# Title page

**Article title**: `NATDB`: An `R` package that downloads species trait data, but is Not A Trait DataBase

**Running head**: `NATDB`: Not A Trait DataBase

**Authors:** William D. Pearse[1,2,*], Maxwell J. Farrell[3], Konrad Hafen[4], Mallory Hagadorn[1], Spencer B. Hudson[1], Sylvia Kinosian[1], Ryan McCleary[1], Alexandre Rego[1], & Katie Welgarz[1]

[1]Department of Biology, Utah State University, 5305 Old Main Hill, Logan UT, 84322, USA

[2]Ecology Center, Utah State University, 5305 Old Main Hill, Logan UT, 84322, USA

[3] Biology Department, McGill University, 1205 Docteur Penfield, Montreal, QC H3A 1B1, Canada

[4]Department of Watershed Science, Utah State University, 5305 Old Main Hill, Logan UT, 84322, USA

[*]To whom correspondence should be addressed: `will.pearse@usu.edu`

**Keywords**: traits, database, `R`, open science, taxonomy

**Word-count**: XXX

# 1 Abstract


1. Ecologists and evolutionary biologists often wish to make use of species trait data, either as ancillary data, such as in community ecology, or as the primary focus of a study, such as macro-evolutionary modelling.

2. Such biologists are often hampered by the difficulties of collecting sufficient trait data from published sources.

3. We present `NATDB`, an `R` package that automatically downloads species trait data from existing sources.

4. NATDB collates trait data from over 100 publications across $\sim 120,000$ species, and at the time of writing downloads over 3.5 million individual trait measurements.

5. NATDB is *Not A Trait DataBase*: it circumvents issues over the intellectual ownership of data by distributing no data, and merely giving users automated tools to create their own database from data that scientists have already agreed to share. We hope to establish a community around this package that will add additional data sources and cleaning routines.

6. `NATDB` can be installed by typing `library(devtools);install_github("willpearse/natdb")` at an `R` console. We will submit the package to CRAN once it is accepted at a journal.

# 2 Introduction

Ecologists and evolutionary biologists have long recognised the importance of (functional) traits in their work (Díaz & Cabido 2001). Biodiversity is a multi-faceted concept (Purvis & Hector 2000), but it is intuitively obvious that if we are to understand patterns in biodiversity and the processes that govern it, we must have quantitative data on the properties of species themselves. Large datasets of plants (Kattge *et al.* 2011), mammals (Jones *et al.* 2009), and birds (Wilman *et al.* 2014) have opened the door to analyses of the evolution (Harmon *et al.* 2010; Pennell *et al.* 2015) and global distribution (Kattge *et al.* 2011; Gross *et al.* 2017) of trait diversity. Species' traits help us better predict how species will respond to land use (Mayfield *et al.* 2010) and climate change (Estrada *et al.* 2016), allowing us to generalise and compare across species to find general biodiversity patterns.

Yet, despite its importance, it is often difficult to find data on species' functional traits. We suggest there are three main reasons for this: (1) it is difficult to obtain trait data, (2) it is difficult to collate trait data, and (3) trait data are not always published. (1) Often the most functionally important species traits are the most difficult to measure (Cornelissen *et al.* 2003; Violle *et al.* 2007), and even when measuring a trait is simple, finding a suitable specimen is often not. Usefully measuring and defining species' traits requires specialist knowledge and expertise, and is difficult to do properly. If a solution exists to this problem, it is likely a re-allocation of funding towards training biologists capable of measuring species' traits. While advances in machine learning improve the prospects of automated trait collection (*e.g.*, Pearse *et al.* 2016), methods for easing data collection will most likely vary across species and traits: there will be no single silver-bullet solution.

The last two of the problems above, however, may be possible to tackle across all species and traits simultaneously. (2) Creating and maintaining large databases is complex: the nomenclatures for species and traits are not universal (Kattge *et al.* 2011; Hudson *et al.* 2017), and resolving discrepancies across datasets requires detailed knowledge of species and their traits. A potential solution, however, is to develop a framework within which it is possible for this specialist-driven, intensive discrepancy resolution to be parallelised across many scientists. Focusing on the development not

of a database *per se*, but rather on the tools that enable the creation of databases, may allow this. The modern data scientist makes frequent use of tools such as `GitHub` (reviewed in Ram 2013), which are designed to ease decentralised effort on common tasks such as nomenclature resolution. (3) Unlike other other kinds of data such as DNA sequences (Benson *et al.* 2013), the publication of species trait data has been controversial (*e.g.*, Moles *et al.* 2013; Poisot *et al.* 2013). The most compelling argument against the publication of data is a concern that a database creator, with a few minutes' work, could grab data that took decades to collect, and receive all the citation credit for that data as released in what we call a "*pseudo-new*" dataset. This objection would be resolved if scientists didn't use (or create) such pseudo-new datasets, and instead cited the sources of the original data that they make use of.

We present here `NATDB`, an `R` package that collates over 3.5 million trait records for $\sim$120,000 species, making existing species trait data more widely available for use by ecologists and evolutionary biologists. We argue that `NATDB` is a prototype for a new way of making data more accessible that avoids concerns about data ownership and pseudo-new datasets: `NATDB` is *Not A Trait DataBase*. `NATDB` is a software package that simplifies the process of collating data the user already has access to, and so obviates any concerns over attribution in pseudo-new dataset because it simply retrieves data whose collectors have already publicly released. `NATDB` contains no data, and so users must cite the sources of data when using the package. This model both liberates the vast trait-based knowledge that already exists in the literature, and protects the intellectual contributions of those who collected the data in the first place.

# 3 Usage

NATDB consists of a series of internal functions, each of which downloads a single dataset from a published source. Typically, a user will download a set of data and then subset that down to only the species or traits that they require. Note that, by default, NATDB waits ten seconds between downloading datasets to minimise impact on journals' servers. For example, the following would download all the data in NATDB, and then subset that down to only two kinds of traits for two species:

```
library(natdb)
data <- natdb(taxon)
species <- c("Quercus_robur", "Pinus_sylvestris")
traits <- c("specific_leaf_area", "height")
subset.data <- data[species, traits]
```

NATDB can cache whatever it downloads during an R session. So, for example, if the user were to realise that they wanted data on an additional species or trait after executing the code above, they could run the entire script again and NATDB would not download any more data. To use this option, a user must specify a directory when invoking NATDB so that they can save their searches between sessions. The following code, for example, would cache results between sessions, and would add additional data to that cache as new versions of NATDB were released. This is the recommended way to use NATDB, as it saves the user time and reduces server load.

```
data <- natdb(cache="~/.natdb_cache/")
subset.data <- data[c("Phocoena_phocoena","Tursiops_truncatus"),]
```

NATDB has a single class for dealing with trait data, called (unimaginatively) natdb. This class has head, print, summary, and as.data.frame methods to make it easier for the user to work with their data. Internally, NATDB distinguishes between, and convert all data into, numeric and character types, and *melts* (*sensu* Wickham 2007) all data within these types. This makes it easy to add new data to an existing NATDB object, and keeps the memory requirements manageable. If NATDB were to store data in a data.frame-like format, it would require $\sim 14,400,000,000$ cells ($\sim 120,000$

species, $\sim 1,200$ traits) to store all its data, $\sim 2\%$ of which would be empty. Users are encouraged to explore options for summarising the data in `NATDB`, as the means and modes reported for the `numeric` and `character` data, respectively, may not be best-suited for their particular needs.

Ready access to meta-data is important in any database. The databases `NATDB` can build are complex, in that the meta-data that different source datasets provide can vary a great deal. We follow the general approach of *FigShare* (`https://figshare.com/`) and *DataDryad* (`http://datadryad.org/`) in not enforcing rigid meta-data requirements, but placing the meta-data of each source dataset within a comparable framework so as to allow users to interrogate the meta-data in their own way. Thus while we do standardise some aspects of the data (*e.g.*, ensuring all latitude and longitude measurements, where present, are termed `latitude` and `longitude`), users must check whatever subset of data they have to see what meta-data are available. For example:

```
simplified.data <- as.data.frame(subset.data)
simplified.metadata <- metadata(subset.data)
plot(simplified.data$specific_leaf_area ~ metadata$latitude)
```

We make no guarantee that the taxonomy or units of the data within `NATDB` are internally compatible: users are responsible for checking the validity of the data they have collated. However, we have made attempts to harmonise trait names within `NATDB`, and provided wrapper scripts that use `taxize` and `convertr` to harmonise taxonomy and units, respectively. For example:

```
data <- natdb(taxon)[c("Quercus_robur", "Pinus_sylvestris"), ]
# Basic cleaning (of trait, species, and unit names)
clean.data <- clean.natdb(data)
# Clean taxonomy using Global Names Resolver through taxize
clean.data <- clean.natdb.names(clean.data)
# Convert traits to a single unit of measurement
clean.data <- clean.natdb.units(clean.data)
```

Finally, it is important that those who generated the data `NATDB` downloads are appropriately cited. It is simple to generate BibTeX files to help with citations; for example:

```
citations(clean.data)
```

# 4 Coverage and scope

As table 1 shows, NATDB downloads data from over 100 published papers, and covers a reasonably wide range of taxonomic groups (including, but not limited to, plants, birds, mammals, amphibians, and reptiles). In terms of species- and trait-level coverage, over 200 traits have data on over 1000 species (see figure 1). We emphasise that more rigorous taxonomic cleaning (and trait definition checking) may somewhat alter these figures (see also below). Critically, NATDB has been designed, from the ground-up, to be easy to extend. Adding a publication's data to the package requires no knowledge other than the basic structure of data to be added. Many of the functions that load data structure into NATDB are fewer than ten lines of R code, in part because we have contributed code to the R package fulltext (Chamberlain 2015) to automate the download of data from published papers. Since NATDB uses reflective programming to determine what datasets are available for download, all a user need do to add a dataset to NATDB is submit a 'pull request' on GitHub with a function. This represents a major advantage to NATDB: it is a living package that will, we hope, grow as authors add their own publications to it. We provide detailed instructions on how to contribute data sources to NATDB in the package's vignette and on the package's GitHub wiki.

The flexibility and scope of NATDB, however, means its output has not been as carefully cleaned and checked as most published datasets typically are. This is by design: NATDB is fundamentally different, and we use the TRY dataset (Kattge *et al.* 2011) to illustrate this. TRY is a carefully-collated dataset that required thousands of person-hours to create, and to reflect this and ensure that the data is used correctly, its authors require that many data contributors and the two lead authors of the database are offered co-authorship on any publication making use of TRY data. We consider this reasonable request given the amount of effort involved in producing a database like TRY, and the feedback and data-validation that these additional co-authors provide to a finished manuscript. But NATDB is not a database and does not follow this model: the data are provided 'as-is' and neither we, nor the original data publishers, require co-authorship for use of the package.

# 5 Future directions

We actively encourage code contributions, and the package's online vignette contains a detailed set of instructions on how to contribute functions that download data from new sources. Our intention is to make the process as simple as possible, and so encourage authors to release the data underlying their analyses and integrate them into the package. We hope that this will speed scientific development by making data more widely available for analysis, and increase the rate at which those who collect data can be acknowledged and cited for their effort. This, in part, is the reason we have few formal checks on meta-data and units within `NATDB`: the first hurdle we must overcome is getting data 'out there' in a useable format, and everything else is a problem for the future. We hope that, using `NATDB` as a base, others will develop cleaning and checking routines that can be applied to the data the package downloads. Whether these will be incorporated into `NATDB` itself, or released as separate companion package(s), remains to be seen.

Without data there can ultimately be no science, and we hope `NATDB` will make it easier to access data and to acknowledge those who collected it. `NATDB` is an experiment, and its long-term success or failure will depend on whether we can develop around it a community of scientists willing to share their data through it.

# Bibliography


Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, & E. W. Sayers (2013). GenBank. *Nucleic Acids Research* 41.D1, D36–D42.

Cayuela, L., Í. Granzow-de la Cerda, F. S. Albuquerque, & D. J. Golicher (2012). Taxonstand: An R package for species names standardisation in vegetation databases. *Methods in Ecology & Evolution* 3.6, 1078–1083.

Chamberlain, S. (2015). fulltext: Full text of 'scholarly' articles across many data sources. R package version 0.1.4.9000.

Cornelissen, J. H. C., S. Lavorel, E. Garnier, S. Díaz, N. Buchmann, D. E. Gurvich, *et al.* (2003). A handbook of protocols for standardised and easy measurement of plant funcitonal traits worldwide. *Australian Journal of Botany* 51, 335–380.

Díaz, S. & M. Cabido (2001). Vive la différence: plant functional diversity matters to ecosystem processes. *Trends in Ecology & Evolution* 16.11, 646–655.

Estrada, A., I. Morales-Castilla, P. Caplat, & R. Early (2016). Usefulness of species traits in predicting range shifts. *Trends in ecology & evolution* 31.3, 190–203.

Gionata, B. (2015). TR8: an R package for easily retrieving plant species traits. *Methods in Ecology and Evolution* 6.3, 347–350.

Gross, N., Y. Le Bagousse-Pinguet, P. Liancourt, M. Berdugo, N. J. Gotelli, & F. T. Maestre (2017). Functional trait diversity maximizes ecosystem multifunctionality. *Nature Ecology & Evolution* 1, 0132.

Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, *et al.* (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution* 64.8, 2385–2396.

Hintze, C., F. Heydel, C. Hoppe, S. Cunze, A. König, & O. Tackenberg (2013). D3: the dispersal and diaspore database–baseline data and statistics on seed dispersal. *Perspectives in Plant Ecology, Evolution and Systematics* 15.3, 180–192.

Hudson, L. N., T. Newbold, S. Contu, S. L. Hill, I. Lysenko, A. De Palma, H. R. Phillips, T. I. Alhusseini, F. E. Bedford, D. J. Bennett, *et al.* (2017). The database of the PREDICTS (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems) project. *Ecology and Evolution* 7.1, 145–188.

Jones, E. I., R. Ferrière, & J. L. Bronstein (2009). Eco-evolutionary dynamics of mutualists and exploiters. *The American naturalist* 174.6, 780–94.

Kattge, J., S. Díaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bönisch, *et al.* (2011). TRY - a global database of plant traits. *Global Change Biology* 17.9, 2905–2935.

Mayfield, M. M., S. P. Bonser, J. W. Morgan, I. Aubin, S. McNamara, & P. A. Vesk (2010). What does species richness tell us about functional trait diversity? Predictions and evidence for responses of species and functional trait diversity to land-use change. *Global Ecology and Biogeography* 19.4, 423–431.

Moles, A., J. B. Dickie, & H. Flores-Moreno (2013). A response to Poisot et al.: publishing your dataset is not always virtuous. *Ideas in Ecology and Evolution* 6.2.

Pearse, W. D., J. Cavender-Bares, S. E. Hobbie, M. Avolio, N. Bettez, R. R. Chowdhury, P. M. Groffman, M. Grove, S. J. Hall, J. B. Heffernan, *et al.* (2016). Ecological homogenisation in North American urban yards: vegetation diversity, composition, and structure. *bioRxiv*, 061937.

Pennell, M. W., R. G. FitzJohn, W. K. Cornwell, & L. J. Harmon (2015). Model adequacy and the macroevolution of angiosperm functional traits. *The American Naturalist* 186.2, E33–E50.

Poisot, T., R. Mounce, & D. Gravel (2013). Moving toward a sustainable ecological science: don't let data go to waste! *Ideas in Ecology and Evolution* 6.2, 11–19.

Purvis, A. & A. Hector (2000). Getting the measure of biodiversity. *Nature* 405.6783, 212–219.

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine* 8.1, 7.

Violle, C., M.-L. Navas, D. Vile, E. Kazakou, C. Fortunel, I. Hummel, & E. Garnier (2007). Let the concept of trait be functional! *Oikos* 116.5, 882–892.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software* 21.12, 1–20.

Wilman, H., J. Belmaker, J. Simpson, C. de la Rosa, M. M. Rivadeneira, & W. Jetz (2014). EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology* 95.7, 2027–2027.

Wright, I. J., P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, F. Bongers, J. Cavender-Bares, T. Chapin, J. H. Cornelissen, & M. Diemer (2004). The worldwide leaf economics spectrum. *Nature* 428.6985, 821–827.

# Figures

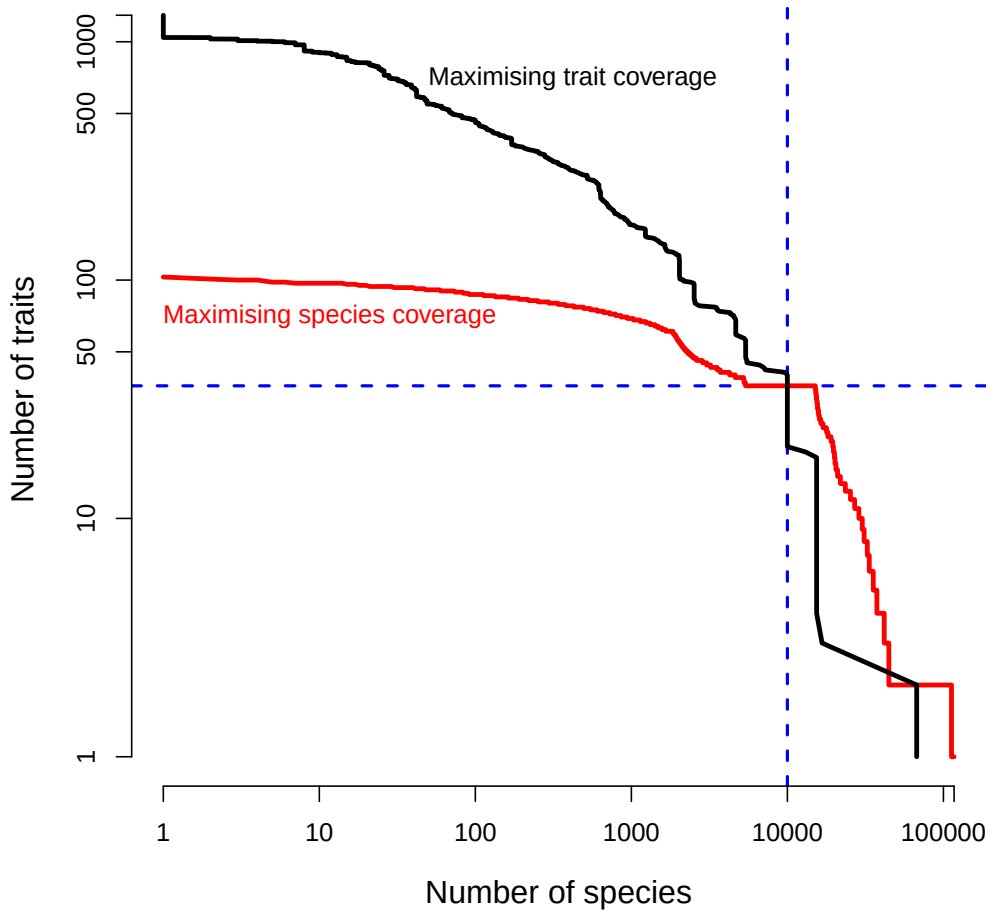


Figure 1: Data coverage within `NATDB`. Within `NATDB`, no species has data for every trait, and no trait has data for every species. Thus while `NATDB` downloads over 3.5 million pieces of data, this number is not necessarily representative of the species- or trait-coverage can expect to work with. Since it is possible to select groups of species or traits within `NATDB` maximising the number of species or traits for which data are available, we plot coverage curves maximising species (in red) and traits (in black). The point of intersection between the two curves (at $10,000$ species and 36 traits) is shown with blue dashed lines. This plot was produced using data that had been run through `NATDB`'s `clean.natdb` function to harmonise trait and (to a limited extent) species names.

# Tables

| Taxonomic group | # Species | # Traits | % Complete | Citations |
|---|---|---|---|---|
| Plants | | | Wright *et al.* (2004) | |
| Mammals | | | Jones *et al.* (2009) and Wilman *et al.* (2014) | |
| Birds | | | Wilman *et al.* (2014) | |
| ...TBC... | | | | |

Table 1: Overview of data available for download within `NATDB`. Overall, the package downloads XXX data points, covering XXX species and XXX separate functional traits. XXX% of these trait values have some form of meta-data associated with them.