Towards a harmonization of distributed trait datasets

Florian D. Schneider[1], Malte Jochum[2], Gaëtane Le Provost[3], Andreas Ostrowski[4], Caterina Penone[2] and Nadja K. Simons[5]

[1] *Institute of Linguistics and Literary Studies, Technische Universität Darmstadt, Darmstadt, Germany;*

[2] *Institute of Plant Sciences, University of Bern, Bern, Switzerland;*

[3] *Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt am Main, Germany;*

[4] *Department of Mathematics and Computer Science, Friedrich-Schiller-Universität Jena, Jena, Germany;*

[5] *Department for Ecology and Ecosystem Management, Center for Life and Food Sciences Weihenstephan, Technische Universität München, Freising, Germany*

What are trait data?

Observation data on

phenotypic characteristics (or attributes, properties) of

species (or species occurrences, individual specimens, higher taxa).

→ Entity-Quality model*

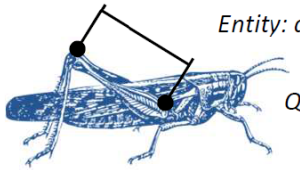
Various sub-classes of trait data have been defined:

Fitness traits, performance traits, life-history traits, morphometric traits, locomotion traits, environmental traits, phenological traits, genetic traits, behavioural traits, …

* Garnier et al. 2017 Towards a thesaurus of plant characteristics: An ecological contribution. Journal of Ecology 105:298–309

Trait-data heterogeneity

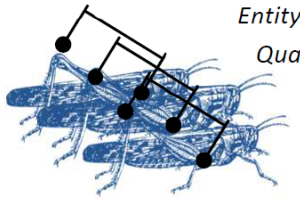
a) Measured quantitative data:



Entity: an individual x of taxon z

Quality: femur length equals 14.1 mm

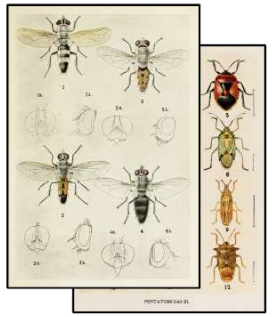
b) Aggregate quantitative data:



Entity: taxon x

Quality: average adult femur length is 12.2 ± 2.3 mm

d) Quantitative literature data:



Entity: males of taxon x

Quality: min. body length of 8 mm

c) Qualitative literature or database data:

Entity: genus x

Quality: herbivore

- small and large focal scale (system, landscape, biogeography)
- methodological heterogeneity (sampling methods, measurement methods)
- different disciplines, different taxa, different research questions
- …

→ ***Harmonization is labour-intensive; risk of misinterpretation if done on the user side***

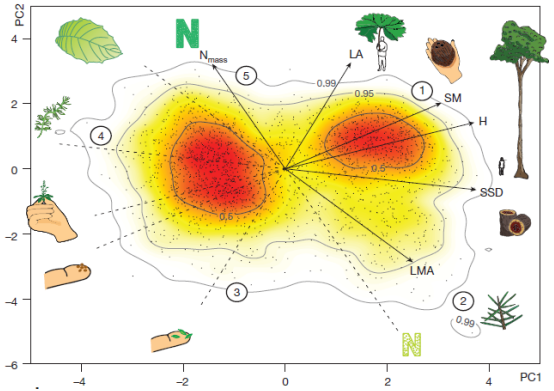
Trait-based ecology is on the rise

ARTICLE

doi:10.1038/nature16489

The global spectrum of plant form and function

Sandra Díaz[1], Jens Kattge[2,3], Johannes H. C. Cornelissen[4], Ian J. Wright[5], Sandra Lavorel[6], Stéphane Dray[7], Björn Reu[8,9], Michael Kleyer[10], Christian Wirth[2,3,11], I. Colin Prentice[5,12], Eric Garnier[13], Gerhard Bönisch[2], Mark Westoby[5], Hendrik Poorter[14], Peter B. Reich[15,16], Angela T. Moles[17], John Dickie[18], Andrew N. Gillison[19], Amy E. Zanne[20,21], Jérôme Chave[22], S. Joseph Wright[23], Serge N. Sheremet'ev[24], Hervé Jactel[25,26], Christopher Baraloto[27,28], Bruno Cerabolini[29], Simon Pierce[30], Bill Shipley[31], Donald Kirkup[32], Fernando Casanoves[33], Julia S. Joswig[2], Angela Günther[2], Valeria Falczuk[1], Nadja Rüger[3,23], Miguel D. Mahecha[2,3] & Lucas D. Gorné[1]



PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2011) **366**, 2536–2544
doi:10.1098/rstb.2011.0024

Research

Understanding global patterns of mammalian functional and phylogenetic diversity

Kamran Safi[1], Marcus V. Cianciaruso[2,*], Rafael D. Loyola[2], Daniel Brito[2], Katrina Armour-Marshall[3] and José Alexandre F. Diniz-Filho[2]

[1]*Max Planck Institute for Ornithology, Vogelwarte Radolfzell, Schlossallee 2, 78315 Radolfzell, Germany*
[2]*Departamento de Ecologia, ICB, Universidade Federal de Goiás, CP. 131, 74001-970 Goiânia, Brazil*
[3]*Imperial College London, Silwood Park, Ascot, Berkshire SL5 7PY, UK*

Trait-based ecology is on the rise

e.g.

- in „functional ecology“ linking biodiversity and ecosystem functions → functional diversity
- a cross-cutting theme in evolutionary biology and ecology,
  e.g. trait-matching in species interactions, or mapping to phylogenies
- a major data type in digitization in museum collections (fossil and current)
- promising solution to taxonomic description-gap: know species function from its form
- …

→***Relies heavily on the availability, taxonomic and regional coverage, and harmonization of data***

The open data „problem“

Demand for open data publications by funding agencies and journals, and community standards

- Response: Use of general-purpose data repositories
  - providing citable DOI & long-term stability
  - Cost-free
  - Clear re-use and re-publication policies (CC by or CC 0 licenses)
  - low thresholds for data standardisation and documentation

→ ***Lots of distributed trait-data of small and intermediate research projects, but not interoperable for use in databases and future computer-aided analysis***

(the I in FAIR guiding principles for scientific data management and stewardship, Wilkinson et al. 2016 Scientific Data 3:160018)

Readying trait-data for the semantic web

→ ***Aim: Shifting harmonization effort on data-provider side, i.e. standardise before upload!***

Incentives and tools are needed:

- Data publications must be recognised as publications
- Awareness for data accessibility and interoperability (i.e. metadata, reference to taxonomic and other ontologies)
- Consensus building on trait definitions and methods (handbooks and thesauri, ontologies)
- Standard Terminology for Trait-data labelling } ETS

OpenTraits.org

The Ecological Trait-data Standard

https://terminologies.gfbio.org/terminology/?ontology=ETS

A standard vocabulary, *i.e.* a set of terms for

1. labelling own trait-data for publication on general-purpose file servers or project-specific and internal databases (distributed data)
2. harmonizing and assembling datasets from distributed sources (aggregation)
3. building input and output interfaces for software tools and webservices dealing with trait data (tools)

gfbio Terminology Service

biodiversity exploratories functional biodiversity research

www.biodiversity-exploratories.de

Ecological Trait-data Standard Vocabulary

Florian D. Schneider, Malte Jochum, Gaëtane LeProvost, Caterina Penone, Andreas Ostrowski, Nadja K. Simons

v0.8.0, released: 29 May 2018

Glossary of terms

This defined vocabulary aims at providing all essential terms to describe datasets of functional trait measurements and facts for ecological research. The vocabulary builds on the Darwin Core Standard and it's extensions (terms of DWC are referenced thus in field 'Refines'; the full Darwin Core Standard can be found here: http://rs.tdwg.org/dwc/terms/index.htm).

The glossary of terms is ordered into a **core section** with essential columns for trait data, extensions which are allowing to provide additional layers of information, as well as a vocabulary for **metadata** information of particular importance for trait data. Another section provides defined **terms for trait definitions** to be included in the metadata or published along with the dataset.

We provide four **extensions** of the vocabulary, that allow for additional information on the trait measurement.

- the `Taxon` extension provides further terms for specifying the taxonomic resolution of the observation.
- the `Occurrence` extension contains information on the level of individual specimens, such as date and location and method of sampling and preservation, or physiological specifications of the phenotype, such as sex, life stage or age.
- the `MeasurementOrFact` extension takes information at the level of single measurements or reported values, such as the original literature from where the value is cited, the method of measurement or statistical method of aggregation.
- The `BiodiversityExploratories` extension provides columns for localisation for trait data from the Biodiversity Exploratories sites (www.biodiversity-exploratories.de).

This glossary of terms is available as

- this human-readable reference (html file), including commentaries and further definitions,
- a csv table file (the 'source' file, TraitDataStandard.csv),
- a machine readable RDF ontology file, compliant with semantic web standards accessible via an API (produced by and hosted on GFBio Terminology Server, **coming soon!**).

The rationale for developing the Ecological Trait-data standard has been cast in a paper that is available as pre-print:

> Schneider, F.D., Jochum, M., Le Provost, G., Ostrowski, A., Penone, C., Fichtmüller, D., Gossner, M.M., Güntsch, A., König-Ries, B., Manning, P. and Simons, N.K. (2018) Towards an Ecological Trait-data Standard, biorxiv.org DOI: 10.1101/328302

Table of contents

Core traitdata terms

For the essential primary data (trait value, taxon assignment, trait name), it is recommended to report the original naming and value scheme as used by the data provider. However, to ensure compatibility with other datasets, the original data provider's information should be duplicated into standardized columns indexed by appending `Std` to the column name. This ensures compatibility on the provider's side and transparency for data users on the reported measurements and facts, and enables checking for inconsistencies and misspellings in the complete dataset provided by the author. If provided, the standardized fields allow merging heterogeneous data sources into a single table to perform further analyses. This practice of double bookkeeping of trait data has successfully established for the TRY database on plant traits, for instance (Kattge et al. 2011. TRY – a global database of plant traits. Global Change Biology, 17, 2905–2935).

By linking to (public) ontologies via the field `taxonID`, further taxonomic information can be extracted for analysis. Alternatively, `taxonID` may also link to an accompanying datasheet that contains information on the taxonomic resolution or specification of the observation.

Similarly, linking to published trait definitions in public thesauri or ontologies via the field `traitID` allows an unambiguous interpretation of the trait measurement. If no online ontology is available, an accompanying data table should specify the trait definitions by making use of terms provided in the section 'Traitlist' below.

`scientificName`

go to top | direct link

	scientificName
Definition	Original character string provided as species name by the data owner (kept for reference and continuity)
Comment	Can be equal to scientificNameStd. Authors may use abbreviations, or use underscores to separate genus and species name.
valueType	character
Identifier	http://ecologicaltraitdata.github.io/ETS/#scientificname
DateIssued	2017-07-07
FirstIssuedIn	v0.8
DateModified	2018-05-29
Refines	http://rs.tdwg.org/dwc/terms/scientificName
Replaces	NA
Deprecated	NA
ReplacedBy	NA

How to apply the vocabulary?

- Use vocabulary terms as column names
- ensure the minimal information (core data)

! Undefined width of table, dataset-specific columns

! Unified, well-defined columns

! Entity-Quality pairs

How to apply the vocabulary?

- Link to accepted taxonomies and other ontologies (using URI)
- Use terms as columns names of the dataset (for two-dimensional data) or the core table + extensions (in relational databases)

c) Standardized names and URIs
(added as columns to core table)

! measurement-level definition of unit

scientificNameStd	traitNameStd	traitValueStd	traitUnitStd	traitID	taxonID	measurementID	occurrenceID
Agonum ericeti	body_length	5.87	mm	http://t-sita.cesab.org/BETSI_vizInfo.jsp?trait=Body_length	http://www.gbif.org/species/5755044	1	001
Agonum ericeti	antenna_length	3.74	mm	http://t-sita.cesab.org/BETSI_vizInfo.jsp?trait=Antenna_length	http://www.gbif.org/species/5755044	2	001
Agonum gracile	body_length	4.80	mm	http://t-sita.cesab.org/BETSI_vizInfo.jsp?trait=Body_length	http://www.gbif.org/species/5755080	3	002
...	...	...	...		..	...	...

+

! unambiguous reference to taxon

d) Extensions
(added as columns, mapped to identifiers)

Taxon

taxonID	taxonRank	order
http://www.gbif.org/species/5755044	species	Coleoptera
http://www.gbif.org/species/5755044	species	Coleoptera
http://www.gbif.org/species/5755080	species	Coleoptera
..	...	...

Measurement or Fact

measurementID	basisOfRecord	measurementMethod	measurementResolution	references	...
1	PreservedSpecimen	Digital caliper	0.1 mm	NA	
2	LiteratureData	NA	genus	https://doi.org/10.1038/sdata.2015.13	
...	...	...	...		

Occurrence

occurrenceID	sex	lifeStage	samplingProtocol	eventDate	country	habitat	...
001	f	adult	Pitfall trap	2008-06-12	DE	forest	
002	m	adult	Pitfall trap	2008-06-12	DE	forest	
...	...	...	...	...	...	...	

The Ecological Trait-data Standard

- Designed with combined expertise of researchers in:
  - Empirical biodiversity researchers (data providers)
  - Synthesis researchers (data users)
  - Biodiversity informatics researchers (data managers)
- Compatible with existing structures of major trait databases (TRY, TraitBank)
- Build on Terms of Darwin Core Standard and its Extensions
- ETS is FAIR: findable (GFBio), accessible (documentation), interoperable (URIs), re-useable (CC by)
- Open Development: Invites contributions, submissions, discussions at https://github.com/EcologicalTraitData/ETS for upcoming v1.0

www.biodiversity-exploratories.de

Package ‘traitdataform’ - harmonizing ecological trait data in R

This package assists in handling functional trait data and transferring them into the Trait Data Standard (Schneider et al. in preparation).

There are two major use cases for the package:

- preparation of own trait datasets for upload into public data bases, and
- harmonizing trait datasets from different sources by moulding them into a unified format.

The toolset of the package includes

- transforming species-trait-matrix or occurence table data into a unified long-table format
- mapping column names into terms provided in a standard trait vocabulary
- matching of species names into GBIF Backbone Taxonomy (taxonomic ontology server)
- matching of trait names into a user-provided traitlist, i.e. a thesaurus of traits
- unifying trait values into target unit format and legit factor levels
- saving trait dataset into a desired format using templates (e.g. for BExIS)

Installation

The package can be installed from Github via the ‘devtools’ package

```
install.packages('devtools')
devtools::install_github('EcologicalTraitData/traitdataform')
```

Links

Browse source code at https://github.com/fdschneider/traitdataform

Report a bug at https://github.com/fdschneider/traitdataform/issues

License

Full license

MIT

Developers

Florian D. Schneider

Author, maintainer, author

All authors...

https://ecologicaltraitdata.github.io/traitdataform/

Usage

```
data(carabids)

thesaurus <- as.thesaurus(
       body_length = as.trait("body_length",
         expectedUnit = "mm",
         identifier = "length"
         ),
       antenna_length = as.trait("antenna_length",
         expectedUnit = "mm",
         identifier = "antenna"
         ),
       metafemur_length = as.trait("metafemur_length",
         expectedUnit = "mm",
         identifier = "metafemur"
         ),
       eyewidth = as.trait("eyewidth_corr",
         expectedUnit = "mm",
         identifier = "eyewidth"
         )
)

traitdataset1 <- standardize(carabids,
            thesaurus = thesaurus,
            taxa = "name_correct",
            units = "mm"
            )
```

https://ecologicaltraitdata.github.io/traitdataform/

Take home

- Facilitate trait-data standardisation on data-provider side
- Incentivise application of FAIR guiding principles for distributed data
- Standard terminology provided by Ecological Trait-data Standard
    1. labelling own trait-data for publication on general-purpose file servers or project-specific and internal databases (distributed data)
    2. harmonizing and assembling datasets from distributed sources
    3. building input and output interfaces for software tools and webservices dealing with trait data

Thanks to:

- Co-authors Birgitta König Ries, Martin Gossner, Pete Manning
- David Fichtmüller and Anton Güntsch (GFBio)
- BExIS
- Matthias Biber, Kristin Bohn, Diana Bowler, Klaus Birkhofer, Runa Boeddinghaus, Catrin Westphal, Markus Fischer and Jens Kattge for comments on the MS; Biodiversity Exploratories members that responded to online survey
- Will Pearse, Josh Madin, and other participants of the Open Traits Workshop, New Orleans, August 2018, and organisers Brian Enquist, Rachel Galagher, Brian Maitner

www.biodiversity-exploratories.de

Terminology Service

OpenTraits.org

Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Towards an Ecological Trait-data Standard

Florian D Schneider, Malte Jochum, Gaëtane Le Provost, Andreas Ostrowski, Caterina Penone, David Fichtmüller, Anton Güntsch, Martin M. Gossner, Birgitta König-Ries, Pete Manning, Nadja K. Simons

doi: https://doi.org/10.1101/328302

Paper available at:
https://www.biorxiv.org/content/early/2018/05/31/328302

Ecological Trait-data Standard:
https://terminologies.gfbio.org/terminology/?ontology=ETS

R-package ‚traitdataform‘:
https://ecologicaltraitdata.github.io/traitdataform/

FAIR guiding principles for scientific data management and stewardship

To be FAIR, data must be

- Findable: register data and metadata, central repositories, enable web browsers, metadata and appropriate labelling
- Accessible: open access, long-term accessible (using DOI), human-readable
- Interoperable: applying global resource identifiers (URI) and terminologies, machine-readable assignment of contents
- Re-usable: clearly stating the conditions for re-use (for humans and machines), e.g. using Creative Commons Licenses

→ ***Prerequisites for a re-use in computer-aided big-data analyses and integration into the semantic web of biodiversity data***