Ten (mostly) simple rules to future-proof trait data in ecological and evolutionary sciences

Alexander Keller1, Markus Ankenbrand2, Helge Brueelheide3, Stefanie dekeyzer4, Brian J Enquist5, Mohammad Bagher Erfanian6, Daniel Falster7, Rachael V Gallagher8, Jen Hammock9, Jens Kattge10, Sara D Leonhardt11, Joshua Madin12, Brian Maitner13, Margot Neyret14, Renske E Onstein3, William D Pearse15, Jorrit H Poelen16, Roberto Salguero-Gomez17, Florian D Schneider18, Anikó B Tóth19, and Caterina Penone20

1Cellular and Organismic Networks, Faculty of Biology, Ludwig-Maximilians-Universität München
2Center for Computational and Theoretical Biology, Julius-Maximilians-Universität Würzburg
3German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig
4Flanders Marine Institute (VLIZ)
5Department of Ecology and Evolutionary Biology, University of Arizona, Tucson
6Department of Biology, Faculty of Science, Ferdowsi University of Mashhad
7Evolution & Ecology Research Centre, University of New South Wales Sydney
8Hawkesbury Institute for the Environment, Western Sydney University
9National Museum of Natural History, Smithsonian Institution
10Max Planck Institute for Biogeochemistry, Jena
11Plant-Insect Interactions, TUM School of Life Science Systems, Technical University of Munich
12Hawaii Institute of Marine Biology, University of Hawaii at Manoa
13University at Buffalo, Department of Geography
14Senckenberg Biodiversity and Climate Research Center (SBik-F)
15Department of Life Sciences, Imperial College London
16Ronin Institute for Independent Scholarship
17Department of Zoology, University of Oxford
18ISOE - Institute for Social-Ecological Research
19School of Biological Earth and Environmental Sciences, University of New South Wales
20Institute of Plant Sciences, University of Bern

July 23, 2022
Ten (mostly) simple rules to future-proof trait data in ecological and evolutionary sciences

Alexander Keller¹,⁵, Markus Ankenbrand², Helge Bruelheide³,⁴, Stefanie Dekeyzer⁵, Brian J. Enquist⁶, Mohammad Bagher Erfanian⁷, Daniel Falster⁸, Rachael V. Gallagher⁹, Jen Hammock¹⁰, Jens Kattge⁴,¹¹, Sara D. Leonhardt¹², Joshua Madin¹³, Brian Maitner¹⁴, Margot Neyret¹⁵, Renske E. Onstein⁴, William D. Pearse¹⁶, Jorrit H. Poelen¹⁷,¹⁸, Roberto Salguero-Gomez¹⁹, Florian D. Schneider¹⁵,²⁰, Anikó B. Tóth²¹, Caterina Penone²², &

¹ Cellular and Organismic Networks, Faculty of Biology, Ludwig-Maximilians-Universität München, 82152 Planegg-Martinsried, Germany
² Center for Computational and Theoretical Biology, Julius-Maximilians-Universität Würzburg, 97070 Würzburg, Germany
³ Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, Halle (Saale), 06108, Germany
⁴ German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstrasse 4, Leipzig, 04103, Germany
⁵ Flanders Marine Institute (VLIZ), InnovOcean Site, Wandelaarkaai 7, 8400 Oostende, Belgium
⁶ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, United States and The Santa Fe Institute, Santa Fe, NM 87501
⁷ Department of Biology, Faculty of Science, Ferdowsi University of Mashhad, Mashhad, Iran
⁸ Evolution & Ecology Research Centre, University of New South Wales Sydney, Australia
⁹ Hawkesbury Institute for the Environment, Western Sydney University, Richmond NSW, Australia
¹⁰ National Museum of Natural History, Smithsonian Institution, Washington, DC, USA
¹¹ Max Planck Institute for Biogeochemistry, Jena, Germany
¹² German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
Plant-Insect Interactions, TUM School of Life Science Systems, Technical University of Munich, Hans-Carl-von-Carlowitz-Platz 2, 85354 Freising

Hawai‘i Institute of Marine Biology, University of Hawai‘i at Manoa, Kāne‘ohe, Hawai‘i, USA

University at Buffalo, Department of Geography, 120 Wilkeson Quadrangle, 14261 Buffalo, NY, United States. Department of Environment and Sustainability. Department of Environment and Sustainability, 602 Clemens Hall, North Campus, 14260 Buffalo, NY United States

Senckenberg Biodiversity and Climate Research Center (SBik-F), Frankfurt, Germany

Department of Life Sciences, Imperial College London, Ascot SL5 7PY, United Kingdom

Ronin Institute for Independent Scholarship, Montclair, New Jersey, United States

Cheadle Center for Biodiversity and Ecological Restoration, UC Santa Barbara, Santa Barbara, California, United States

Department of Zoology, University of Oxford, 11a Mansfield Road, OX1 3SZ, Oxford, UK

ISOE - Institute for Social-Ecological Research, Frankfurt am Main, Germany

Centre for Ecosystem Science, School of Biological Earth and Environmental Sciences, University of New South Wales, Sydney Australia

Institute of Plant Sciences, University of Bern, Switzerland

Corresponding author: Alexander Keller, Cellular and Organismic Networks, Faculty of Biology, Ludwig-Maximilians-Universität München, 82152 Planegg-Martinsried, Germany, keller@bio.lmu.de

& Corresponding author: Caterina Penone, Institute of Plant Sciences, University of Bern, Switzerland, caterina.penone@ips.unibe.ch

Running head: Ten rules to future-proof trait data
Abstract

1. Traits have become a crucial part of ecological and evolutionary sciences, helping researchers understand the function of an organism's morphology, physiology, growth and life-history, with effects on fitness, behaviour, interactions with the environment, and ecosystem processes. However, compiling and analysing trait data comes with data-scientific challenges due to the complex nature of trait data.

2. We offer 10 (mostly) simple rules, with some detailed extensions, as a guide in making critical decisions that consider the entire life cycle of trait data.

3. This article is particularly motivated by its last rule, i.e. to propagate the good practices, and has the intention of bringing awareness of the different facets of a trait's life cycle to the community.

4. When it comes to working with trait data, we gain particularly as an interdisciplinary community of field biologists, synthesis ecologists, computer scientists and database managers. We hope these basic guidelines can be useful as a starter for active communication in disseminating such integrative knowledge and how to make trait data future-proof.

Keywords: data life cycle, data science, FAIR principles, good practices, metadata, open science, phenotype, trait data
Introduction

As early as 300 BC, Greek philosophers such as Theophrastus forged the first formal systems defining and classifying organisms by their combination of morphological, physiological, behavioural, and phenological characteristics, i.e., their traits (Weiher et al., 1999). Having knowledge about an organism’s traits often allows deeper understanding of its life-history, behaviour, fitness, interactions and potential responses to and effects on ecosystem processes (Violle et al., 2007). Trait-based research questions in ecology and evolution often focus on the processes that drive trait patterns (Sutherland et al., 2013), and view the organism as a functional rather than a taxonomic unit. Traits offer a bridge between different dimensions: from organismal biology, e.g. population abundance (Webb et al., 2010), species distribution (Sporbert et al., 2021), and phylogeny (Junker et al., 2015; Tucker et al., 2018) to ecosystem functioning (Lavorel & Garnier, 2002; Wright et al., 2016).

Traits are commonly defined as a measure of an entity (Garnier et al., 2017), where the entity can be the whole individual, or a specific organ or tissue (e.g. a tail fin, a leaf or wood) and the quantity is an observable characteristic of that entity (e.g. the length of a fish, its colour, the length of a tail fin, or the density of wood tissue). Together, all traits of an individual organism represent its phenotype resulting from the historical evolution of the genotype and potential current interactions with the environment. Therefore, a trait record can inform not only about the entity that was observed (e.g. taxonomic classification or age) and the quantity/characteristic that was measured, but also about the environment in which the individual has developed that trait (e.g. where a fish was caught, where a tree lived or the soil depth where an invertebrate was observed)(Kattge et al., 2011).
There are many ways to define and measure traits of organisms (Kearney et al., 2021; Walker et al., 2022). For example, a plant leaf can be described by several hundred diverse individual measurable characteristics. These include surface area, sodium concentration, phenology, and maximum photosynthetic rate (see e.g. Kattge et al., 2020). On the one hand, different traits of an individual are often correlated, e.g., for a tree to grow tall it usually needs a thick stem. It is essential to recognise these correlations in the way the data are collected (e.g. on the same tree) and stored. In this case, for a trait record to be meaningful it needs to be connected to a complex combination of multiple measurements. In contrast, a trait record can also be rather simple, if the given trait is well defined, if it depends only on the genotype and is not affected by current interaction with the environment.

In essence, trait data are a special kind of data: they are diverse, often not well defined, relatively simple or potentially complex, independent of one-another or correlated, and range between cheap and costly to collect. However, they are very informative as they represent the evolutionary adaptation or developmental acclimation of the individual organisms to their environment and allow for quantitative and predictive ecology and biodiversity research. Therefore - if collected, stored and published in a meaningful way - organismal trait data have an extraordinary value for reuse, which is e.g. indicated by the >20,000 data requests to the TRY Plant Trait Database since 2015 (Kattge et al., 2020).

To enable the reuse of these trait data beyond their original research campaign, to make them meaningful in other contexts and to avoid data degradation, observation records must be clearly defined, where possible the environmental and mereological context given, as well as provenance and protocols for collection documented (Michener, 2006). Recent efforts to expand trait knowledge across the Tree of Life (Gallagher et al., 2020) call for datasets which
are open and FAIR (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al., 2016), key principles at the heart of the emergent Open Science movement (Nosek et al., 2015). Global and local datasets of organismal traits have rapidly grown in scope since the 1990’s (e.g. Herberstein et al., 2022; Kattge et al., 2020; J. S. Madin et al., 2016). However, these datasets bear a variety of new challenges linked to harmonisation, biases, expertise and communication (Salguero-Gómez et al., 2021). These challenges result in a major trade-off between investing in collection of new trait data or reusing open trait data (Westoby et al., 2021). Indeed, many studies in trait-based research reuse available trait data or collect additional trait data and/or assemble new data (e.g. examples in Kattge et al., 2020). Thus, these studies also often involve linking to different types of data, which requires interoperability between datasets (Feng et al., 2022; Gallagher et al., 2020).

These key aspects are just a few of many dimensions illustrating how and why researchers not only have to make biological decisions, but also a wide range of data-science decisions when working with traits. Multiple complexities of trait data structure and manipulation are not obvious at a first glance (Michener, 2006). For instance, there is sometimes confusion and lack of awareness of trait standards, measurement units, and trait data are particularly prone to errors in recording, translation, and understanding. By offering a larger perspective, the trait data life cycle can help clarify these confusions and inform about good practices when working with trait data (Fig.1). In this article, we bring awareness to some common pitfalls in the usage of trait data and offer 10 rules for making critical decisions that consider the entire life cycle of trait data. We start each rule with a general and simple statement and develop the complexity of each rule within more detailed subsections.
**Rule 1: Define your trait**

Let your study question or hypothesis determine both the trait(s) to be used and how those traits are collected and analysed. Clear upfront definitions of traits will avoid bias through e.g. confusion of scales and definitions, data gaps or inclusion of inadequate traits (Dawson et al., 2021; González-Suárez et al., 2012; Hulme et al., 2013; Messier et al., 2017).

**Follow your hypothesis:** Increasingly, trait data describing organisms of interest are publicly available for reuse. Vast availability extends the potential scope of what is possible with limited resources (e.g. Falster et al., 2021; Kattge et al., 2020). However, when reusing trait data, we relinquish control of what variables are collected, which species are sampled, and
the methods used for collection (Koricheva et al., 2013). Undirected fishing expeditions for
traits can yield large datasets, but these may not necessarily be appropriate to answer a given
research question, for various reasons (e.g. coverage, distribution, meaningfulness and
resolution, Violle et al., 2015). Furthermore, the wealth of available trait data may distract
from initial hypotheses, risking random exploration of the available traits and fishing for
significant relationships without a clear focus. Thus, trait selection and collection should in
most cases be primarily tethered to a concrete hypothesis, not defined by availability of
existing data. This rule does, however, not completely exclude extensive data exploration and
data-driven discovery within a given range as relevant to the research question and
subsequent streamlining (Violle et al., 2015).

Consider the scale: Research questions define the appropriate hierarchical level for sampling:
a continental-scale study of thousands of species may treat the intra-specific variation as
statistical noise. In contrast, this variation may be the study focus in locally scaled projects.
There is no "correct" scale, either in terms of spatial grain (e.g., km², m²), temporal duration
(e.g., seconds, years), or taxonomic coverage (e.g., clade, species, population or individual),
but not every scale will be appropriate for every question. So, when defining the traits of
interest, it is important to define the scale at which these need to be collected or aggregated to
match the research question (Messier et al., 2017).

Be aware of existing trait definitions and homologies: Much effort has already gone into
creating definitions and protocols for traits collection. Yet, trait naming and corresponding
definitions may differ between studies and trait databases (Ankenbrand et al., 2018; Dawson
et al., 2021; Kunz et al., 2022). For example, specific leaf area (SLA) and leaf mass per area
(LMA) are essentially the same trait, one being the inverse of the other. Similarly, for discrete
traits, values may differ between resources (e.g. “therophyte” and “annual” are synonyms).

Furthermore, when comparing traits and trait states across organisms, it is important to be aware of the ‘homology’ of the character. Homologous traits share similarity of structure, physiology, or development (often by common evolutionary ancestry), whereas non-homologous (or analogous) characters may perform a similar function, but differ in structure, physiology, or development.

Be pragmatic: In a perfect trait research world, we could measure or retrieve the exact traits for the exact scale and organisms needed to answer our specific question. This vision is rarely real. Instead, we often need to work with proxies of traits that are difficult to measure (e.g. reproductive output is often used as a proxy for fitness (McGraw & Caswell, 1996)) or are incomplete in a database (e.g. diet or behavioural traits are less complete than morphological traits, Oliveira et al., 2017). There is a common understanding of these technical or financial limitations in the scientific community; ultimately, we need to be pragmatic to advance research questions. However, it is crucial to explain and justify the choice of traits, especially when these are used as proxies or “best available data”.

Rule 2: Do not reinvent the wheel

Build on existing trait resources to reduce the likelihood of redundancy and ensure compatibility with current data. The decision when to collect new trait data is generally based on the research question, the scope of the analysis (e.g. local, global) and the availability of the existing data. Financial and geographic constraints may also influence the decision to use current trait data instead of embarking on a measurement campaign. However, the existing trait data must be ‘fit for purpose’ to avoid compromising the capacity to answer the research question.
Check public data sources: Most data probably exist decentralised as individual trait datasets in the form of raw data attachments to publications, data papers, or data uploads to unspecific public databases (e.g. Zenodo https://zenodo.org, DataDryad https://datadryad.org). However, these datasets can be challenging if not registered at central hubs (e.g. https://opentraits.org). To counter this challenge, dedicated centralised trait databases have been and continue to be developed (e.g., TRY (Kattge et al., 2020), Encyclopedia of Life (EOL) TraitBank (Parr et al., 2015), Marine Traits Portal of the World Register of Marine Species (WoRMS) (Marine Species Traits editorial board), AusTraits (Falster et al., 2021). Common to these efforts is the fact that they contain already harmonised, error-checked, and standardised values. These resources usually provide user-friendly interfaces for searches and dynamic, up-to-date aggregations of data. Particularly for studies of larger scale (e.g. many taxa, many bioregions), it often makes sense to consult these existing big databases and data registries.

Identify data origins: Trait data are not always raw or first-hand: they can be created and perhaps aggregated from original observations and measurements (e.g. Kattge et al., 2020), but also mobilised from literature or undigitised legacy trait data (e.g. Parr et al., 2015), synthesised as imputed trait data (e.g. Penone et al., 2014), reused from data publications (e.g. Kattge et al., 2020), or mined from texts with automated algorithms or other contexts (Thessen et al., 2018). Thus, when reusing trait data, it is important to check and report information about the source to downstream analyses and subsequent publications. Importantly, providing this information also gives credit to the original trait data collectors.

Fill the gaps: Existing databases are taxonomically and biogeographically biased, ‘gappy’, and traits assigned to the same species are rarely collected in the same locations or conditions.
Despite the presence of large trait databases, new trait collections continue to remain valuable. When collecting new data, we encourage researchers to first check available trait databases, identify such gaps, and contribute to the wider trait community by filling these gaps even if this collection goes beyond the current project. Additional traits may be easily collected with little extra effort yet provide the possibility to close gaps in trait coverage. Filling gaps may be especially valuable in biodiverse but hard to access regions (Etard et al., 2020), for rare but functionally important species which may be less likely to have traits documented (Leitão et al., 2016), or for threatened species which will benefit from functional approaches to their conservation (Gallagher et al., 2021).

**Rule 3: Rely on measurement protocols and know your units**

To ensure comparability, data reuse and synthesis, relate primary measurements of your traits to the wider body of published trait data. Conform your measurement procedures to existing trait measurement protocols, or – if no such standard protocols exist – document with precision and build upon unambiguous concepts. Also, confusion and errors in terms of recording and reporting of units can be propagated through large trait compilations. Thus, define your units clearly; they are essential for harmonising different trait data sets, approximations and uncertainties.

*Beware of ambiguities:* In most cases, researchers of a domain (e.g. plants) have adopted sufficiently specific trait definitions to allow comparison of widely used measurements and enable synthesis within the field. However, some difficulties in measurement remain. To illustrate, specific leaf area (SLA) is the ratio of the surface area to leaf biomass of an individual leaf. However, the application of the concept of SLA may differ between research
contexts, because the value reported may relate to measurements of individual leaves or an
average of all leaves on the shoot, for one or both sides of the leaf, including or excluding the
petiole, and focus on the leaf or leaflet (example taken from Garnier et al., 2017). While fully
justified in the specific research setting, identifying and dealing with semantic
disambiguation is a major challenge in trait-based synthesis.

Adhere to existing standards: Methodological handbooks for trait measurements have been
proposed, e.g. for plant (Cornelissen et al., 2003), macrofungi (Dawson et al., 2019) or for
terrestrial invertebrate functional traits (Moretti et al., 2017). These handbooks provide
precise, taxon-specific definitions and recommended methods for trait measurement,
measurement precision and replication. They also provide considerations and warnings of
misconception and error, and point to the key literature debating the methodology. Taking
formalisation of trait concepts even one step further are thesauri of trait concepts (Garnier et
al., 2016, 2017), e.g. TOP. The bottom line is: research that provides original trait
measurements should consider existing measurement protocols, make an explicit choice, and
describe any deviations from or additions to protocols. When such handbooks do not exist it
is good practice to also report accompanied the specific measurement protocol in the
metadata (see Rule 4). For instance, how the length of a fish has been measured and if
potential extensions of the tail fin were taken into account.

Understand your units: Trait data are necessarily ‘rich in dimensions’. That is, trait data may
require multiple SI base units and may also be measured and reported in multiple alternative
configurations of units. For example, photosynthetic rate involves three SI base units e.g.
mass per area per time, and is often reported in units of µmol CO₂m⁻²s⁻¹ or an amount per unit
area per unit time. Simple measures of size, area, and time are often reported in differing
units, though all can be related to more fundamental base units. All metric trait data can be
reduced to the seven base units as defined by the SI standard (m, g, s, K, A, cd, mol).

Significant data management effort is needed to record units accurately, preserve them through metadata, and convert them correctly to avoid propagating errors (Calder, 1982).

**Rule 4: Context is crucial**

Always pair your data points with metadata. Sampling protocols ideally also define metadata that can be considered as covariates of the measurement procedure or inform the user about the provenance of the trait data. Together with the trait measurements, metadata defines an observation and its context (Madin et al., 2008). While such metadata may already be important for the proximate research question, it further helps future users to better understand and reproduce the methods and correctly interpret the trait values. The reuse value of existing datasets increases with the quantity and quality of metadata, so that datasets with sufficient context information are more likely to be reused in future synthesis analyses or included in larger databases.

**Define at least the minimum context:** Some metadata are considered essential and universal between all domains, such as unique ID for observations to cross-references to other measures, geolocation, time and date, life stage (e.g. juvenile), health status, scale (e.g. leaf), habitat type (e.g. natural habitat) and measurement details (e.g. following standards, devices used, etc.) (Schneider et al., 2019). Further metadata must include the source and authorship of the trait measurement. To permit effective reuse, authorship attributes should include the original data collectors and the databases where these data were gathered, as they may have undergone processing therein (Rule 2).
Cover the domain-specific standard, if possible: Deciding which further metadata to collect often involves a trade-off between which data are commonly collected in a specific domain and the time and expense involved in collecting or processing such data. Metadata preferably include detailed documentation and code of how traits were measured (e.g. manufacturer and version of devices used) and processed (e.g. standardizations or species means). We recommend checking existing well-used datasets and databases of the specific domains before collecting new trait data to determine which common metadata should be covered.

Link to other data by metadata: A good practice is to directly link the data with publications (e.g. by DOI) for the scientific context and further information in the materials and methods sections, as well as identification of trait data providers (e.g. by ORCID) to provide opportunities for feedback and requests for additional information. Often traits are measured with the objective to also collect other data, such as ecosystem function (e.g. Bongers et al., 2021) or species composition (e.g. Breitschwerdt et al., 2018). In these cases, functions measured, and species composition recorded would be part of the metadata, or links to that data in other repositories.

Rule 5: Structure trait data

Do not underestimate the importance of the structure of your dataset. It might sound trivial at first glance to think about how to structure the data, but poorly structured data may become a nightmare to work with in downstream analyses, or to reformat for publication, deposit in a public database, or synthesise in meta-analyses. It thus makes sense to consider structural aspects even in the early stages of a project using traits.
**Minimum trait data standards:** The minimal, essential information for a trait record includes taxon name, trait name, observation ID, trait value, unit (if applicable), and source. Several standards are available to help structure this minimal information set (Fegraus et al., 2005; Kattge et al., 2011; J. Madin et al., 2007; Parr et al., 2015; Schneider et al., 2019; Wieczorek et al., 2012). A good start for data structuring is to adopt one of these well-established schemes.

**Preserve metadata and further observations recorded together:** A complex aspect of structuring trait data is how to keep metadata and data links. This linkage is critical for various questions that address intra-specific variation, derived traits, or multivariate modelling on individuals. A good practice is to have unique identifiers for every entity that requires relations. Using such IDs allows linking to other traits and further data types, like community records of the plot in which the individual was found, DNA barcodes or information on the experiment that involved the individual. One approach is the Extensible Observation Ontology (OBOE) (Madin et al., 2007), which various large databases have also adopted (e.g. TRY, Kattge et al., 2020).

**Apply version control:** The process from gathering to analysing trait data is long, which might result in many different versions of a trait dataset. Thus, it is important to keep track of these different versions. It is recommended to keep the raw trait dataset and the processed trait dataset used for analysis as separate copies. Version names should be structured in a logical way, e.g. combining project acronyms, researchers’ initials, short names of the trait dataset, version numbers, file status and/or dates. Use of versioning or change control systems like Git (Spinellis, 2012) is highly recommended to keep track of changes to data.
Also, conventions that include a current (e.g. my_data_current.tsv), as well as versioned copy (e.g. my_data_1.0.tsv) help automated systems by providing a stable interface for indexing.

**Rule 6: Check and process**

Rigorously check your data quality, integrity and compatibility during each step of data processing. Trait-based analyses, mainly when data are consolidated from different sources, can harbour various inherent incompatibilities that may cause biases and severe scientific misinterpretations. For trait compilations, data usually need to be harmonised, subset, transformed, derived and/or aggregated into comparable formats to fit the research question.

It is crucial that steps are wherever possible scripted, and by that directly reproducible, and where not that manual steps are well documented.

**Harmonise trait data:** If trait data originate from multiple sources, each source may identify the same entities or concepts differently (Kunz et al., 2022). Harmonisation is crucial to reconcile equivalent entities and explicitly connect related entities by “similar” or subclass relationships. Ideally, these entities or concepts should be identified by standard identifiers (see Rule 5). Manual harmonisation may be necessary to detect and reconcile spelling variations, before text strings are mapped to identifiers. But for common classes of data there are a variety of services available that allow automated and reproducible harmonisation, e.g. for taxonomic names (B. Boyle et al., 2013; Chamberlain & Szöcs, 2013; Global Names Architecture.) (reviewed by Grenié et al., 2022), units (Gama, 2014) or geographic names (Boyle et al., 2022). Other covariates and categorical trait values may be semantically reconciled where appropriate ontologies exist (Kunz et al., 2022; Violle et al., 2015).
Filter where needed and double-check data contexts: Not all trait data are equally suitable for all purposes. Erroneous or duplicate data points need to be identified and removed before analyses (Kattge et al., 2020). As with other kinds of data, outlier detection and data visualisation provide valuable methods for the detection of such data errors (de Bello et al., 2013). For trait data mainly compiled from different sources, other reasons may also render data points inappropriate. For example, if metadata suggest that the observation is from a cultivated occurrence such as a botanical garden, greenhouse, zoo or farm, values might not be representative for wild specimens (Gering et al., 2019). Observations stemming from introduced or experimental populations may violate assumptions as well. Observations can be collected from different subsets of the population (e.g. adult vs juvenile, healthy vs diseased), at different times of year (e.g. breeding season vs overwintering), in different contexts (e.g., experimental temperature treatments), and using other protocols. It is essential to exclude unsuitable observations, usually by making use of the associated metadata.

Derive traits from raw data: Research questions may concern composite or derived traits, such as ‘hand-wing index’ (the ratio between wing length and body length in birds). Where possible it is advisable to calculate derived traits directly from the raw data to avoid bias and allow for new calculations. This procedure may not always be possible because of data gaps; in this case the calculation can be done at a higher level (e.g. at the taxonomic level of interest).

Aggregate trait data: Trait data may come at different levels of resolution. A dataset may include multiple measurements per individual, per population, or species, or even higher taxonomic levels. Such structures may imply first aggregating (e.g. to calculate average trait values) within individuals, then populations, then species derived from a particular data
source, and then across data sources if the species is represented in several of these 
(Schneider et al., 2019). The way trait values were aggregated has to be precisely described, 
in particular when data transformation is involved. For example, when it is desired to express 
leaf area on a log scale, it might make a big difference to take the log before or after 
aggregating the data. Importantly, if multiple successive steps of aggregation are necessary, 
there is the need to properly measure the uncertainty of the final trait values and assess the 
effect of aggregation on the results and conclusions, e.g., by sensitivity analyses with 
different aggregated datasets (Kunz et al., 2022).

**Transform and standardise:** Likewise for other types of data, transformations such as the 
natural logarithm or square root may be essential to conform with the requirements of 
analytical models. Beyond these, data challenges include how to combine binary, categorial 
and continuous traits into the same analysis (de Bello et al., 2021). It is thus very useful to 
explore transformation and standardisation options applied in current trait scientific literature. 
For example, to compare the effects of several explanatory traits on a specific response in a 
linear model approach, values can be standardised for each trait to range between 0 and 1, or 
by scaling their mean to 0 and their standard deviation to 1 or 0.5 (in case of making 
continuous traits comparable with categorical traits, (Gelman, 2008)).

**Work with relative errors:** Units are important when we deal with approximations, 
uncertainties, and errors (Langtangen & Pedersen, 2016). As an example, a trait measurement 
where the length scale is typically measured in mm and has an approximation of 12.5 m to 
the exact value 12.52 m with an error of 0.02 m. Switching units to mm leads to an error of 
200 mm. A study working in mm would report $2 \times 10^2$ as the error, while a study working in 
m would report 0.02 as the error. As a result, knowing the original measurement units is
essential and the downstream use of the unitless relative error is recommended (Langtangen & Pedersen, 2016).

**Rule 7: Know the limitations**

Follow the latest developments for best practices in trait data analyses. As the downstream part of data analysis is directly linked to the research question, generalisation of analytical methods is rarely possible. Given the diversity of research questions, the analytical steps can thus broadly diverge. However, some best practices can help to avoid common mistakes made with trait data due to their nature.

*Mind the level:* Traits encompass different levels: organ, individual, population, species, and community (Violle et al., 2007), and this structure determines the tools used for data analyses. For instance, trait-environment relationships investigated at the species or community level require different analysis types (*e.g.* comparative models vs. simple linear models, see below). It is important to choose the appropriate level early in the research program to fit the target scientific question and to be able to analyse the data correctly.

*Be careful with categorical traits:* By having fewer possible values, categorical traits might influence the outcome of statistical analyses. In particular, categorical traits might have disproportionate effects when aggregated with continuous traits in a common metric (*e.g.* functional diversity). For instance, when computing Gower distances of traits between species pairs, a categorial trait treated as a binary trait (0/1) for each possible categorial value will only result in distances of 0 or 1. In contrast, for a continuous trait, only the species pair with the highest/lowest trait values will have a value of 1. When averaging Gower distances of these two traits, the categorial one will have more influence than the continuous one.
The joint use of continuous and categorical traits thus needs particular scaling (Gelman, 2008). In general, it is important to be aware of this issue and to account for it with existing proposed methods (see de Bello et al., 2021).

**Do not confuse trait signals for richness and abundance signals:** Metrics aggregating traits at the community level (e.g. functional diversity or community-weighted means - CWM), are influenced by the richness, the abundance of species and the overall species composition of the community. In general, choosing metrics unrelated to abundance (e.g. unweighted means) or null models (Hawkins et al., 2017) is necessary to separate species abundance, composition or richness signals from trait information.

**Handle correlations with care:** Traits are often correlated causing issues with statistical analyses (e.g. collinearity in linear models when traits are explanatory variables). Often, these correlations are due to biological constraints (e.g. allometries), or “strategies” (Díaz et al., 2016). In some cases, it is possible to use multivariate analyses (e.g. principal component analysis) to reduce trait space dimensionality by use of axes in further analyses. Keeping the original traits is advised when possible, as the biological meaning can be lost when using multivariate principal component axes. On another level, since the start of trait analyses at the community level, numerous metrics have been proposed to characterise functional diversity, analogous to species diversity. When choosing a metric for a specific analysis, it is essential to be aware that several of these functional diversity metrics are highly correlated among themselves and with species richness (see de Bello et al., 2021).

**Consider correction for phylogenetic relatedness:** When analysing data from multiple species in trait-trait correlations or using traits as responses and depending on whether the
focus of the question is ecological or evolutionary, it may become necessary to account for
the fact that species are not independent units (Pillar et al., 2021). The whole field of
comparative analyses tackles this issue and proposes tools to account for phylogenetic
relatedness in trait analyses (e.g. see Garamszegi, 2014), though care should be taken to
justify the use of such analytical corrections relative to the aims of the research question
(Freckleton, 2000; Westoby et al., 1995).

Account for variability and uncertainty: Very often, intra-specific data are aggregated at the
species level to obtain one trait value per species. All information on variability and
measurement uncertainty is then lost. When information on variability is available and
reasonable in the scope of the study, it is possible to include it, e.g., by weighting species-
level measures in functional diversity metrics (de Bello et al., 2021) or by explicitly including
it when inferring trait evolution across lineages (Kostikova et al., 2016; Purschke et al.,
2017). This can be an issue, especially if variability is phylogenetically structured
(Garamszegi, 2014; Paterno et al., 2018).

Rule 8: Publish trait data

Openly publish trait data to facilitate answering yet unknown questions beyond their original
study, lay the groundwork for understanding ecological processes beyond clearcut niches
(Elton, 1927; Schneider et al., 2019) and democratise access to valuable trait datasets
(Soranno et al., 2015). Each data point of trait measurements has a considerable value for the
scientific community and future generations working on trait-related research questions.

Consider the stakeholders: As our scholarly processes evolve to find better, access, integrate,
and reuse scientific data, we face the communal task of treating trait datasets as first-class
research citizens. However, doing so is not easy as it involves different stakeholders: publishers have to make their publications open and FAIR (Wilkinson et al., 2016), scientists have to improve their skills to publish and reuse datasets, and funding agencies have to find ways to reward exemplary projects. A welcome development is that many publishers now consider trait data papers (e.g. Falster et al., 2021; Guerrero-Ramírez et al., 2021; Tobias et al., 2022; Vandvik et al., 2020), which allow for a detailed methodological and context description, open access, and at the same time, accreditation of trait data collectors by citations.

Accept the additional responsibility: Erroneous data might not only bias a current project but also the future works of others. There are currently no common established practices on how peer review is also extended to trait data. A way to ensure that a dataset conforms to community standards is to submit it to an established curated database (e.g. TRY (Kattge et al., 2020) for plant traits; Coral Traits (Madin et al., 2016)).

Aim for redundancy: Public trait data suffer from the same generic issues as other data, e.g. hardware failures, linkrot (URLs not fully reliable) or content-drift (content changes, but URLs do not) (Koehler, 1999). To mitigate such issues and reliably preserve data in the long term, data can be submitted to multiple repositories, e.g. beside trait databases, also in general storage platforms such as FigShare (https://figshare.com) or Zenodo (https://zenodo.org). This procedure however requires systematic methods to track changes and separately citable versions e.g. by unique DOIs.

Make data accessible for machines and humans: Trait data often carry large numbers of references and republished original data. Tracing both call for systematic and automated
methods (Elliott et al., 2020) that rely on machine-readable data. This issue becomes particularly relevant when reporting already published data, as future studies might run the risk of using the same trait from independent sources, thus resulting in pseudo-replication of measurements. Also, in order to facilitate reuse, machine-readable, non-proprietary data formats should be preferred (i.e. plain csv over excel or pdf).

Register trait data: Independent of the choice of actual data deposition, it is of great importance that datasets are also registered in a trait data registry (e.g. https://opentraits.org) to allow fellow scientists to easily find the data.

Rule 9: Review data and code like the research itself

“Be polite, fair, specific, and constructive”: best practices in peer review have already been discussed in detail (Roberts, 2004; Spigt & Arts, 2010), but can perhaps be summarised with such statement. A reviewer should provide information for the editorial team to decide; this process also applies to the data. Specifically for trait-based papers, it includes considering the entire life cycle of the trait data:

1. First, are the traits themselves appropriate for the question being asked? It should be considered how these traits have been used in the past and how they fit into biological theory. Are they being contextualised appropriately, and are they fit for the purpose to which they are being used?

2. How were the data collected? Does the protocol conform to current standards, bearing in mind that the purpose of many papers is to improve standards and so they may not? Is the collection of new data well justified? Are units and metadata properly provided?
3. How were the data processed? Consider not just quality assurance and quality control but also how the traits were generally processed into a format that can be analysed. Ensure that relevant code (ranging from simple cleaning scripts to full-featured analysis pipelines and models) is openly available, functional, and conforms to community standards. Guidelines for source code review like those by the rOpenSci community ([https://ropensci.org/software-review](https://ropensci.org/software-review)) or The Journal of Open Source Software ([https://joss.readthedocs.io/en/latest/reviewer_guidelines.html](https://joss.readthedocs.io/en/latest/reviewer_guidelines.html)) can be helpful.

4. Ensure that a distinction is made between the ‘raw’ data collected, and the ‘clean’ final product used for analysis, and that both forms of data are released.

5. Ensure that all data sources are appropriately cited, and the provenance of data is explained.

While these standards apply to all manuscripts, a distinction should be made between data papers and research papers. Standards for data release are the same for both types of publication. However, the expectations for demonstrating data cleaning and conceptual novelty likely differ between the two (but are, ultimately, determined by the journal’s guidelines).

**Rule 10: Propagate the good practices**

Inform the community and the next generation of ecologists about the issues discussed here and in other resources (*e.g.* de Bello et al., 2021). As we saw above, there are several issues to be aware of when collecting, handling, analysing, and publishing trait data, *i.e.* the life cycle of trait data. Some of them might be straightforward; others require more technical knowledge or extensive reading of existing resources. In many cases, good procedures are not applied simply because ecological or evolutionary scientists are unaware that they exist, *e.g.*
when it comes to trait quality control or using a standard structure. Educating can be done from the small scale of an informal conversation with a colleague, to teaching a large undergraduate class, up to participating in the collective creation of open access materials in several languages, accessible from any part of the world. Integrating trait-data-specific sections into ecology textbooks and modules in ecology courses could become a standard practice, which will undoubtedly be made easier in the future by disseminating open-access material by the trait scientific community. You are welcome to use these ten rules as a starter when teaching your students, colleagues and friends. For diving deeper into each single topic, we encourage you to also check more comprehensive resources such as the Handbook of Trait-Based Ecology (de Bello et al., 2021) or the activities of the Open Traits Network (https://opentraits.org) (Gallagher et al., 2020).

**Train students:** Courses specific to trait-based research are often lacking at both the undergraduate and graduate levels. Where courses or modules are taught, the focus may be limited to a subset of the trait data life cycle (e.g. Collection and Analysis; Fig. 1), leaving students lacking critical skills (Feng et al., 2020). The use of Open Educational Resources, including those built using incubators (Ryder et al., 2020), is one promising method for making the implementation of such courses and modules easier. In particular, authentic teaching experiences provide a number of benefits over traditional lectures or “cook-book” experiments (Brownell et al., 2012), and seem well suited to trait-based ecology given that many traits can be collected easily and inexpensively, and that there are a host of tools available (see e.g. de Bello et al., 2021). One example of such authentic teaching experiences, the TraitTrain plant functional trait courses (https://plantfunctionaltraitscourses.w.uib.no/), has provided training across the entire trait data life cycle to hundreds of participants and has
created scientific (Henn et al., 2018), data (Vandvik et al., 2020), methodological (Maitner et al., 2021), and pedagogical (Geange et al., 2021) publications.

Train colleagues: Making colleagues aware of important developments in trait-based research via either formal (e.g. publishing protocols, giving talks) or informal means (e.g. conversations, social media, email) is an important way of helping to advance the field. Further, trait-based research is an integrative field and provides many opportunities for collaboration and idea sharing across branches of life science and so discussing traits with a wide variety of colleagues is useful.

Train the world: There is an urgent need for more comprehensive trait data across the globe and the tree of life (Feng et al., 2022) and thus increasing access to training globally is critically important. Open access publications, tools, data, and educational resources provide ways to help lower the barriers to participation (Evans & Reimer, 2009). Further, due to the relative ease, low cost, and tangible nature of many functional traits, they are well-suited to inclusion in elementary education and citizen science (e.g. Isaac & Martin, 2019; Schiller et al., 2021).

Conclusions

This ten rules document is particularly motivated by its last rule and has the intention of bringing awareness of the different facets of a trait's life cycle to the community. Most scientists working with traits usually are experts on only one or few aspects of the life cycle of trait data, which are numerous, ranging from biological theory and field research to computer sciences as well as publishing and review ethics. When it comes to working with trait data, we gain particularly as an interdisciplinary community of field biologists, synthesis
ecologists, computer scientists and database managers. We hope these basic guidelines can be useful as a starter for active communication in disseminating such integrative knowledge and how to make trait data future-proof.

Conflict of Interest statement

All authors declare to have no conflicts of interest.

Acknowledgements

The manuscript was drafted during a workshop by participants of the sDiv working group sDevTrait hosted by the German Centre for Integrative Biodiversity Research (iDiv) (DFG FZT 118). We appreciate funding (grant number W7.15 to AK and MA) and support before, during and after the workshop by the iDiv, particularly staff members Marten Winter, Carolin Kögler and Doreen Brückner. SD acknowledges additional support from LifeWatch Belgium. Funding to keep the World Register of Marine Species (WoRMS) is currently provided through the LifeWatch Belgium project. WoRMS constitutes a major contribution to the LifeWatch Species Information Backbone. CP acknowledges support from the DFG Priority Program 1374 "Biodiversity- Exploratories" (DFG-193921238).

Author contributions

AK and MA conceived the idea. CP, JP, MA and AK led the writing of the manuscript. AK, MA, BE, HB, SD, DF, REO, RG, JH, JK, JM, BM, WP, CP, FS and JP contributed in writing the first draft during a workshop session and thereafter. All authors contributed critically to the drafts and gave final approval for publication.

Data availability

Not applicable.
References


https://doi.org/10.1016/j.ecoinf.2020.101132


https://www.biodiversitylibrary.org/item/31642


https://doi.org/10.1111/geb.13184


Gama, J. (2014). *NISTunits: Fundamental physical constants and unit conversions from NIST.*


Different sets of traits explain abundance and distribution patterns of European plants at different spatial scales. *Journal of Vegetation Science, 32*(2), e13016.
https://doi.org/10.1111/jvs.13016

https://doi.org/10.1111/1365-2745.12025


