TRY Data Integration Workflow (TRY version 5, 2019-08-15)

Two basic decisions when designing the TRY database (1: data contribution without barriers, 2: data release ready-for-use by non-experts) caused the need for substantial data curation in the context of the TRY database. This is specifically relevant, as plant traits are characterized by high level of idiosyncrasy, as trait measurements are in most cases not sampled for systematic screening of biodiversity with common and well-standardized measurement protocols, but with respect to specific research questions, which often have specific requirements for trait measurements. The development of a systematic data-curation workflow was therefore an essential part of the first phase of the TRY initiative. The workflow was then gradually improved as standardized external data became available, i.e. trait definitions, taxonomic backbone and functional classification of plant species.

The TRY data curation workflow (or the integration of new datasets into the TRY database) consists of the key aspects: data complementation, consolidation, quality assurance and preparation for data releases. The guiding principle in this context is to preserve the original data and annotate them with complementing and consolidated information and data quality attributes. Changing the original data is the exception and in collaboration with dataset custodians.

All datasets and all additional information submitted to TRY are stored and archived on the file system of the Max-Planck-Institute for Biogeochemistry.

Data contribution:

New datasets are in general contributed via the TRY website: https://www.try-db.org/TryWeb/Submission.php

In the context of data contribution we ask for two things:

(1) To contribute as much additional information (auxiliary data, meta data) about plant growth and trait measurement conditions as the data contributor – the expert for the given dataset – assumes necessary to understand and correctly interpret the data. If the additional information is provided as text, the individual meta-data are extracted before they are entered into the database.
(2) To contribute original, disaggregated trait records rather than aggregated averages. Disaggregated data provide a better representation of trait variation, i.e. intraspecific variation, compared to aggregated data. So far the size of in-situ measured trait datasets was not limiting and posterior aggregation, e.g. in the context of analyses, is possible, while disaggregation is not.

Data integration is organized in seven major steps:

1: Data complementation 1

Before data are imported into the TRY database the dataset is checked for missing information, e.g. geographic references. If possible, this is added, i.e. from the publication related to the dataset. In this context structured meta-data are extracted from unstructured textual information.

2: Data consolidation

Data consolidation is based on the following (sub)steps: structural data integration; semantic integration of taxonomy, trait names and names of metadata; standardization of well represented numerical traits and the most relevant meta data and categorical traits

2.1: Structural integration

Data consolidation is based on several steps to fully integrate new datasets. In the context of dataset import into the TRY database all data are transformed to the entity-attribute-value (EAV) data model of the TRY database. All trait records and auxiliary data are stored in one long table with the three principal columns for entity, attribute and value. All trait measurements and all auxiliary data measured on the same entity (e.g. the traits leaf area, leaf dry mass and SLA and the auxiliary data for geo-reference, measurement date, etc. measured on the same leaf at the same date) are combined to an observation with unique identifier. This data model is consistent with two fundamental framework ontologies in trait based ecology: the EAV model is consistent with the entity-quality model (Mungal et al. 2010, Garnier et al. 2017), which defines the trait of an organism to be the 'quality of an entity'. (2) The aggregation of different measurements on the same entity is consistent with the OBOE framework ontology (Madin et al. 2007), which conceptualizes an observation as combination of several measurements on the same entity.

2.2: Semantic integration: taxonomic names, names of traits and ancillary data

2.2.1: Plant taxonomy

Plant taxonomy is consolidated using the Taxonomic Names Resolution Service (TNRS) developed by iPlant (http://tnrs.iplantcollaborative.org/, Boyle et al. 2013) against a species backbone with a taxonomic backbone based on the Plant List (http://www.theplantlist.org), Missouri Botanical Garden's Tropicos database (http://www.tropicos.org), the Global Compositae Checklist (https://www.compositae.org/checklist), the International Legume Database and Information Service (http://www.ildis.org), and USDA's Plants Database (http://plants.usda.gov). TNRS suggests accepted or at least known taxonomic names in case of misspellings and resolves synonyms.

Before submission of the species list at the TNRS website, taxonomic names are cleaned for letters not compatible with TNRS. For taxonomic names resolution we use the default settings at the TNRS website:

- Processing Mode: Perform Name Resolution
- Match Accuracy: Allow partial matches, selected minimum threshold: 0.05
- Sources: TPL, GCC, ILDIS, TROPICOS, USDA
- Family classification: TROPICOS

Selection process

- We receive "detailed" results.
- We select the 'best estimates' only: highest agreement of provided and suggested names.

- We accept suggested names, if only the species epithet is changed and the epithet agreement is >0.85.
- We do not accept suggested names, if the genus part of the name was changed.
- When the 'accepted name' is provided by TNRS, i.e. resolving synonyms, we accept it.
- Else, if a known name is provided, we accept it.
- If we do not accept suggested changes of the names, or no accepted or no known taxonomic names are provided by TNRS, we use the original name provided to TRY.

2.2.2: Trait names

Trait names and definitions are consolidated across all datasets, based on the TOP thesaurus of plant characteristics (Garnier et al., 2016) (http://top-thesaurus.org) or the plant trait handbook (Pérez-Harguindeguy et al., 2013), if possible.

2.2.3: Names of Ancillary data

The names of all ancillary data are consolidated across all datasets.

References for trait data contributions and primary references of trait data contributed by already integrated datasets are consolidated.

2.3: Standardization of trait and ancillary data

2.3.1: Trait data

For numerical traits with more than 1000 records standardized units and values are added and trait values are recalculated if necessary.

2.3.2: Ancillary data

Most relevant ancilary-data (geo-reference, measurement date, exposition, maturity, health) are standardized:

- geo-reference: decimal degree latitude and longitude, altitude in meters
- measurement date: ISO 8601 (YYYY-MM-DD)
- exposition: natural environment, glasshouse, climate chamber, etc.
- maturity: juvenile, mature
- health: stressed, healthy

3: Data complementation 2

After the consolidation of data in the TRY database, additional trait values are derived from contributed trait data where possible, e.g. leaf nitrogen content per area can be calculated from leaf nitrogen content per leaf dry mass and leaf area per leaf dry mass if measured on the same entity; the categorical trait plant woodiness (woody, non-woody) can be logically derived from the trait 'plant growth form' (tree, shrub, herb).

4: Data quality assurance

4.1: Numerical trait values

For numerical traits with >1000 trait records errors, outliers and duplicates of trait records are identified by consistency checks of consolidated trait values across all datasets (probabilistic approach).

4.1.1 First check and correction of systematic errors

After consolidation of trait names, plant taxonomy, units and values numerical trait records are identified as systematic potential errors, if most records for this trait of a dataset are out of range across all datasets. This is a strong indication for a unit mismatch, which is corrected, if possible.

4.1.2 Second check and correction of systematic errors

After this initial quality check data are transformed to approximate normal distributions and errors and outliers are identified as z-scores (the number of standard deviations a trait record is away from the group mean). A z-score >3(4) or <-3(4) indicates low (0.3%, 0.006%) probability to be a true representative of the respective normal distribution. Z-scores are calculated based on all data of a trait and after grouping the data of a trait at species, genus and family level and according to plant growth form. As the number of trait records per group is in many cases not sufficient to calculate a robust standard deviation for the group, we use the mean standard deviation across all groups of the respective level to calculate the z-scores (see Kattge 2011). The mean z-score of all records for a trait of a dataset is used to check again for potential systematic errors within datasets, which can then be corrected.

4.1.3 Identification of individual outliers

After correction of systematic errors, z-scores are recalculated and published (https://www.try-db.org/TryWeb/Data.php#25). The maximum absolute z-sore is released with the trait data records to indicate outliers or potential errors in individual trait records.

4.1.4: Identification of duplicates

Duplicates are flagged, for same consolidated species, same consolidated trait and similar consolidated trait value, if neither geo-reference, measurement date, nor original publication indicate a difference.

4.2: Identification of errors in georeferences

Errors in geo-references are identified by comparison against a terrestrial land mask and flagged.

5: *Dataset Custodians feedback*

After a dataset has been integrated in the TRY database the dataset custodian is asked for feedback, i.e. if trait names are appropriate and values correct.

6: Reformatting for output and string consistency check

Finally data are cashed for data release and format consistency is checked in the cached data, errors in the output format (i.e. line breaks in database cells) are corrected before data are released from the database.

Data release format:
Datasets are released as machine-readable tab delimited text with 'UTF-16 Latin1 swedish ci' encoding (MySQL standard). The released datasets are organized according to the entity-attribute-value (EAV) model, where the original data (species name, trait or meta-data name, trait or metadata unit and value) are annotated and enriched by the Observation ID and consolidated information for plant taxonomy, names of traits and meta-data, consolidated units and trait values, indicators for outliers and duplicates, and the contribution reference.

7: Additional information provided at the TRY File Archive (https://www.try-db.org/TryWeb/Data.php)

- Climate, soil, biome information of TRY measurement sites
- Categorical traits relevant to determine PFTs
- Primary references