely the trait measurement is linked to individual, independently of the population | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| measuretrait\_pk |  | PK | Primary key of the association Trait Individual |  |
| individual\_pk |  | FK | FK reference to individual entity |  |
| trait\_pk |  | FK | FK reference to Trait entity |  |
| trait\_value |  | decimal(5,2) | Numerical value given to the trait measurement |  |
| trait\_unit |  | varchar(64) | Standard unit used for the type of measurement |  |
| trait\_part |  | varchar(256) | Part of morphology if defined | wing |
| measurement\_date |  | DATE | Date if known of when the measurement was taken |  |
| date\_describe |  | DATE | ISO date format |  |
| date\_iso |  |  |  |  |
| ~~trait\_label~~ |  | ~~varchar(64)~~ | ~~NOT HERE, MOVE to TRAIT Label if trait information is specific to individual. Check this might be trait entity~~ |  |

**Comments**

* Population\_id in this version is not part of individual attributes as it contains relationship now contains this association. Let's test if this works better for adding more studies in sharetrait and filling in the annotation
* Individual contains trait values which are relevant of the individual when is been collected. This is a choice as these values could be defined as Traits and then build the mapping via the trait\_measure. I will think about this to see what works best
* However we include some features of individual that are defined slightly differently than sharetrait. These are: life\_stage\_general\_initial, life\_stage\_general\_final, life\_stage\_specific\_initial and life\_stage\_specific\_final. Then size\_value, size\_type, size\_unit are general attributes that refer to offspring\_size\_value, offspring\_size\_type, offspring\_size\_unit, parental\_size\_value, parental\_size\_type, parental\_size\_unit as used for development trait type, is not just without distinction of parent and offspring type, that becomes an attribute on its own.

**Category ORGANISM v0.2.3**

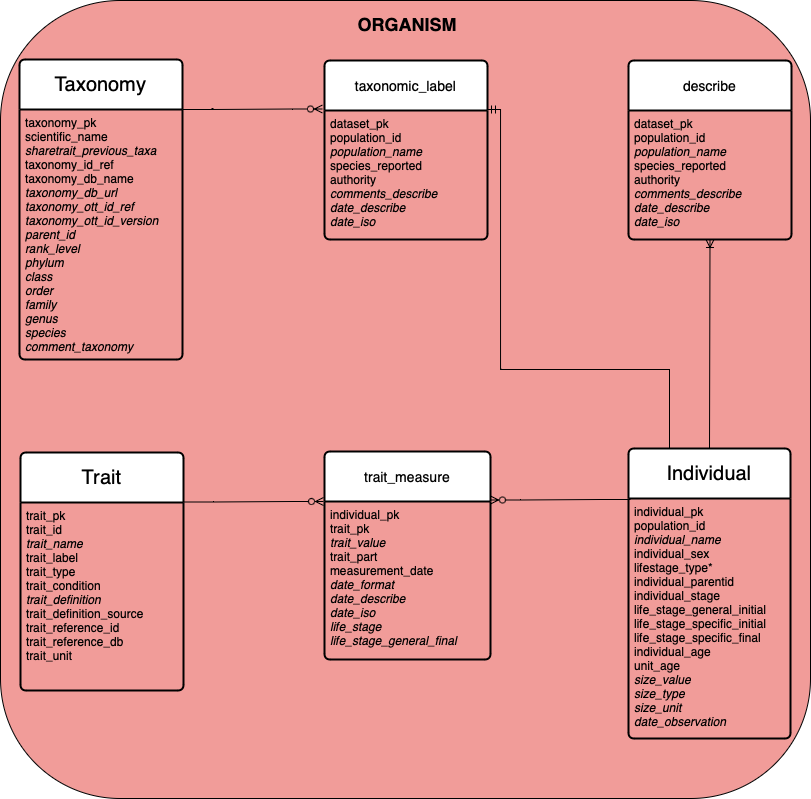
This version solves the following:

* Population is the species labelled and is associated to the assigned reference name
* Identifies the population(s) that provide information that is an aggregate measurement of a number of individuals. Currently sharetrait excludes studies that contain these values and only relies on individual measurements.
* Solves the possibility of having the foreign key constraint for individual - dataset for population\_id.
* Includes the possibility of having more than one reference name provided for a population
* Includes the possibility of searching data only based on distinct measurements. This supports as well it increased the integration strategy for including more trait based datasets.
* Gives a good outline of a proper classification of traits.

| **Category** | Organism | | | |
| --- | --- | --- | --- | --- |
| **Definition** | Primary Relevant source of information regarding the individual species that is being collected and described based on a condition and/or during a study of interest. | | | |
|  | | | | |
| **Table Name** | Individual | | | |
| **Table Type** | Entity | | | |
| **Definition** | Information about the physical individual observed and collected for the conduction of the study and the description of the trait. | | | |
| **Notes** | The information provided in this entity belongs to the condition of the individual at the moment that has been collected. For example, age of individual is based on the moment of collection, but age can be also an attribute of experiment condition, in which an individual has a certain age when it was mated | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| individual\_pk |  | PK | PK of the individual | TRAIND000001 |
| individual\_name |  | varchar(256) | a name provided to the specimen, this can be associated to a dataset reference | IND454 |
| individual\_sex | trait.sex | varchar(64) | Sex type if determined | female |
| lifestage\_type |  | varchar(64) | Specify if it is a egg, larvae, adult stage when individual is collected = this is probably the same as individual stage | adult |
| individual\_type |  | varchar(64) | indicates if it is part of offspring generation or parent type | yes |
| individual\_parentid |  | ID | specification if the specimen has a related specimen in the collection, such as parent is known. If parent is unknown, then leave blank | TRAIND0000001 |
| individual\_parentid2 |  | ID | specification another parent is known. Very rare case, for now this attribute is defined | TRAIND0000001 |
| individual\_stage |  | varchar(256) | Current stage of the individual when it is obtained from its environment. \*\*Check if this is the same as lifestage\_type |  |
| individual\_stage\_general | trait.life\_stage\_general\_initial |  |  |  |
| individual\_stage\_specific | trait.life\_stage\_specific\_inifial |  |  |  |
| individual\_age | trait.parental\_age | INT | age of the specimen - if specimen is parent type, then specimen age is parental age |  |
| unit\_age |  | varchar(64) | units used to define the age of the specimen |  |
| size\_value | trait.size\_value | DECIMAL(8,2) | Current size of the individual when collected |  |
| size\_type | trait.size\_type | varchar(64) | Type of size metric |  |
| size\_unit | trait.size\_units | varchar(64) | Standard unit of measurement for the size metric of the individual collected |  |
| mass\_value |  | DECIMAL(8,2) | Mass of the individual when it is being collected |  |
| mass\_unit |  | varchar(64) | Standard unit of measurement of the mass |  |
| date\_observation |  | DATE | Date when individual actual measurements are being observed |  |
| date\_ISO |  | DATE | Standard date |  |
| description\_individual |  | TEXT | free description and remarks relevant to the individual |  |
|  |  |  |  |  |
| **Table Name** | Population | | | |
| **Table Type** | Entity | | | |
| **Definition** | Information about the population observed and studied during the research. | | | |
| **Notes** | A population contains distinct information obtained from the dataset | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| population\_pk |  | PK | PK of the individual | TRAPOP000001 |
| population\_name |  | varchar(256) | Reference to dataset naming of the collection/population | popdata\_001 |
| *population\_label\** |  | varchar(32) | Defines if population reference provides measurements that are derived from a collection rather than individuals. Note that if population is dataset specific (1,1), then it makes sense that population label is an attribute. However if population measurement is specific for certain condition, then it makes sense that contains relation contains the attribute pop\_label | yes/no |
| pop\_description |  | text | Any context that can be relevant for the population | Population x occurs only during extreme conditions such snow storms |
| species\_reported | taxonomy.species\_reported | varchar(256) | species name given based on the study | aphidius\_platensis |
| authority | taxonomy.authority | text | comment containing year and name noted to the species |  |
| population\_facts |  | text | Any relevant information specific for population if mentioned by author/article | rare |
| comments |  | text | any relevant remarks regarding the species described | Different populations of Aphidius platensis |
|  |  |  |  |  |
| **Table Name** | Describe | | | |
| **Table Type** | Relation | | | |
| **Definition** | The reference of the dataset to a populationl observed and described for a study | | | |
| **Notes** | While ShareTrait defines the attribute species reported as part of the Taxonomy Entity, it is an author contribution of the report, thus is obtained from the reference dataset. Additionally, a dataset may contain more than 1 species reported, thus, this attribute cannot be part of the entities (either Dataset, or Taxonomy), but is part of a relationship, that builds the association with the population ID  Describe3: Tuple id works for dataset\_pk, populatio\_id as foreign key.  Describe2: use individual\_pk for association could be the best possibility. For now we select describe3 f  Update: species\_reported is part\_of Population Entity. This can only be true based on the definition that A Population belongs to only one species. | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| dataset\_pk |  | FK | reference to resource | TRADAT001 |
| population\_pk |  | FK | reference to population. This ID is generated | TRAPOP001 |
| date\_describe |  | date | Date when a population is described | 19 July 2018 |
| date\_iso |  | date | Standard ISO date | 2018/07/18 |
| year |  | date | Standard ISO year | 2018 |
|  |  |  |  |  |
| **Table Name** | Taxonomy | | | |
| **Table Type** | Entity | | | |
| **Definition** | The reference taxonomic labelling for the individual observed | | | |
| **Notes** | We define the taxonomic ranking attributes for facilitating the aggregation of data for the query (such as group by based on specific taxonomic levels) | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| taxonomy\_pk |  | PK | PK of taxonomic name |  |
| scientific\_name | taxonomy.species\_ott | varchar(256) | complete full scientific name of the taxonomic name |  |
| sharetrait\_id\_taxa | taxonomy.id\_taxonomy | ID | reference to sharetrait id |  |
| taxonomy\_id\_ref | taxonomy.species\_ott | ID | taxonomic id of the reference taxonomy. The taxonomy\_id is then generic as the new attribute tax\_db will contain ott db type. (ott = open tree of life) |  |
| taxonomy\_db\_name |  | varchar(256) | taxonomic db full name |  |
| taxonomy\_db\_url |  | text | taxonomic db url |  |
| taxonomy\_db\_version |  | varchar(64) | taxonomy reference db version |  |
| parent\_id |  | ID | parent id of taxonomic name if exists |  |
| rank\_level |  | varchar(128) | taxonomic level type of the scientific name |  |
| phylum | taxonomy.phylum | varchar(128) | phylum of the scientific taxonomic name |  |
| class | taxonomy.class | varchar(128) | class level of the scientific taxonomic name |  |
| order | taxonomy.order | varchar(128) | order level of the scientific taxonomic name |  |
| family | taxonomy.family | varchar(128) | family level of the scientific taxonomic name |  |
| genus | taxonomy.genus | varchar(128) | genus level of the scientific taxonomic name |  |
| species |  | varchar(128) | Species level name |  |
| comments\_taxonomy | taxonomy.comments\_taxonomy | text | Any specific comments on the taxonomy naming |  |
|  |  |  |  |  |
| **Table Name** | taxonomic\_label | | | |
| **Table Type** | Relation | | | |
| **Definition** | The reference taxonomic labeling for the individual observed | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| population\_pk |  | PK | reference to population (FK) |  |
| taxonomy\_pk |  | PK | reference to taxonomy (FK) |  |
| naming\_date |  | date | ISO date used for taxonomic labelling |  |
| reference\_label |  | varchar(64) | Reference label given to the taxa db |  |
| version\_taxa |  | varchar(128) | version of the taxonomic mapping |  |
|  |  |  |  |  |
| **Table Name** | Trait | | | |
| **Table Type** | Entity | | | |
| **Definition** | characteristic for a morphological trait | | | |
| **Notes** | Note that there is no reference to identifiers of sharetrait\_id. This is a complete entity table that merges the 3 main traits entity tables that are currently separated. These are "Development", "Fecundity", "Respiration" | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| trait\_pk |  | PK | PK of the morphological trait |  |
| trait\_id |  | ID | name given to the trait |  |
| trait\_name |  | varchar(256) | this will be size, developmental stage, life stage |  |
| trait\_label |  | varchar(64) | Label if trait is derived |  |
| *trait\_type\** |  | varchar(64) | this will include Development, Fecundity, Respiration. There can be more. Note that type refers to the 3 entity tables of sharetrait "Development", "Fecundity", "Respiration" | development |
| trait\_condition |  | text | Formula reference to calculate the trait |  |
| trait\_recommended\_unit |  | varchar(64) | Recommended unit if exists |  |
| trait\_parentid |  | ID | reference to any other major part of trait |  |
| trait\_definition |  | text | Definition of trait |  |
| trait\_definition\_source |  | text | person/documentation reference that gave the definition of the trait |  |
| trait\_reference\_id |  | ID | trait reference id |  |
| trait\_reference\_db |  | varchar(256) | name of reference database for trait |  |
|  |  |  |  |  |
| **Table Name** | trait\_measure | | | |
| **Table Type** | relation | | | |
| **Definition** | trait description for an individual | | | |
| **Notes** | Most likely the trait measurement is linked to individual, independently of the population | | | |
| **attribute\_name** | **attribute\_ref\_wilco** | **type** | **definition** | **example** |
| measuretrait\_pk |  | PK | Primary key of the association Trait Individual |  |
| individual\_pk |  | FK | FK reference to individual entity |  |
| trait\_pk |  | FK | FK reference to Trait entity |  |
| trait\_value |  | decimal(5,2) | Numerical value given to the trait measurement |  |
| trait\_unit |  | varchar(64) | Standard unit used for the type of measurement |  |
| trait\_part |  | varchar(256) | Part of morphology if defined | wing |
| measurement\_date |  | DATE | Date if known of when the measurement was taken |  |
| date\_describe |  | DATE | ISO date format |  |
| date\_iso |  |  |  |  |
| ~~trait\_label~~ |  | ~~varchar(64)~~ | ~~NOT HERE, MOVE to TRAIT Label if trait information is specific to individual. Check this might be trait entity~~ |  |

**Logical Schema for Organism Category**

**Version 0.2.3 - approved for the constraints mentioned**

****

**Association Rules for the Category Organism**

**Relationship Dataset - Individual (Table describe)**

* A dataset is a mandatory entity for ShareTrait as it must contains information about study of trait(s) belonging to an individual or a group of individuals.
* An individual is an entity that contains information and specific characteristics and description of the individual organism.
* A dataset describes a population, a collection of individuals belonging to the same species. However, a dataset may describe multiple populations. (1,N)
* A population must be described in at least in one dataset, but the same population may be described in more than one dataset. (1,N)
* An individual is specific and can be part of for only one population (attribute of Individual) (1,1)
* A dataset can describe a set of individuals without specifying a population. This set of individuals belongs to a population if collected from the same region and condition. In this case, the set of individuals is grouped together with a population identifier that is defined by the db.
* A population contains only one type of individuals belonging to the same species.
* A population can be defined as population when it belongs to the same species and must have at least one individual in the collection

**Relationship Individual - Taxonomy (Table taxonomic\_label)**

* A population (group of individuals) can be identified with a species name by the researchers contributing to the study
* An individual must be labelled and have an assigned official reference taxonomic name (1,1)
* A taxonomic scientific name can be assigned to an individual, but the same reference name can also be used to label many different individuals (0,N)

**Relationship Individual - Trait (Table trait\_measure)**

* Each individual has specific characteristics that are dependent on the current status when it was observed and collected.
* An individual can have a morphological trait measured, but can also have many traits observed and measured. (0,N)
* A trait can be measured for an individual but can also be used to measure many individuals, independently of the population. (0,N)

**ShareTrait dataset reference**

| id | title | DOI | dataset\_repo |
| --- | --- | --- | --- |
| 1 | Alfaro\_Tapia\_et\_al\_2022 | <https://doi.org/10.5281/zenodo.7774767> | zenodo |
| 2 | Burraco\_et\_al\_2020 | <https://doi.org/10.6084/m9.figshare.9992201.v1> | figshare |
| 3 | Castaneda\_and\_Nespolo\_2013 | <https://doi.org/10.6084/m9.figshare.22561378.v1> | figshare |
| 4 | Hermaniuk\_et\_al\_2016 | <https://doi.org/10.1086/684974> | no dataset |
| 5 | Jerbi\_Elayed\_et\_al\_2022 | <https://doi.org/10.5281/zenodo.4721086> | zenodo |
| 6 | Jorissen\_et\_al\_2023 | <https://doi.org/10.48804/M9FUPT> | dataverse |
| 7 | Le\_Lann\_et\_al\_2014 | <https://doi.org/10.5281/zenodo.7775047> | zenodo |
| 8 | Saeed\_et\_al\_2021 | <https://doi.org/10.5281/zenodo.7764457> | zenodo |
| 9 | Semsar-kazerouni\_et\_al\_2022 | <https://doi.org/10.17026/dans-22w-ygkz> | dans |
| 10 | Shameer\_et\_al\_2018 | <https://doi.org/10.5281/zenodo.7811831> | zenodo |
| 11 | Tougeron\_et\_al\_2020 | <https://doi.org/10.5281/zenodo.7763931> | zenodo |
| 12 | van\_Dis\_et\_al\_2021 | <https://doi.org/10.5061/dryad.15dv41nx0> | dryad |
| 13 = 1 | Alfaro\_Tapia\_et\_al\_2022 same as ID1, but contains other info that is not associated to ID1 | <https://doi.org/10.5281/zenodo.7774767> | zenodo |
| 14 | Gebauer\_et\_al\_2007 | <https://doi.org/10.5281/zenodo.7313889> | zenodo |
| 15 | Hoffer\_et\_al\_2012 | <https://doi.org/10.48338/VU01-0OCUNQ> | yoda |
| 16 | Jerbi\_Elayed\_et\_al\_2021 | <https://doi.org/10.5281/zenodo.7763964> | zenodo |
| 17 | Keinan\_et\_al\_2017 | <https://doi.org/10.1111/een.12422> | no dataset |
| 18 | Le\_Goff\_et\_al\_2021 | <https://doi.org/10.6084/m9.figshare.13187642.v1> | figshare |
| 19 - 7 | Le\_Lann\_et\_al\_2014 | <https://doi.org/10.5281/zenodo.7775047> | zenodo |
| 20 | Mesas\_and\_Castaneda\_2023 | <https://doi.org/10.5061/dryad.2v6wwpzsg> | dryad |
| 21 | Molinet\_et\_al\_2023 | <https://doi.org/10.1016/j.fishres.2020.105803> | no dataset |
| 22 | Saeed\_et\_al\_2021 | <https://doi.org/10.5281/zenodo.7764457> | zenodo |
| 23 | Tougeron\_et\_al\_2021 | <https://doi.org/10.5281/zenodo.5342664> | zenodo |
| 24 | Tuni\_et\_al\_2016 | <https://doi.org/10.5061/dryad.72g06> | dryad |
| 25 | Vogels\_et\_al\_2021 | <https://doi.org/10.17026/dans-zsa-f3y9> | dans |
| 26 | Barneche\_et\_al\_2019 | https://doi.org/10.5281/zenodo.2634100 | zenodo |
| 27 = 3 | Castaneda\_and\_Nespolo\_2013 | https://doi.org/10.6084/m9.figshare.22561378 | figshare |
| 28 | Gomez\_Isaza\_et\_al\_2020\_1 | https://doi.org/10.14264/fe5e038 | UQ eSpace |
| 29 | Hermaniuk\_et\_al\_2017 | https://doi.org/10.1086/689408 | no dataset |
| 30 | Hermaniuk\_et\_al\_2021 | https://doi.org/10.5061/dryad.2280gb5qw | dryad |
| 31 = 7 | Le\_Lann\_et\_al\_2014 | https://doi.org/10.5281/zenodo.7775047 | zenodo |
| 32 | Leiva\_et\_al\_2015 | https://doi.org/10.5281/zenodo.7273461 | zenodo |
| 33 | Mesas\_and\_Castaneda\_2023 | https://doi.org/10.5061/dryad.2v6wwpzsg | dryad |
| 34 | Shokri\_et\_al\_2022 | https://doi.org/10.17605/OSF.IO/56TNH | OSF |
| 35 | vandePol\_unpublished |  | no dataset |
| 36 | Verberk\_and\_Bilton\_2015 | https://doi.org/10.5281/zenodo.7772187 | zenodo |
| 37 | Wu\_et\_al\_2015 | <https://zenodo.org/record/7597994#.Y9t3QuxByAl> | zenodo |
| 38 = 37 | Wu\_et\_al\_2018 | https://github.com/nicholaswunz/old-data | Github |
| 39 | Wu\_et\_al\_2022 | https://doi.org/10.5061/dryad.v6wwpzgxm | dryad |

selfNOTES

Last edit: 02-

**Definitions:**

* A study must describe at least one individual recognized and identified with a taxonomic name
* A study can describe a group of individuals that belong to a population.
* A population is a group of individuals belonging to the same species.
* A population is a group of individuals observed in a region of interest
* A dataset can describe many populations, for each one belonging to different locations
* A study can describe different populations belonging to the same species. I have not identified a case for this
* A dataset describes a population unknowing which individuals it contains. (in other words, you cannot trace an individual, a record has no id/ref to population)
* An individual can be related to another individual
* An individual can be derived from another individual. In this case, the individual offspring will be associated to a parent individual.
* So far there was no case of an individual containing more than one parent.
* An individual contains a population\_id relationship.
* An individual contains a parent\_id relationship
* A population\_id is a group\_id
* Describe table will contain foreign key referencing the primary key of dataset and foreign key referencing population id of individual table.

**Examples:**

**Dataset id1:**

* Reference: Alfaro\_Tapia\_et\_al\_2022, <https://doi.org/10.5281/zenodo.7774767>
* Note that sharetrait contains two datasets referring to the same zenodo DOI. These correspond to the two files but they are not associated
* Contains 5 chileans populations, however it is difficult to know which individual belongs to who
* This dataset contains two files which were described in different times and never associated the size of eggs to populations
* The second tab of this dataset is defined in dataset id 13
* Independently of this issue, we could probably define the following
* Sharetrait contains 3693\* records (the first rows 1-3693) (row 1-3693 of tab dev\_time of <https://doi.org/10.5281/zenodo.7774767>), 112 attributes
* From these 112 attributes, the dataset and manuscript information are unique (1 value only) we know the following:
* 26 attributes are unique (for example, doi is an attribute shared with all 3693 rows)
* 82 attributes have empty values (NA)
* 4 attributes contain important measurement information (in this case temperature and amount of days) in relation to population (5 types) and environmental condition (in this case 5 elevation specific for location)
* The way the dataset is provided in ShareTrait you do not know which individuals are measured, reference to egg size based on developmental stage is lost (second table of original dataset of Alfaro)
* If we rely on the way ShareTrait is providing the data (we are replicating), then we only have 1 trait : number of days, which is trait type development. We are then going to define traits with types and measurement is going to have the value (i.e. number of days), the trait entity will have the unit, in this case unit value of development day is day. We then see how to associate size of eggs based on trait type, but at the moment we cannot build the association developmental stage (in days) vs size of egg in specific condition (such as temperature and elevation)

Species reported is not part of dataset attribute because a dataset can report multiple species. Species reported is a describe attribute and is part of population information

**Trait and Individual "common" attributes**

An individual has a size type, size value and size units, this is the actual (initial) value size. This is a characteristic of individual.

The trait has an attribute named size with its own definition.

* Trait entity has only defined attributes with reference to possible standards
* The trait measure is the relationship of traits with individuals. This is not based on population association because each individual will have its own condition and value.
* If a dataset provides trait values based on aggregation, average measurements, then this measurement is provided for the individual table with a label specifying that is a derived measurement.
* An individual can have a size and mass as

[update 2023-12-11]

* Mass and size of individual will not be part of individual attributes, but rather these will be trait definitions (part of the Trait entity) and the relationship between individual and Trait will contain the value and date. Listing all individual trait values belonging to a specific trait (i.e. size) with a specific date will help to trace the change over time (development)

Example: two versions of dealing with the size of an individual over time (such as development)

Option 1:

Size is not seen as a characteristic of an individual, but rather it is a measurement that can be calculated.

If trait name is size and has the pk TRAIT001, and individual is identified with IND001, then the relation "trait\_measure" can be the following:

| Trameasure\_pk | Individual\_pk | Trait\_pk | Value | stage | date |
| --- | --- | --- | --- | --- | --- |
| MEA001 | IND001 | TRAIT001 | 200 | egg | 2023/12/01 |
| MEA002 | IND001 | TRAIT001 | 220 | egg | 2023/12/02 |
| MEA003 | IND001 | TRAIT001 | 225 | egg | 2023/12/03 |
| MEA004 | IND001 | TRAIT001 | 230 | egg | 2023/12/04 |

Then we can say that individual 1 has 4 measurements collected during the egg stage with the values: 200, 220, 225, 230. The final egg size is of 230 g.

Option2:

Say now, size is a characteristic of the individual and specific metadata of the date is not provided. The development trait can be defined as the following:

Entity Trait

| Trait\_pk | trait\_name | trait\_type | description |
| --- | --- | --- | --- |
| TRAIT001 | Development time | development | development time = age at stage 2 - age at stage 1 |
| TRAIT002 | Development rate | development | development rate = 1 / development time |
| TRAIT003 | Age |  | Age of an individual based on days |
| TRAIT004 | Egg size |  | Measurement of an egg |

Entity Individual

| Individual\_pk | pop\_id | stage | Value | parent\_id | age |
| --- | --- | --- | --- | --- | --- |
| IND001 | POP001 | adult | 200 |  | 3 |
| IND002 | POP001 | adult | 220 |  | 3 |
| IND003 | POP001 | egg | 225 |  | 37 |

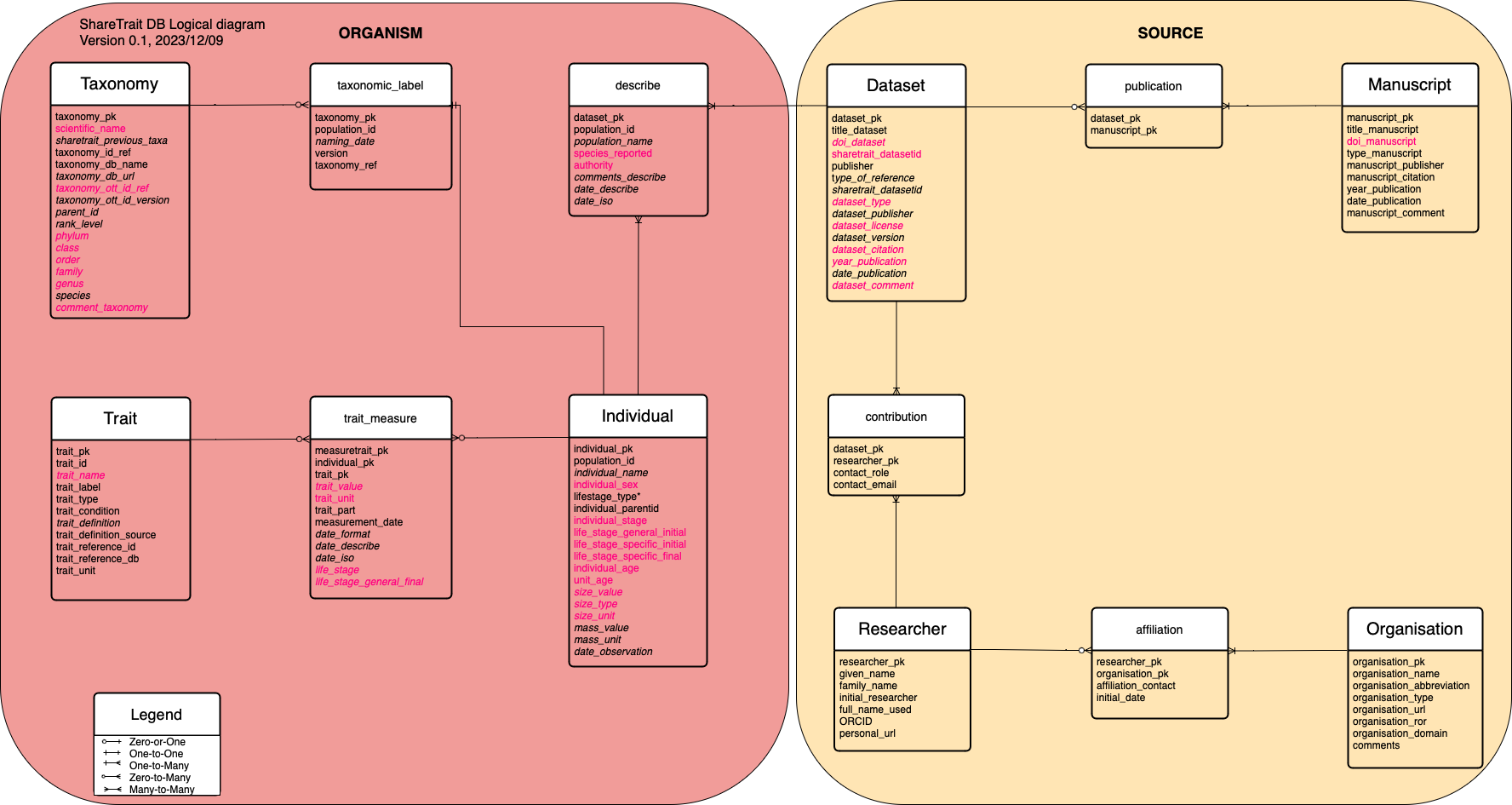
Trait\_measurement

| Trameasure\_pk | Individual\_pk | Trait\_pk | Value | stage | date |
| --- | --- | --- | --- | --- | --- |
| MEA001 | IND003 | TRAIT003 | 200 | egg | 2023/12/01 |
| MEA002 | IND003 | TRAIT003 | 220 | egg | 2023/12/02 |
| MEA003 | IND003 | TRAIT003 | 225 | egg | 2023/12/03 |
| MEA004 | IND001 | TRAIT001 | 230 | egg | 2023/12/04 |

| Individual | trait\_measure | Trait |
| --- | --- | --- |
| individual\_pk | size | trait\_pk |
| individual\_name | Part\_of = | Trait\_name = size, development time |
| sex | date | trait\_type |
| size\_type | value | trait\_unit |
| size\_value |  | trait\_definition |
| size\_units |  | Trait\_parentid = ref to another trait \* |
| collection\_date |  |  |
| life\_stage |  |  |
| age\_collection |  |  |
| date\_age |  |  |
| ISO\_date\_age |  |  |

[update 2023-12-12]

**Reference diagram (Logical diagram, v0.2 - Categories Source and Organism)**

****

**Notes after meeting with Jacintha, Wilco, Matty, Brett**

* Confirmed that a population contains group of individuals belonging to the same species
* Population derived measurements (for example the average size of egg) needs to be included as there will be datasets (studies) that provide measurements of traits based on a collection (group of individuals). For example the average size of egg is 0.5 cm in diameter for a population of 10 individuals.
* The information above can be contained in ShareTrait as long as this annotation is flagged. This should work and we will test it for the next meeting. (//label population if it contains average measurements)
* Type of questions: which traits? Which traits are considered in a study? Which traits are specific for a taxonomic group? What is the average developmental stage for an egg to turn into an adult for the x species? Are there any studies in sharetrait which contain an average measurement already provided? If so which datasets and what are these traits measured
* Still need to understand if specific individual characteristics should be defined as traits
* The Trait measurement relationship between Trait and Individual must include a numerical value and a standard unit. Ideally also a date should be part of the annotation
* Classification of Traits: is there a reference? And can we define a trait based on a specific phase of a stage? Would this simplify the query and make an easier selection of specification of the phylogeny?

**For next meeting (proposed date 29-01):**

* OK Taxonomy entity is associated to Study and not to the Individual (mapping is population name - scientific name)
* A concrete first list of Trait definitions should be added. These definitions should contain a type and specifications (for example, are they defining part\_of stages, part of morphological characteristics..)
* All associations of Traits are with individuals. The association are measurement numerical values of traits.
* OK For the case that a given measurement is provided by an average (thus you are not able to trace to which individual it belongs to, then population\_id refers to one individual\_id and specify that measurement is based on an average. In this way it is easy to distinguish populations that have distinct individual records measurements with the ones that are just provided as a single value based on 100 individuals for example
* OK Provide table templates (in sql format), sort out the foreign key issues (//share with Brett the db)
* We will populate the tables listed of Source and Organism attributes to include real sharetrait values and also include artefact values for average measurements for population
* List of query like user questions to see if we can replicate the sharetrait dataset (selected attributes) and to see if we can distinguish with average measurements based on individuals and average measurements provided based on population.

[18-12-2023]

* If a population contains a group of the SAME species, then it makes more sense to map taxonomy reference to source and NOT to individual.
* Species reported cannot be an attribute of dataset, as a dataset can have many species reported. Species\_reported becomes a