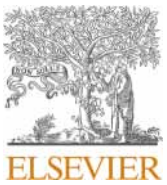
Ecological Informatics 62 (2021) 101206



Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Integration and harmonization of trait data from plant individuals across heterogeneous sources

Tim P. Lenters [a], Andrew Henderson [b], Caroline M. Dracxler [a], Guilherme A. Elias [c], Suzanne Mogue Kamga [d], Thomas L.P. Couvreur [e], W. Daniel Kissling [a,*]

[a] Department of Theoretical and Computational Ecology (TCE), Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, P.O. Box 94240, 1090 GE Amsterdam, the Netherlands
[b] The New York Botanical Garden, Bronx, NY 10458-5126, USA
[c] Programa de Pós-Graduação em Ciências Ambientais, Universidade do Extremo Sul Catarinense UNESC, Av. Universitária, 1105, 88806-000 Criciúma, SC, Brazil
[d] Université de Yaoundé I, Ecole Normale Supérieure, Département des Sciences Biologiques, Laboratoire de Botanique systématique et d'Ecologie, B.P. 047 Yaoundé, Cameroon
[e] IRD, DIADE, Univ Montpellier, Montpellier, France

ARTICLE INFO

Keywords:
Data mobilization
Data science
Functional traits
Standards
Ontology
Semantics

ABSTRACT

Trait data represent the basis for ecological and evolutionary research and have relevance for biodiversity conservation, ecosystem management and earth system modelling. The collection and mobilization of trait data has strongly increased over the last decade, but many trait databases still provide only species-level, aggregated trait values (e.g. ranges, means) and lack the direct observations on which those data are based. Thus, the vast majority of trait data measured directly from individuals remains hidden and highly heterogeneous, impeding their discoverability, semantic interoperability, digital accessibility and (re-)use. Here, we integrate quantitative measurements of verbatim trait information from plant individuals (e.g. lengths, widths, counts and angles of stems, leaves, fruits and inflorescence parts) from multiple sources such as field observations and herbarium collections. We develop a workflow to harmonize heterogeneous trait measurements (e.g. trait names and their values and units) as well as additional information related to taxonomy, measurement or fact and occurrence. This data integration and harmonization builds on vocabularies and terminology from existing metadata standards and ontologies such as the Ecological Trait-data Standard (ETS), the Darwin Core (DwC), the Thesaurus Of Plant characteristics (TOP) and the Plant Trait Ontology (TO). A metadata form filled out by data providers enables the automated integration of trait information from heterogeneous datasets. We illustrate our tools with data from palms (family Arecaceae), a globally distributed (pantropical), diverse plant family that is considered a good model system for understanding the ecology and evolution of tropical rainforests. We mobilize nearly 140,000 individual palm trait measurements in an interoperable format, identify semantic gaps in existing plant trait terminology and provide suggestions for the future development of a thesaurus of plant characteristics. Our work thereby promotes the semantic integration of plant trait data in a machine-readable way and shows how large amounts of small trait data sets and their metadata can be integrated into standardized data products.

1. Introduction

The integration and harmonization of data from heterogeneous sources is one of the biggest challenges in current ecological research (Farley et al., 2018). Like many other branches in biology, ecology has seen a strong increase in data availability over the past few decades (Farley et al., 2018). This increase corresponds to a general trend in the accumulation of data volumes, exponentially increasing in the past decade (Chen et al., 2014). Such 'big data' provide many opportunities for studying ecological systems at much larger spatial, temporal and taxonomic scales than has been previously possible (Dietze, 2017; LaDeau et al., 2017). Besides massive data volumes, there is also a large number of small datasets gathered within the ecological sciences which are often described as 'long-tail data' (Heidorn, 2008). These data are usually collected by individual researchers, over relatively small spatial and temporal scales and with funding models that often provide little resources for data curation and sharing (Heidorn, 2008; LaDeau et al., 2017).

* Corresponding author.
E-mail address: wdkissling@gmail.com (W.D. Kissling).

https://doi.org/10.1016/j.ecoinf.2020.101206
Received 14 August 2020; Received in revised form 24 October 2020; Accepted 9 November 2020
Available online 27 February 2021

Such long-tail data are thus often not collected, indexed or shared in a standardized way, limiting their findability and usability for other researchers. Moreover, when data collectors move on to different research projects or ultimately retire, a loss of information content and data degradation is inevitable (Michener et al., 1997). Developing methods to improve the integration of such long-tail data therefore enhances their re-use in other ecological research projects or meta-analyses (Gerstner et al., 2017).

One aspect within ecology that benefits from such data integration is trait-based research, which has sharply increased in recent years (Gallagher et al., 2020; Kattge et al., 2020). For plants, traits are critical to plant form and function (incl. growth, survival and reproduction) and therefore shape fundamental aspects of population and ecosystem dynamics as well as ecosystem services (de Bello et al., 2010; Díaz et al., 2016). Studies have shown that traits such as leaf size and seed mass determine how species respond to environmental factors (Campetella et al., 2011; Wright et al., 2017). Other traits (e.g. below ground biomass, spines, leaf nutrients, fruit sizes and colours) serve as a link between biodiversity and ecosystem functioning (Lavorel and Garnier, 2002; Wilke and Snapp, 2008) or mediate interactions with animal mutualists and antagonists (Nascimento et al., 2020; Onstein et al., 2017; Tielens and Gruner, 2020). Traits have thus become an integral part of predictive ecology and global change biology (Díaz et al., 2016; McGill et al., 2006; Schleuning et al., 2020; Westoby and Wright, 2006). Traits further show inter- and intraspecific variation, and both are relevant for assessing global change impacts on biodiversity (Bjorkman et al., 2018; Díaz et al., 2013; Kissling et al., 2018). However, intraspecific trait variation across space and time is often not widely analyzed in ecological research, despite its importance for biodiversity responses to global change, for instance in regions with low diversity or when species are widely distributed (Siefert et al., 2015). Measuring intraspecific trait variation over time is also essential for monitoring biodiversity change (Kissling et al., 2018) as it can inform about policy targets such as the 20 'Aichi Biodiversity Targets' and their post-2020 successors, or the 17 'Sustainable Development Goals' (Geijzendorffer et al., 2016). However, the limited availability of individual trait measurements makes it difficult to use species traits for assessing progress towards policy goals (Kissling et al., 2018).

Many projects that combine trait measurements from different sources aggregate data at the species level (Gallagher et al., 2020; Parr et al., 2016). This is easier because publications (e.g. books, monographs and taxonomic revisions) often only report trait means and extremes (minimum and maximum values) rather than the raw data measured directly from individual organisms. However, this also leads to a loss of information content regarding individual-level trait variation, as intraspecific variability and spatial (or temporal) variation are not captured (Guralnick et al., 2016; Kissling et al., 2018). One challenge is that many individual-level trait measurements are only available from the researchers themselves, e.g. from taxonomists writing monographs and taxonomic revisions or from ecologists who have done the field work but only store the data on their personal computers. Another source are herbarium collections which often contain individual-level trait information, sometimes written on the specimen labels (Guralnick et al., 2016; Miller-Rushing et al., 2006; Robbirt et al., 2011). These records are therefore often not findable, accessible, interoperable and reusable, and thus do not adhere to the 'FAIR Data Principles' (Wilkinson et al., 2016). If the FAIR Data Principles are used as a guideline for data management in ecological research, the individual-level trait measurements from heterogeneous sources could substantially accelerate trait-based ecological science and policy applications (Gallagher et al., 2020; Guralnick et al., 2016; Kissling et al., 2018).

Several methods and guidelines already exist to aid ecological data integration and interoperability using the FAIR Data Principles (Hardisty et al., 2019; Kissling et al., 2018; Michener et al., 1997). For instance, in ecology different thesauri (i.e. synonym dictionaries) have been developed to provide standardized names and definitions for biodiversity-related metadata terms. Examples are the 'Darwin Core' (DwC; Wieczorek et al., 2012), the 'Ecological Trait-data Standard' (ETS; Schneider et al., 2019) and the 'Ecological Metadata Language' (EML; Fegraus et al., 2005). They consist of controlled vocabularies (lists of consensus terms for a given concept) which are linked to their associated definitions and other information. Beyond thesauri, integrating data from disparate sources also requires ontologies. These are semantic models that allow formal descriptions of the relationships among concepts and vocabulary terms. They consist of controlled vocabularies with associated definitions, but also define the relationships between different terms (Gruber, 1995). For plant traits, three prominent examples are the 'Plant Ontology' (PO; Jaiswal et al., 2005), the 'Plant Trait Ontology' (TO; Arnaud et al., 2012) and the 'Thesaurus Of Plant characteristics' (TOP; Garnier et al., 2017). Ontologies, as well as thesauri, make use of semantic standards, allowing machine-readability and interoperability. The terminologies are also linked to Uniform Resource Identifiers (URIs), which are unambiguous codes that refer to a single term in a thesaurus or ontology. Incorporating these ontologies and thesauri in ecological research makes the data better findable (e.g. via extensive metadata that are assigned to URIs), interoperable (via controlled vocabularies and standardized terminologies) and reusable (via detailed descriptions of metadata following semantic standards) (Garnier et al., 2017; Parr et al., 2016; Schneider et al., 2019). Although these tools are now increasingly implemented in ongoing research projects, most available data from past research do not follow these principles. Integrating trait data from heterogeneous sources therefore benefits from using semantic standards and standardized terminologies.

Here, we develop a workflow for the integration and harmonization of quantitative plant trait measurements (e.g. lengths, widths, counts and angles of stems, leaves, fruits and inflorescence parts) from heterogeneous sources. We illustrate this with palms (Arecaceae), a model system for tropical biodiversity science and ecological and evolutionary research due to their global distribution, their ecologically representativeness, ample taxonomic and systematic research and their importance for animals as food resources (Couvreur and Baker, 2013; Eiserhardt et al., 2011; Henderson, 2002; Muñoz et al., 2019). As with other trait-based research, most of the available trait data for palms is limited to species-level measurements such as averages, minimum and maximum values (Kissling et al., 2019). Individual-level trait measurements do exist (e.g. underlying taxonomic revisions of palm genera), but are typically not digitally available in a standardized data format nor do they follow semantic and terminology standards or have links to existing ontologies. We use unstandardized spreadsheets from palm taxonomists and ecologists containing individual palm trait measurements and develop a metadata form and an R script to automatically integrate and standardize these data. We follow the suggestions from the ETS (Schneider et al., 2019) and provide as output a core table containing species names, trait names, trait values and trait units, and several extension tables containing information about taxonomy, measurement or fact and occurrence. All tables are linked through identifiers and follow semantic standards from the ETS, DwC, TO and TOP. Our workflow enables the semantic integration of plant trait data in a machine-readable way and aims to increase the discoverability, semantic interoperability, digital accessibility and re-use of long-tail trait data such as those collected by plant ecologists and taxonomists.

2. Materials and methods

2.1. Workflow

We developed a workflow for integrating quantitative plant trait data consisting of three main parts: input, data integration and output (Fig. 1).

The first part ('input') takes information from two types of external sources. The first are (Excel) spreadsheets from domain researchers which contain individual-level measurements of plant traits and its metadata. These spreadsheets have to be provided by the data provider. The second source are open-access ontologies such as the 'Thesaurus Of Plant characteristics' (TOP; Garnier et al., 2017) and the 'Plant Trait Ontology' (TO; Arnaud et al., 2012). These provide standardized trait names, metadata information and an unambiguous identifier (URI) for each term described in

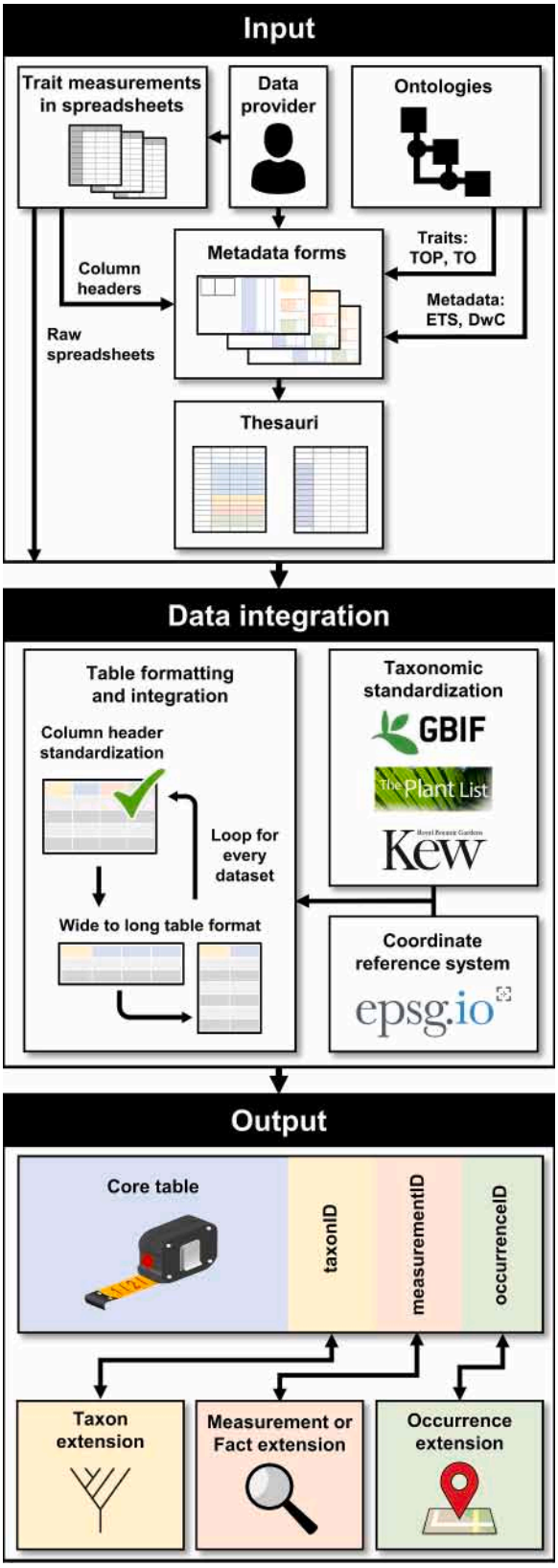


Fig. 1. Workflow for integrating individual plant trait measurements from heterogeneous sources. The workflow comprises three main steps (input, data integration and output) and requires spreadsheets with trait measurements, metadata forms and ontologies for data integration. Abbreviations: TOP = Thesaurus of Plant characteristics; TO = Plant Trait Ontology; ETS = Ecological Trait-data Standard; DwC = Darwin Core.

the final output dataset. Additional metadata information can be linked to the 'Darwin Core' (DwC; Wieczorek et al., 2012) and the 'Ecological Trait-data Standard' (ETS; Schneider et al., 2019). These ontologies are already semantically linked to a metadata form and require no additional input from the data provider. Column headers from the spreadsheets are linked to quantitative trait terms derived from the ontologies through a metadata form (Fig. 1). We focus on quantitative traits (i.e. continuous measurements) because they provide numeric trait values to capture intraspecific variation, in contrast to qualitative traits (categorical or binary data) which usually vary less within species than quantitative traits, and often come with many verbatim trait descriptions (e.g. colours) that are difficult to standardize. The metadata form is filled-out by the dataset provider and enables to specify two thesauri (a metadata thesaurus and a units thesaurus). These thesauri facilitate the automated standardization and integration of each input spreadsheet (Fig. 1).

The second part of the workflow ('data integration') enables the integration of the different trait spreadsheets. This requires to standardize column headers and to change the table format from a wide to a long table. It further standardizes measurement units (into centimeters) and the taxonomy, e.g. by using the backbone taxonomy from the 'Global Biodiversity Information Facility' (GBIF.org, 2020), from the 'The Plant List' (Kalwij, 2012) or from the 'World Checklist of Vascular Plants' (WCVP, 2020). These three databases identify the use of synonym names and replace them with accepted names. Additionally, the coordinate reference systems are harmonized using the Geodetic Parameter Dataset of the 'European Petroleum Survey Group' (EPSG; Bivand et al., 2019). This is one of the most widely used spatial reference system in geographic information systems. The simple numeric codes of the EPSG make it machine-readable and less prone to errors than other reference systems. This system is also often updated to include the most recent coordinate references. All data integration in the workflow is built on scripts and functions using the open source programming language R (R Core Team, 2013; v. 4.0.2). File reading and saving and data frame manipulations are done using memory efficient functions from the 'data.table' and 'dplyr' packages (Dowle and Srinivasan, 2019; Wickham et al., 2019). This allows for fast computation even in the case of many input spreadsheets.

The third part of the workflow ('output') divides the obtained long-table dataset into four different tables (following Schneider et al., 2019): (1) a core table providing original and standardized quantitative trait terms, (2) a taxon extension table with standardized taxonomic information, (3) a measurement or fact extension table with the reference and basis of record, and (4) an occurrence extension table capturing spatial information on observation and sampling. The four tables are linked via IDs and provided as comma-separated values (CSV) files, as this is compatible with most spreadsheet software and programming languages.

2.2. Data sources

To illustrate our data integration workflow, we use individual trait measurements of palms as collected by taxonomists and ecologists (Table 1). These trait measurements are typically stored in Excel spreadsheets and come from different sources, such as herbarium specimens and ecological field measurements. The datasets cover a wide range of taxa and traits. Data entries, measurements, units, column headers and other metadata are typically not standardized across data providers. Besides using terms from the TOP and the TO, we also summarize the trait measurements in categories ("Whole plant", "Shoot," "Leaves" and "Reproductive organs") that have been used in the context of the TRY plant trait database (Kattge et al., 2020). Besides taxonomic information and trait measurements, the spreadsheets often also contain geographic information (e.g. geographic coordinates of sampling events, country information) and additional information about the individual measurements (e.g. collector name, collection institute, sampling date) which we also integrate and harmonize using DwC and the ETS.

Table 1
Summary information of palm trait datasets used to test and develop the data integration workflow. The categorization into types of traits follows the TRY plant trait database (Kattge et al., 2020).

Taxonomic or thematic focus	Source of measurement	Data provider	Types of traits	Species	Reference
Multivariate analysis of *Hyospathe*	Herbarium specimens	A. Henderson	Whole plant, shoot, leaves, reproductive organs	6	Henderson (2004)
Multivariate study of *Calyptrogyne*	Herbarium specimens	A. Henderson	Shoot, leaves, reproductive organs	18	Henderson (2005)
Taxonomic revision of *Desmoncus*	Herbarium specimens	A. Henderson	Whole plant, shoot, leaves, reproductive organs	24	Henderson (2011a)
Taxonomic revision of *Geonoma*	Herbarium specimens	A. Henderson	Whole plant, shoot, leaves, reproductive organs	68	Henderson (2011b)
Taxonomic revision of *Leopoldinia*	Herbarium specimens	A. Henderson	Whole plant, shoot, leaves, reproductive organs	3	Henderson (2011c)
Taxonomic revision of *Pholidostachys*	Herbarium specimens	A. Henderson	Shoot, leaves, reproductive organs	7	Henderson (2012)
Taxonomic revision of *Chuniophoenix*	Herbarium specimens	A. Henderson	Shoot, leaves, reproductive organs	3	Henderson (2015)
Taxonomic revision of *Rhapis*	Herbarium specimens	A. Henderson	Shoot, leaves, reproductive organs	11	Henderson (2016)
Morphometric study of *Synechanthus*	Herbarium specimens	A. Henderson	Shoot, leaves, reproductive organs	2	Henderson and Ferreira (2002)
Taxonomic revision of *Welfia*	Herbarium specimens	A. Henderson	Whole plant, shoot, leaves, reproductive organs	2	Henderson and Villalba (2013)
Taxonomic revision of *Attalea*	Herbarium specimens	A. Henderson	Shoot, leaves, reproductive organs	30	Henderson (2020a)
Taxonomic revision of *Calamus*	Herbarium specimens	A. Henderson	Shoot, leaves, reproductive organs	411	Henderson (2020b)
Phytosociological study in the Brazilian Atlantic Forest	Ecological field measurements	G. A. Elias	Whole plant	8	Elias et al. (2019)
Taxonomic revision of *Raphia*	Herbarium specimens	T.L.P. Couvreur & S. Mogue	Whole plant, shoot, leaves, reproductive organs	23	Unpublished
Seed measurements of Arecaceae	Ecological field measurements	C.M. Dracxler	Reproductive organs	1	Unpublished

3. Results

3.1. Input data

The provided input data represented different long-tail palm trait spreadsheets (Table 1) which often followed different formats and sometimes contained mistakes which prevented an automated data integration. The most common mistakes included formatting issues such as placeholder values (e.g. 'n.m' or 'NA'), measurement units together with trait values, ranges rather than single values, and the use of different decimal separators within the same spreadsheet (Fig. 2). We provide examples and suggestions for how to avoid these common mistakes (Fig. 2).

Before the spreadsheets are automatically integrated, the filled out metadata forms are validated in the script. This automated step checks for each of the four common mistakes described in Fig. 2. It also checks if the dataset provider has entered column headers in the metadata form that are not present in the measurement spreadsheet, mitigating the risks of typing errors. If a mistake is found, the user will receive a message in which column and in which spreadsheet the mistake has to be corrected to make the spreadsheet suitable for automated integration.

3.2. Metadata form

The metadata form was developed to facilitate the automated integration of individual trait measurements (Fig. 1) because manually changing column headers in Excel spreadsheets into standardized terms from ontologies is more prone to errors. The validation step described in the previous section identifies possible mistakes while at the same time matching column headers to standardized trait names. The metadata form also enables to add metadata from the ontologies to the used terms. The dataset providers therefore need to fill out a metadata form consisting of three main parts (Fig. 3a–c). In the first part (Fig. 3a), basic information on name of dataset provider, reference, basis of record (e.g. living specimen, preserved herbarium specimen) and coordinate reference system (if known, otherwise "Unknown" is automatically entered) are filled in. This information facilitates the automated integration of the spreadsheets by providing a reference to the data origin and source (provenance) and specifying the usage of a particular coordinate system. The second part (Fig. 3b) enables trait data integration and consists of a list of standardized quantitative trait names, to which the names of column headers from the spreadsheets are added, as well as their measurement unit. For the 15 input datasets (Table 1), which included a total of 50 trait terms (e.g. "Rachilla_thickness", "Plant_height", "Stem_length" or "Fruit_width"), 25

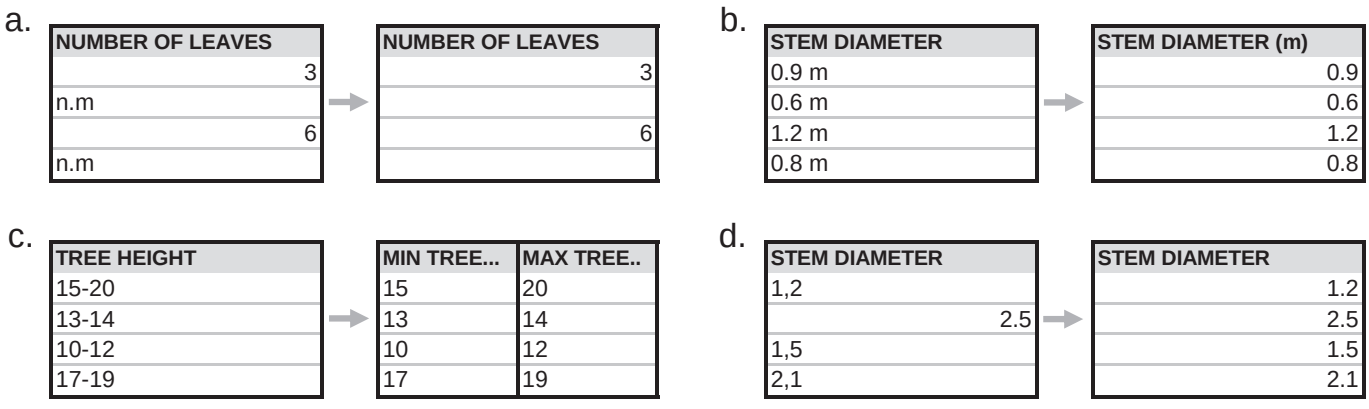


Fig. 2. Suggestions of how to avoid common mistakes in trait measurement spreadsheets which can prevent an automated integration of long-tail data. a. Cells with no measurement value should not be filled with a placeholder, but kept empty instead. b. Measurement units should not be stated in the value cell, but rather in the metadata or column header. c. Variability in measurements should not be given as a range, but rather as separate minimum and maximum values. d. Decimal separators should be homogeneous across all individual measurements in a spreadsheet.

trait names (50%) were present in both the TO and TOP and linked via URIs (Appendix Table A.1). The other 25 trait names could not be linked to ontologies via URIs. The third part of the metadata form (Fig. 3c) consists of three categories ("taxon", "measurement or fact" and "occurrence"), each corresponding to the metadata categories (hence not the traits) proposed by the ETS (Schneider et al., 2019). For those, we asked the dataset providers to match names of column headers (i.e. those not covering traits) to the term with the same definition in the ETS or DwC. Most information (94% of the terms) came from the DwC because the ETS often uses terms from the DwC. The data provider could encounter several issues when filling out the metadata form. For example, a measurement spreadsheet could include traits or metadata terms not included in the metadata form. The solution to this and other issues for both the data provider and workflow user are explained in the accompanying "bookdown" tutorial (see 'Data availability').

3.3. Combining datasets using metadata information

The original spreadsheets together with the filled out metadata forms result in a metadata thesaurus and a units thesaurus (Fig. 4). This results in a core table representing the integrated and harmonized trait data information (Fig. 4). This core table consists of standardized scientific names (scientificName), verbatim and standardized trait names (verbatimTraitName, traitName), standardized trait measurement values and units (traitValue, traitUnit), the URIs linking the traits to the TO or TOP (traitID), and the ID-columns (taxonID, measurementID, occurrenceID) which make the link to the three extension tables ("Taxon", "MeasurementOrFact" and "Occurrence") (Fig. 4).

3.4. Integrated dataset

The integration of the 15 long-tail spreadsheets of individual palm traits (Table 1) resulted in a final core table with 138,993 individual trait measurements covering 50 standardized traits and a total of 551 unique standardized palm species names (based on the GBIF taxonomy, and if no match was found with TPL and the WCVP). Although only 17 of the 181 currently recognized genera of palms are captured, the 551 included species represent 22% of the approximately 2500 currently recognized palm species. Species name standardization led to 549 (89%) name matches with the GBIF taxonomy, 483 (78%) with TPL and 539 (87%) with the WCVP for all verbatim species names in this dataset. The 50 standardized trait names were derived from 158 different verbatim trait names. Most traits represented reproductive organs (64,230 trait measurements, i.e. 46%) followed by leaf traits (49,460 trait measurements, 36%) (Fig. 5a). Examples of such traits are "Rachilla_thickness" and "Fruit_width" for reproductive organs and "Petiole_length" and "Leaflet_number" for leaves. Traits representing the whole plant (e.g. "Plant_height") and shoots (e.g. "Stem_length") were less represented, covering 6118 (4%) and 19,185 (14%) of the measurements, respectively (Fig. 5a).

The 'Taxon' extension table held 2030 unique combinations of 551 species names and 308 infraspecific names (e.g. morphotypes) as described in the original spreadsheets (Table 1). For each species name, higher taxonomic information (from genus to kingdom) was linked via URIs to GBIF, TPL and WCVP. The 'Occurrence' extension consisted of 16,956 unique combinations of occurrence information. This captured the current location of the individual (e.g. record number and herbarium codes) and spatial or contextual information on the origin of each plant (e.g. coordinates, country and elevation). The integrated palm trait dataset held records for 60 different countries, distributed across all tropical regions (Appendix Fig. A.1). The 'Measurement or Fact' extension consisted of 3146 unique combinations of measurement information, e.g. by whom and when the measurement was made. The basis of record (e.g. preserved specimen or living specimen) and the bibliographic reference was also captured for each record.

Of the 50 different trait names present in the final dataset, 25 (50%) were not linked to the TO or TOP by a URI (Fig. 5b). This reflected a total of 53,787 (39%) individual trait measurements. The lack of semantic information (i.e. trait terms with a URI) was especially apparent for leaves and reproductive organs (with URIs only available for 36% and 52% of the trait names, respectively). In contrast, trait names capturing shoots and whole plant traits were covered well by URIs (75% and 100%, respectively), albeit the "Whole plant" category only contained one trait name ("Plant_height").

4. Discussion

We developed a workflow with a metadata form and two thesauri to facilitate the automated integration of quantitative plant trait measurements from heterogeneous sources. This can be applied not only to long-

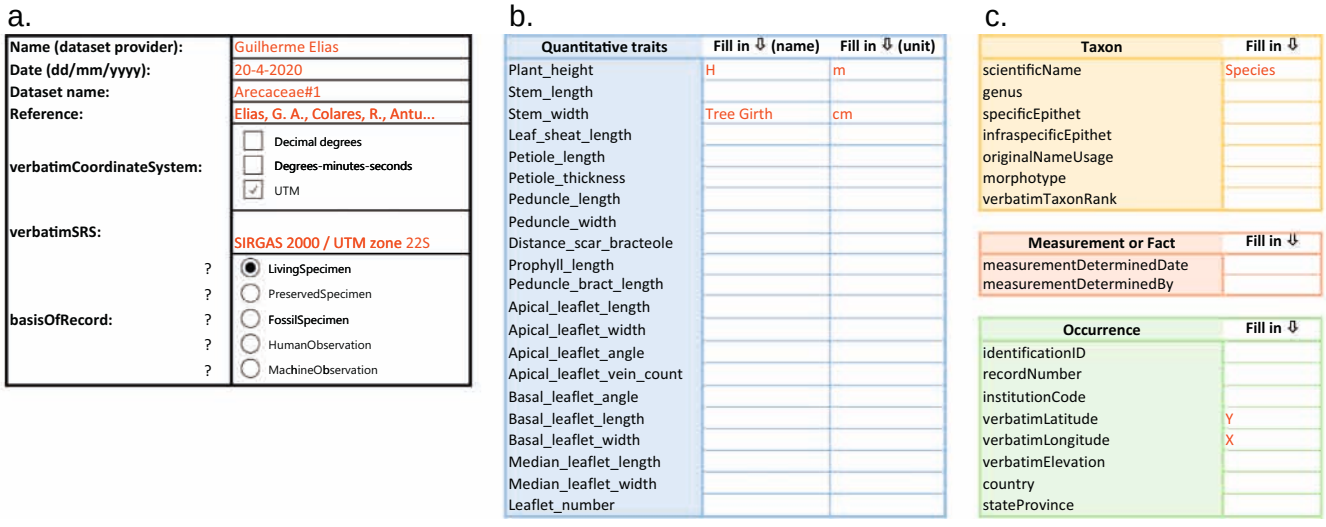


Fig. 3. The metadata form facilitates the automated integration of trait measurements from spreadsheets. The entries in red font illustrate an example from a dataset provider (G. Elias) covering spreadsheet metadata from a phytosociological study in the Brazilian Atlantic Forest (Elias et al., 2019). a. Box for metadata information representing names of data providers, date, reference, coordinate system and basis of record. A "?" can be clicked to obtain information on definitions. b. Examples of standardized trait names from the "Plant Trait Ontology" (TO) for quantitative traits. The column headers and units from spreadsheets need to be entered here, enabling the linking of trait information from spreadsheet columns to the final output dataset. c. Additional metadata from spreadsheets are covered in the taxon, measurement or fact, and occurrence fields. Column header names from spreadsheets need to be entered here to enable linking to standardized terms from the "Ecological Trait-data Standard" (ETS) and the "Darwin Core" (DwC). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

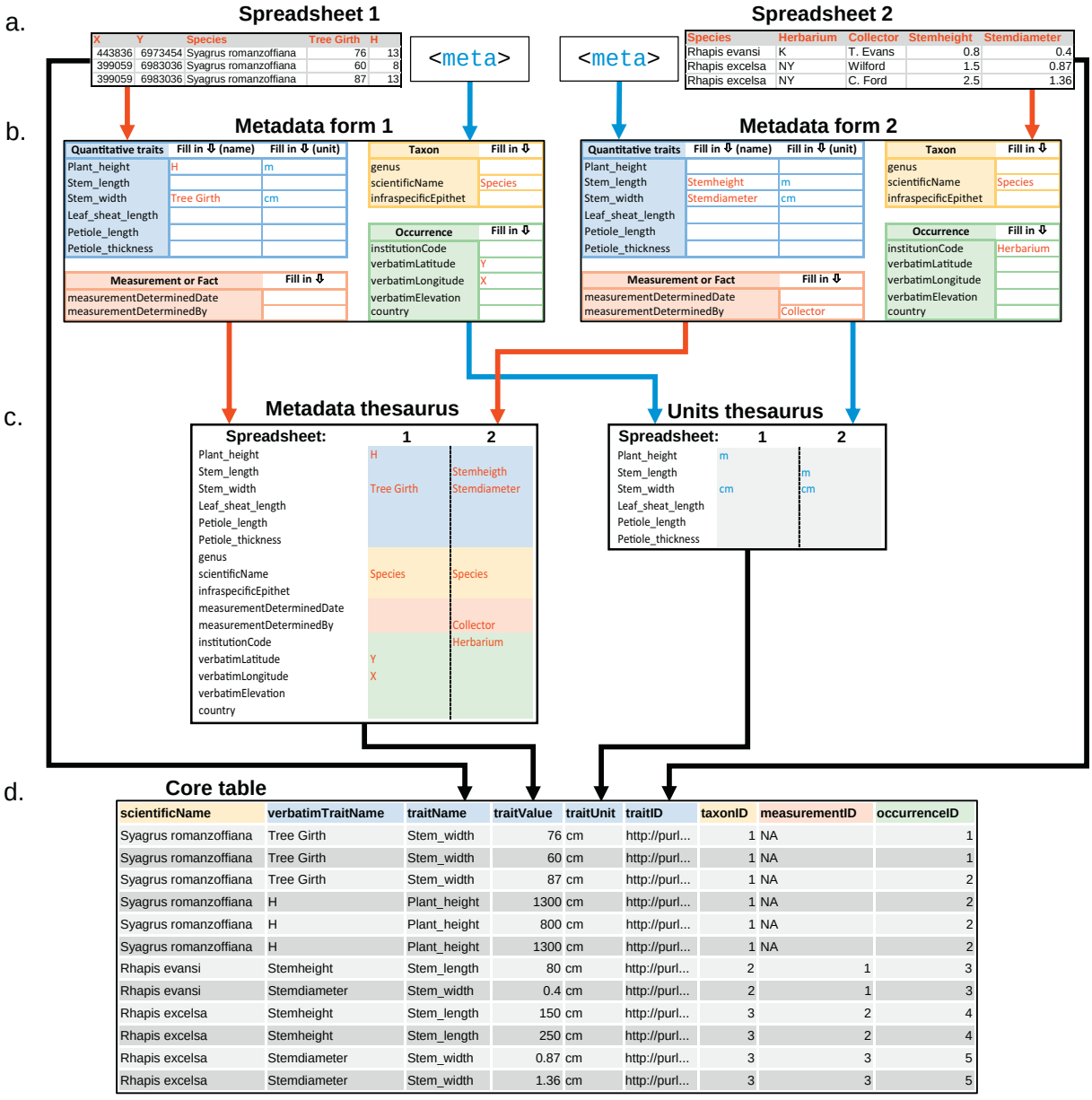


Fig. 4. Example of integrating two heterogeneous spreadsheets with palm trait measurements. The text in red and blue font illustrates how information from column headers and additional metadata (e.g. measurement units) is incorporated. a. Two example spreadsheets with three rows of data and column headers and their metadata. b. The filled out metadata forms for those two spreadsheets (information from column headers in red font and measurement units from the metadata in blue font). c. Resulting thesauri covering metadata (left) and measurement units (right). Each column captures information from one metadata form. d. Resulting core table containing the trait data of the two original spreadsheets and the standardized terminology and units of the two thesauri. This contains the scientific name, verbatim and standardized trait names and ID columns linking the measurements to information in three extension tables. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

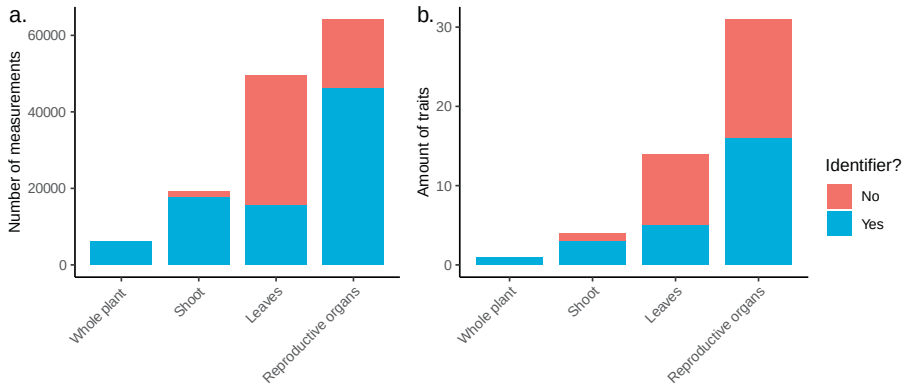


Fig. 5. Representation of traits in the integrated palm trait dataset using four main trait categories (whole plant, shoot, leaves and reproductive organs). a. Total number of individual trait measurements per trait category. b. Amount of trait names per trait category and how many are represented by terms in the "Plant Trait Ontology" or the "Thesaurus Of Plant characteristics". Red shows traits that do not have a semantic identifier (URI) and blue shows traits that do. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tail datasets but also to more standardized datasets. We illustrated this workflow with measurement spreadsheets from palm field observations and taxonomic revisions and thereby mobilized nearly 140,000 individual trait measurements from 551 palm species and 17 palm genera into a harmonized, interoperable and machine-readable format. The resulting core and extension tables are semantically linked to existing ontologies and provide individual measurements of quantitative plant traits from different plant parts (whole plant, shoot, leaves, reproductive organs) and accompanying metadata. The dataset and workflow are open and free to use and together with an open-source bookdown tutorial (https://bookdown.org/timlenters/traitData) available from GitHub (https://github.com/tlenters/traitData). The integrated dataset (with core and extension tables) is also provided on the Dryad Digital Repository (Lenters et al., 2020).

The structure of the core table in our workflow follows suggestions from Schneider et al. (2019) and contains species names, trait names, trait values and trait units. The integrated palm trait dataset that we compiled captured 50 standardized traits in the core table, especially in the "Reproductive organs" and "Leaves" categories. The most represented traits (e.g. "Stem_width", "Petiole_length", "Rachis_length", "Median_leaflet_width" and "Plant_height") were measurements that are typically used in taxonomic revisions to describe and identify species. Some of these traits (e.g. "Plant_height") also represent the most widely measured quantitative plant traits in ecology and the most widely requested data in the TRY plant trait database (Kattge et al., 2020). This reflects their high importance for ecological, evolutionary and global change research (Díaz et al., 2016). However, other widely measured quantitative plant traits which are often used to represent the global spectrum of plant form and function —e.g. seed dry mass, leaf area and leaf nitrogen content (Díaz et al., 2016; Kattge et al., 2020)— were not represented in our dataset. This partly reflects the sources of measurements of palm traits that were accessible to us (mostly herbarium specimens used for taxonomic revisions, see Table 1). It further reflects the difficulty to measure some of these traits for palms in the field, e.g. leaf area because palms have often very large leaves. Nevertheless, our workflow offers the possibility to integrate other sources of trait data in an easy and reproducible way once additional data become available.

The traits described in our workflow were semantically linked to the names and definitions of the 'Plant Trait Ontology' (TO; Arnaud et al., 2012) and the 'Thesaurus Of Plant characteristics' (TOP; Garnier et al., 2017). Although both the TO and TOP are extensive and hold many traits (1554 and 790, respectively), we could only link a semantic source to our trait names in 50% of the cases (25 out of 50 traits). This reflects that many traits captured by the TO and TOP are biochemical and development traits (e.g. "glucose content" and "shoot elongation rate") and that not all morphological plant traits are already represented with terms. Especially traits that require a precise description of the position of a specific plant segment were underrepresented in existing ontologies and thesauri (e.g. "Staminate_rachilla_length" and "Median_leaflet_width"). In our dataset, this was the case for 20 of the 25 traits (80%) that had no semantic source in the TO or TOP. This shows that current plant trait ontologies could be substantially improved by providing more precise trait descriptions, e.g. by including the specific position (apical, basal, median) of leaflets or rachillae (flower-bearing branches). In palms (and many other plants), the morphological measurements (e.g. length, width, thickness, angle) of leaves and inflorescences substantially differ depending on which part is measured (Dransfield et al., 2008). To improve ontologies, we propose as a starting point to define the trait names of the 25 traits that were not represented in existing ontologies (Appendix Table A.2). These definitions follow the hierarchical structure and semantic standards used in current ontologies, and we additionally provide the URI for each term of each trait description, using trait definitions from the palm book 'Genera Palmarum' (Dransfield et al., 2008). This could be used by semantic developers to expand the terminology of existing ontologies and thesauri such as the TO and TOP.

The three extension tables captured metadata of plant trait measurements related to taxonomy, provenance and spatial occurrence. To harmonize taxonomy in an automated workflow requires up-to-date taxonomic databases that are accessible via web services and tools. We used taxonomic information from GBIF, The Plant List (TPL) and the World Checklist of Vascular Plants (WCVP) and compared their taxonomic coverage. Of the 617 unique verbatim palm species captured in our compiled palm trait dataset (roughly 1/5th of all ca. 2500 palm species), about 89%, 78% and 87% of the verbatim species names were found in GBIF, TPL and WCVP databases, respectively. The matched names are provided in the taxon extension table for all three databases. The species names not matched to a taxonomic database were kept with verbatim names, but higher taxonomic levels (kingdom, phylum, class, order) were left blank in the integrated dataset. The WCVP provided by Royal Botanic Gardens, Kew, probably contains the most up-to-date and complete taxonomy for palms because Kew has a long tradition in palm taxonomy and systematics (Baker and Dransfield, 2016; Dransfield et al., 2008) and maintains both the World Checklist of Palms (Govaerts and Dransfield, 2005) and the web portal Palmweb (http://www.palmweb.org). However, neither WCVP nor GBIF or TPL provided subfamily or tribe information although this is in principle available for palms (Govaerts and Dransfield, 2005). Moreover, taxonomic information provided by Kew is only digitally available via the WCVP website without any R package or application programming interface (API). The R script from our workflow therefore incorporates functions that download the database from the WCVP website and loads it in the R environment automatically. This method is not robust, as it uses a static version of the database, which would have to be changed manually in case of an update. In contrast, the two other taxonomic sources (GBIF and The Plant List) provide R packages for an automated integration, but TPL is no longer updated (explaining its lower name matching). Two additional taxonomic databases also provide R packages and are recently updated, namely the Leipzig Catalogue of Vascular Plants (LCVP; Freiberg et al., 2020) and the World Flora Online (WFO; Kindt, 2020). However, we did not include them in the workflow because LCVP currently lacks URIs for the standardized species names (thus not providing

semantic links) and WFO uses a "fuzzy matching" method for name identification which led to a large amount of incorrect species name matches.

The provenance information captured in the measurement or fact extension table provides information on references and sources of measurements (Schneider et al., 2019). Most of the palm trait measurements were captured from herbarium specimen ("PreservedSpecimen" as the basis of record. Measurements obtained in the field ("LivingSpecimen") can differ (i.e. be higher) than those measured in the herbarium because most parts of a plant shrink on drying (Parnell et al., 2013). This may be especially important for measurements of leaves (Queenborough and Porras, 2014) and fruits (Parnell et al., 2013). More comprehensive plant trait databases capturing individual trait measurements from both the field and herbaria (or wet vs. dry) would allow to quantify shrinkage and thus increase the comparability of trait measurements from different sources. Another issue is that measurement protocols for palm traits may differ among researchers. This could be captured at the record level with the DwC field 'measurementMethod', but currently there are no specific trait sampling protocols for palms. While a general handbook for standardized measurements of plant functional traits could serve as a starting point for sampling palm traits (Pérez-Harguindeguy et al., 2013), not all traits are described there and the unique morphology of palms (see glossary of Dransfield et al., 2008) may also require to define particular sampling protocols for palms. This would increase the equivalency and comparability of plant functional traits for palms.

The occurrence extension table of our dataset captured geographic information such as latitude, longitude and country. This information indicated coverage across 60 countries within tropical regions globally (Appendix Fig. A.1). The highest representation was captured in the Neotropics and South-East Asia. These are regions with the highest species richness of palms (Kissling et al., 2012), but many individual trait measurements from other regions (and species) are still missing, despite a relatively good coverage of species-level trait information (i.e. averages, minima, maxima) available from books and published scientific literature (Kissling et al., 2019). Data mobilization from data holders and integration of other unstandardized (long-tail) datasets capturing individual palm trait measurements would therefore strongly increase the geographic coverage and help to fill spatial gaps in large-scale biodiversity knowledge (Hortal et al., 2015).

There remain big challenges for integrating long-tail data of individual trait measurements and making them findable, accessible, interoperable and reusable (FAIR guiding principles; Wilkinson et al., 2016). Many raw datasets are neither published (e.g. via digital repositories) nor otherwise digitally available which inhibits findability and reusability of the data (Heidorn, 2008). The mobilization of long-tail datasets should therefore have high priority before data collectors retire or die which inevitably will lead to the ultimate loss of raw data and metadata (Michener et al., 1997). This requires sociological change such as incentives and reducing barriers to data sharing through citation and use metrics (Costello et al., 2013) and through supporting education and establishing community standards (Kattge et al., 2020; Michener, 2015). Ultimately, this would reduce research costs, improve collaborative efforts and increase research opportunities (Uhlir and Schröder, 2007). Many raw data records and long-tail datasets may not even be available in a digital spreadsheet (e.g. only as field notes), or if they are, contain ample formatting errors (e.g. as highlighted in Fig. 3). This will require a lot of manual work before an automated data integration is possible. Documentation of metadata is also often done insufficiently or in ways that are not machine-readable, e.g. lack of information on sampling protocols or stating the measurement units in the value cell of spreadsheets. Our metadata form facilitates the automated integration of individual trait measurements from spreadsheets but also relies on the willingness of data providers to spend time for entering the metadata. It is therefore important to engage specific scientific communities (e.g. taxonomists and ecologists working on specific plant families) when integrating and harmonizing trait data, and to promote grassroots initiatives and other bottom-up collaborative data integration projects (Aubin et al., 2020).

5. Conclusion

Our workflow provides an open-access resource for integrating and harmonizing individual-level trait measurements of plants into a machine-readable and interoperable format, and the integrated palm trait dataset gives an example of how new plant trait data can be mobilized. Such efforts contribute to mitigating the big shortfalls in large-scale biodiversity knowledge, namely the Raunkiæran shortfall which represents our limited knowledge of species traits (Hortal et al., 2015), especially in tropical regions (Kattge et al., 2020). The current workflow and metadata form can be applied to other plant taxa for harmonizing and integrating long-tail trait datasets. Moreover, it can also be applied to taxa other than plants. This would require to use an empty metadata form and thesauri as template, to enter trait names from relevant ontologies, and to provide links to databases for taxonomic standardization within the workflow. Extending such efforts will ultimately allow to analyse intraspecific trait variability across space and time. This is crucial for many research topics in ecology, evolution and global change biology, including the prediction of biodiversity change (Bjorkman et al., 2018) or assessments of how intraspecific trait variation shapes organisms fitness and performance (Kumordzi et al., 2019) or species interactions (Tielens and Gruner, 2020). Moreover, intraspecific trait variation is also useful for applying trait imputation and gap filling methods (Schrodt et al., 2015). A key aspect of future work should be the extension and further development of ontologies, thesauri and metadata standards which provide the fundamental resources for semantic integration of biodiversity knowledge in the era of big data.

Data availability

The raw data files, the integrated dataset (core table and extension tables), the metadata form, the thesauri, the R script (workflow) and the bookdown tutorial (https://bookdown.org/timlenters/traitData) are available from GitHub (https://github.com/tlenters/traitData). The integrated dataset (with core and extension tables) is also provided on the Dryad Digital Repository (Lenters et al., 2020).

CRediT authorship contribution statement

Tim P. Lenters: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Writing - review & editing. Andrew Henderson: Data curation, Validation, Writing - review & editing. Caroline M. Dracxler: Data curation, Validation, Writing - review & editing. Guilherme A. Elias: Data curation, Writing - review & editing. Suzanne Mogue Kamga: Data curation. Thomas L.P. Couvreur: Data curation, Writing - review & editing. W. Daniel Kissling: Conceptualization, Methodology, Validation, Writing - original draft, Writing - review & editing, Funding acquisition, Supervision.

Declaration of competing interest

None.

Acknowledgements

W. Daniel Kissling was supported by the University of Amsterdam (starting grant), the Faculty Research Cluster 'Global Ecology', and the Netherlands Organisation for Scientific Research (grant 824.15.007).

Appendix A

Table A.1
All 50 traits present in the final palm trait dataset. Standardized trait names are given in the "traitName" column, "category" shows the trait category as derived from TRY plant trait database, "count" gives the number of measurements for each trait and "traitID" the Uniform Resource Identifier (URI) connected to the "Plant Trait Ontology" (TO) or the "Thesaurus Of Plant characteristics" (TOP). Empty fields in the "traitID" column indicate that the trait name had no semantic source in the TO or TOP.

traitName	category	count	traitID
Stem_width	Shoot	10,396	http://purl.obolibrary.org/obo/TO_0001035
Petiole_length	Leaves	7103	http://purl.obolibrary.org/obo/TO_0000766
Rachis_length	Reproductive organs	6665	http://purl.obolibrary.org/obo/TO_0001072
Median_leaflet_width	Leaves	6390	
Plant_height	Whole plant	6118	http://purl.obolibrary.org/obo/TO_0000207
Rachilla_thickness	Reproductive organs	5894	
Median_leaflet_length	Leaves	5752	
Stem_length	Shoot	5700	http://purl.obolibrary.org/obo/TO_0000576
Rachis_thickness	Reproductive organs	5483	http://purl.obolibrary.org/obo/TO_0001061
Rachilla_count	Reproductive organs	5482	http://purl.obolibrary.org/obo/TO_0000954
Rachilla_length	Reproductive organs	5350	http://purl.obolibrary.org/obo/TO_0000972
Leaflet_number	Leaves	5272	http://purl.obolibrary.org/obo/TO_0002636
Peduncle_width	Reproductive organs	5175	http://purl.obolibrary.org/obo/TO_0000649
Inflorescence_branch_count	Reproductive organs	4359	http://purl.obolibrary.org/obo/TO_0000050
Fruit_width	Reproductive organs	4296	http://purl.obolibrary.org/obo/TO_0002627
Fruit_length	Reproductive organs	4295	http://purl.obolibrary.org/obo/TO_0002626
Apical_leaflet_angle	Leaves	4207	
Basal_leaflet_angle	Leaves	4158	
Basal_leaflet_width	Leaves	3658	
Apical_leaflet_length	Leaves	3641	
Peduncle_length	Reproductive organs	3618	http://purl.obolibrary.org/obo/TO_0002691
Pistillate_rachilla_length	Reproductive organs	3361	
Apical_leaflet_width	Leaves	3072	
Distance_scar_bracteole	Reproductive organs	2965	
Basal_leaflet_length	Leaves	2878	
Staminate_rachilla_length	Reproductive organs	1797	
Prophyll_length	Shoot	1649	
Leaf_number	Leaves	1474	http://purl.obolibrary.org/obo/TO_0000241
Peduncle_bract_length	Reproductive organs	1442	
Shoot_axis_internode_length	Shoot	1440	http://purl.obolibrary.org/obo/TO_0000145
Leaf_sheat_length	Leaves	1429	http://purl.obolibrary.org/obo/TO_0002689
Pistillate_inflorescence_length	Reproductive organs	821	
Stamen_count	Reproductive organs	590	http://purl.obolibrary.org/obo/TO_0000225
Staminate_inflorescence_length	Reproductive organs	516	
Seed_width	Reproductive organs	415	http://purl.obolibrary.org/obo/TO_0000149
Staminate_flower_length	Reproductive organs	362	
Basal_rachilla_length	Reproductive organs	275	
Apical_leaflet_vein_count	Leaves	275	
Median_rachilla_length	Reproductive organs	260	
Apical_rachilla_length	Reproductive organs	252	
Seed_count	Reproductive organs	208	http://purl.obolibrary.org/obo/TO_0000445
Bracts_number	Reproductive organs	168	http://purl.obolibrary.org/obo/TO_0006028
Petiole_thickness	Leaves	151	http://top-thesaurus.org/trait#TOP59
Staminate_rachilla_count	Reproductive organs	89	
Pistillate_rachilla_count	Reproductive organs	42	
Basal_rachilla_thickness	Reproductive organs	29	
Seed_length	Reproductive organs	8	http://purl.obolibrary.org/obo/TO_0000146
Inflorescence_length	Reproductive organs	7	http://purl.obolibrary.org/obo/TO_0000271
Inflorescence_width	Reproductive organs	4	http://purl.obolibrary.org/obo/TO_0000804
Apical_rachilla_thickness	Reproductive organs	2	

Table A.2
Proposed definition of 25 traits from the palm trait dataset that had no semantic source in existing ontologies (compare Table A.1). Trait names are given in the "traitName" column and the proposed definitions together with the Uniform Resource Identifiers (URI) in the "definition" column.

traitName	definition
Distance_scar_bracteole	Distance (PATO:0000040) between peduncular bract (PO:0009055) and prophyll (PO:0009042) insertion
Prophyll_length	Length (PATO:0000122) of the prophyll (PO:0009042)
Peduncle_bract_length	Length (PATO:0000122) of empty bracts (PO:0009055) on the main inflorescence axis (PO:0020122) between the prophyll (PO:0009042) and the first rachis (PO:0020055) bract (PO:0009055)
Apical_leaflet_length	Length (PATO:0000122) of the apical (EFO:0001653) leaflet (PO:0020049)
Apical_leaflet_width	Width (PATO:0000921) of the apical (EFO:0001653) leaflet (PO:0020049)
Apical_leaflet_angle	Angle (PATO:0002326) of the apical (EFO:0001653) leaflet (PO:0020049)
Apical_leaflet_vein_count	Count (PATO:0000070) of the apical (EFO:0001653) leaflet veins (BTO:0005515)
Basal_leaflet_angle	Angle (PATO:0002326) of the basal (EFO:0001654) leaflet (PO:0020049)
Basal_leaflet_length	Length (PATO:0000122) of the basal (EFO:0001654) leaflet (PO:0020049)
Basal_leaflet_width	Width (PATO:0000921) of the basal (EFO:0001654) leaflet (PO:0020049)

(*continued on next page*)

Table A.2 (*continued*)

traitName	definition
Median_leaflet_length	Length (PATO:0000122) of the median (EFO:0001660) leaflet (PO:0020049)
Median_leaflet_width	Width (PATO:0000921) of the median (EFO:0001660) leaflet (PO:0020049)
Rachilla_thickness	Thickness (PATO:0000915) of the rachilla (PO:0009080)
Staminate_rachilla_length	Length (PATO:0000122) of the staminate (PO:0025601) rachilla (PO:0009080)
Staminate_rachilla_count	Count (PATO:0000070) of the staminate (PO:0025601) rachilla (PO:0009080)
Pistillate_rachilla_length	Length (PATO:0000122) of the pistillate (PO:0025598) rachilla (PO:0009080)
Pistillate_rachilla_count	Count (PATO:0000070) of the pistillate (PO:0025598) rachilla (PO:0009080)
Basal_rachilla_length	Length (PATO:0000122) of the basal (EFO:0001654) rachilla (PO:0009080)
Basal_rachilla_thickness	Thickness (PATO:0000915) of the basal (EFO:0001654) rachilla (PO:0009080)
Apical_rachilla_length	Length (PATO:0000122) of the apical (EFO:0001653) rachilla (PO:0009080)
Apical_rachilla_thickness	Thickness (PATO:0000122) of the apical (EFO:0001653) rachilla (PO:0009080)
Median_rachilla_length	Length (PATO:0000122) of the median (EFO:0001660) rachilla (PO:0009080)
Staminate_inflorescence_length	Length (PATO:0000122) of the staminate inflorescence (PO:0025601)
Pistillate_inflorescence_length	Length (PATO:0000122) of the pistillate inflorescence (PO:0025598)
Staminate_flower_length	Length (PATO:0000122) of the staminate flower (PO:0025600)

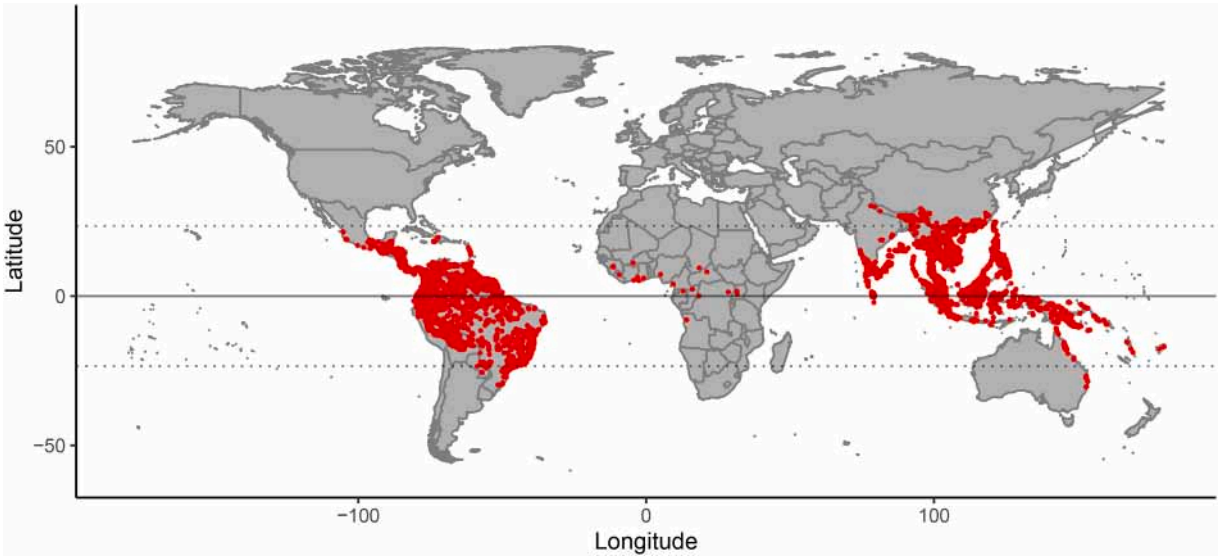


Fig. A.1. Map of the locations (n = 15,054) of all palm trait measurements present in the final palm trait dataset where valid coordinates were provided (89%). Locations are shown in red, the equator is indicated by a solid line and the Tropics of Cancer and Capricorn are indicated by dotted lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

References

Arnaud, E., Cooper, L., Shrestha, R., Menda, N., Nelson, R.T., Matteis, L., et al., 2012. Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes. In: Proceedings of the International Conference on Knowledge Engineering and Ontology Development, KEOD-2012, pp. 220–225.

Aubin, I., Cardou, F., Boisvert-Marsh, L., Garnier, E., Strukelj, M., Munson, A.D., 2020. Managing data locally to answer questions globally: the role of collaborative science in ecology. J. Veg. Sci. 31 (3), 509–517.

Baker, W.J., Dransfield, J., 2016. Beyond Genera Palmarum: progress and prospects in palm systematics. Bot. J. Linn. Soc. 182 (2), 207–233.

de Bello, F., Lavorel, S., Díaz, S., Harrington, R., Cornelissen, J.H., Bardgett, R.D., et al., 2010. Towards an assessment of multiple ecosystem processes and services via functional traits. Biodiversity and Conservation 19 (10), 2873–2893.

Bivand, R., Keitt, T., Rowlingson, B., 2019. rgdal: bindings for the 'Geospatial' data abstraction library. In: R Package Version 1.4-3. https://CRAN.R-project.org/package=rgdal.

Bjorkman, A.D., Myers-Smith, I.H., Elmendorf, S.C., Normand, S., Rüger, N., Beck, P.S., Georges, D., 2018. Plant functional trait change across a warming tundra biome. Nature 562 (7725), 57–62.

Campetella, G., Botta-Dukát, Z., Wellstein, C., Canullo, R., Gatto, S., Chelli, S., et al., 2011. Patterns of plant trait–environment relationships along a forest succession chronosequence. Agriculture, Ecosystems & Environment 145 (1), 38–48.

Chen, M., Mao, S., Liu, Y., 2014. Big data: a survey. Mobile Netw. Appl. 19 (2), 171–209.

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.Q., Bourne, P.E., 2013. Biodiversity data should be published, cited, and peer reviewed. Trends Ecol. Evol. 28 (8), 454–461.

Couvreur, T.L., Baker, W.J., 2013. Tropical rain forest evolution: palms as a model group. BMC Biol. 11 (1), 48.

Díaz, S., Purvis, A., Cornelissen, J.H., Mace, G.M., Donoghue, M.J., Ewers, R.M., et al., 2013. Functional traits, the phylogeny of function, and ecosystem service vulnerability. Ecol. Evol. 3 (9), 2958–2975.

Díaz, S., Kattge, J., Cornelissen, J.H., Wright, I.J., Lavorel, S., Dray, S., et al., 2016. The global spectrum of plant form and function. Nature 529 (7585), 167–171.

Dietze, M.C., 2017. Ecological Forecasting. Princeton University Press, Princeton.

Dowle, M., Srinivasan, A., 2019. data.table: Extension of 'data.frame'. R Package Version 1.12.2. https://CRAN.R-project.org/package=data.table.

Dransfield, J., Uhl, N.W., Lange, C.B.A., Baker, W.J., Harley, M.M., Lewis, C.E., 2008. Genera Palmarum: The Evolution and Classification of Palms. Kew Publishing, London.

Eiserhardt, W.L., Svenning, J.C., Kissling, W.D., Balslev, H., 2011. Geographical ecology of the palms (Arecaceae): determinants of diversity and distributions across spatial scales. Ann. Bot. 108 (8), 1391–1416.

Elias, G.A., Colares, R., Antunes, A.R., Padilha, P.T., Lima, J.M.T., Santos, R., 2019. Palm (Arecaceae) communities in the Brazilian Atlantic forest: a phytosociological study. Floresta Ambiente 26 (4).

Farley, S.S., Dawson, A., Goring, S.J., Williams, J.W., 2018. Situating ecology as a big-data science: current advances, challenges, and solutions. BioScience 68 (8), 563–576.

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bull. Ecol. Soc. Am. 86 (3), 158–168.

Freiberg, M., Winter, M., Gentile, A., Zizka, A., Muellner-Riehl, A.N., Weigelt, A., Wirth, C., 2020. The Leipzig Catalogue of Vascular Plants (LCVP)–an improved taxonomic reference list for all known vascular plants. BioRxiv. https://doi.org/10.1101/2020.05.08.077149.

Gallagher, R.V., Falster, D.S., Maitner, B.S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., et al., 2020. Open Science principles for accelerating trait-based science across the Tree of Life. Nat. Ecol. Evol. 4, 1–10.

Garnier, E., Stahl, U., Laporte, M.A., Kattge, J., Mougenot, I., Kühn, I., Bunker, D.E., 2017. Towards a thesaurus of plant characteristics: an ecological contribution. J. Ecol. 105 (2), 298–309.
GBIF.org, 2020. GBIF Home Page. Available from: https://www.gbif.org. Retrieved 22 July 2020.
Geijzendorffer, I.R., Regan, E.C., Pereira, H.M., Brotons, L., Brummitt, N., Gavish, Y., et al., 2016. Bridging the gap between biodiversity data and policy reporting needs: An Essential Biodiversity Variables perspective. Journal of Applied Ecology 53 (5), 1341–1350.
Gerstner, K., Moreno-Mateos, D., Gurevitch, J., Beckmann, M., Kambach, S., Jones, H.P., Seppelt, R., 2017. Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. Methods Ecol. Evol. 8, 777–784.
Govaerts, R., Dransfield, J., 2005. World Checklist of Palms. Royal Botanic Gardens, London.
Gruber, T.R., 1995. Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum. Comput. Stud. 43 (5–6), 907–928.
Guralnick, R.P., Zermoglio, P.F., Wieczorek, J., LaFrance, R., Bloom, D., Russell, L., 2016. The importance of digitized biocollections as a source of trait data and a new VertNet resource. Database 2016, baw158.
Hardisty, A.R., Michener, W.K., Agosti, D., García, E.A., Bastin, L., Belbin, L., et al., 2019. The Bari Manifesto: An interoperability framework for essential biodiversity variables. Ecol. Inf. 49, 22–31.
Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Libr. Trends 57 (2), 280–299.
Henderson, A., 2002. Evolution and Ecology of Palms. The New York Botanical Garden Press, Bronx.
Henderson, A., 2004. A multivariate analysis of *Hyospathe* (Palmae). Am. J. Bot. 9 (16), 953–965.
Henderson, A., 2005. A multivariate study of *Calyptrogyne* (Palmae). Syst. Bot. 30 (1), 60–83.
Henderson, A., 2011a. A revision of *Desmoncus* (Arecaceae). Phytotaxa 35, 1–88.
Henderson, A., 2011b. A revision of *Geonoma* (Arecaceae). Phytotaxa. 17, 1–271.
Henderson, A., 2011c. A revision of *Leopoldinia* (Arecaceae). Phytotaxa 32, 1–17.
Henderson, A., 2012. A revision of *Pholidostachys* (Arecaceae). Phytotaxa 43, 1–48.
Henderson, A., 2015. A revision of *Chuniophoenix* (Arecaceae). Phytotaxa 218, 163–170.
Henderson, A., 2016. A revision of *Rhapis* (Arecaceae). Phytotaxa 258, 137–152.
Henderson, A., 2020a. A revision of *Attalea* (Arecaceae, Arecoideae, Cocoseae, Attaleinae). Phytotaxa 444 (1), 1–76.
Henderson, A., 2020b. A revision of *Calamus* (Arecaceae, Calamoideae, Calameae, Calaminae). Phytotaxa 445 (1), 1–656.
Henderson, A., Ferreira, E., 2002. A morphometric study of *Synechanthus* (Palmae). Syst. Bot. 27, 693–702.
Henderson, A., Villalba, I., 2013. A revision of *Welfia* (Arecaceae). Phytotaxa 119, 33–44.
Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M., Lobo, J.M., Ladle, R.J., 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. Annu. Rev. Ecol. Evol. Syst. 46, 523–549.
Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E.A., McCouch, S., Pujar, A., et al., 2005. Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp. Function. Genom. 6 (7-8), 388–397.
Kalwij, J.M., 2012. Review of 'the plant list, a working list of all plant species'. J. Veg. Sci. 23 (5), 998–1002.
Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I.C., Leadley, P., et al., 2020. TRY plant trait database–enhanced coverage and open access. Global Change Biology 26, 119–188.
Kindt, R., 2020. WorldFlora: an R package for exact and fuzzy matching of plant names against the World Flora Online Taxonomic Backbone data. bioRxiv. https://doi.org/10.1101/2020.02.02.930719v1.
Kissling, W.D., Baker, W.J., Balslev, H., Barfod, A.S., Borchsenius, F., Dransfield, J., et al., 2012. Quaternary and pre-Quaternary historical legacies in the global distribution of a major tropical plant lineage. Global Ecology and Biogeography 21 (9), 909–921.
Kissling, W.D., Walls, R., Bowser, A., Jones, M.O., Kattge, J., Agosti, D., et al., 2018. Towards global data products of Essential Biodiversity Variables on species traits. Nat. Ecol. Evol. 2 (10), 1531–1540.
Kissling, W.D., Balslev, H., Baker, W.J., Dransfield, J., Göldel, B., Lim, J.Y., et al., 2019. PalmTraits 1.0, a species-level functional trait database of palms worldwide. Sci. Data 6 (1), 1–13.
Kumordzi, B.B., Aubin, I., Cardou, F., Shipley, B., Violle, C., Johnstone, J., et al., 2019. Geographic scale and disturbance influence intraspecific trait variability in leaves and roots of North American understorey plants. Functional Ecology 33 (9), 1771–1784.
LaDeau, S.L., Han, B.A., Rosi-Marshall, E.J., Weathers, K.C., 2017. The next decade of big data in ecosystem science. Ecosystems 20 (2), 274–283.
Lavorel, S., Garnier, E., 2002. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. Funct. Ecol. 16 (5), 545–556.
Lenters, T.P., Henderson, A., Dracxler, C.M., Elias, G.A., Mogue Kamga, S., Couvreur, T.L. P., Kissling, W.D., 2020. Data from: integration and harmonization of trait data from plant individuals across heterogeneous sources. In: Dryad, Dataset. https://doi.org/10.5061/dryad.tdz08kpz0.
McGill, B.J., Enquist, B.J., Weiher, E., Westoby, M., 2006. Rebuilding community ecology from functional traits. Trends Ecol. Evol. 21 (4), 178–185.
Michener, W.K., 2015. Ecological data sharing. Ecol. Inf. 29, 33–44.
Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G., 1997. Nongeospatial metadata for the ecological sciences. Ecol. Appl. 7 (1), 330–342.
Miller-Rushing, A.J., Primack, R.B., Primack, D., Mukunda, S., 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. Am. J. Bot. 93 (11), 1667–1674.
Muñoz, G., Trøjelsgaard, K., Kissling, W.D., 2019. A synthesis of animal-mediated seed dispersal of palms reveals distinct biogeographical differences in species interactions. J. Biogeogr. 46 (2), 466–484.
Nascimento, L.F.D., Guimarães, P.R., Onstein, R.E., Kissling, W.D., Pires, M.M., 2020. Associated evolution of fruit size, fruit colour and spines in Neotropical palms. J. Evol. Biol. 33 (6), 858–868.
Onstein, R.E., Baker, W.J., Couvreur, T.L., Faurby, S., Svenning, J.C., Kissling, W.D., 2017. Frugivory-related traits promote speciation of tropical palms. Nat. Ecol. Evol. 1 (12), 1903–1911.
Parnell, J., Rich, T., McVeigh, A., Lim, A., Quigley, S., Morris, D., Wong, Z., 2013. The effect of preservation methods on plant morphology. Taxon 62 (6), 1259–1265.
Parr, C.S., Schulz, K.S., Hammock, J., Wilson, N., Leary, P., Rice, J., Corrigan Jr., R.J., 2016. TraitBank: practical semantics for organism attribute data. Semantic Web 7 (6), 577–588.
Pérez-Harguindeguy, N., Díaz, S., Garnier, E., Lavorel, S., Poorter, H., Jaureguiberry, P., Bret-Harte, M.S., Cornwell, W.K., Craine, J.M., Gurvich, D.E., Urcelay, C., Veneklaas, et al., 2013. New handbook for standardised measurement of plant functional traits worldwide. Aust. J. Bot. 61, 167–234.
Queenborough, S.A., Porras, C., 2014. Expanding the coverage of plant trait databases–a comparison of specific leaf area derived from fresh and dried leaves. Plant Ecol. Divers. 7 (1–2), 383–388.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria.
Robbirt, K.M., Davy, A.J., Hutchings, M.J., Roberts, D.L., 2011. Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid *Ophrys sphegodes*. J. Ecol. 99 (1), 235–241.
Schleuning, M., Neuschulz, E.L., Albrecht, J., Bender, I.M., Bowler, D.E., Dehling, D.M., et al., 2020. Trait-based assessments of climate-change impacts on interacting species. Trends in Ecology & Evolution 35, 319–328.
Schneider, F.D., Fichtmueller, D., Gossner, M.M., Güntsch, A., Jochum, M., König-Ries, B., et al., 2019. Towards an ecological trait-data standard. Methods in Ecology and Evolution 10 (12), 2006–2019.
Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., et al., 2015. BHPMF–a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. Global Ecology and Biogeography 24 (12), 1510–1521.
Siefert, A., Violle, C., Chalmandrier, L., Albert, C.H., Taudiere, A., Fajardo, A., et al., 2015. A global meta-analysis of the relative extent of intraspecific trait variation in plant communities. Ecol. Lett. 18 (12), 1406–1419.
Tielens, E.K., Gruner, D.S., 2020. Intraspecific variation in host plant traits mediates taxonomic and functional composition of local insect herbivore communities. Ecol. Entomol. https://doi.org/10.1111/een.12923 in press.
Uhlir, P.F., Schröder, P., 2007. Open data for global science. Data Sci. J. 6, OD36–OD53.
WCVP, 2020. World Checklist of Vascular Plants, version 2.0. Facilitated by the Royal Botanic Gardens, Kew. Published on the internet. http://wcvp.science.kew.org/.
Westoby, M., Wright, I.J., 2006. Land-plant ecology on the basis of functional traits. Trends Ecol. Evol. 21 (5), 261–268.
Wickham, H., François, R., Henry, L., Müller, K., 2019. dplyr: A Grammar of Data Manipulation. R Package Version 0.8.1. https://CRAN.R-project.org/package=dplyr.
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al., 2012. Darwin Core: an evolving community-developed biodiversity data standard. PloS One 7 (1), e29715.
Wilke, B.J., Snapp, S.S., 2008. Winter cover crops for local ecosystems: linking plant traits and ecosystem function. J. Sci. Food Agric. 88 (4), 551–557.
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018.
Wright, I.J., Dong, N., Maire, V., Prentice, I.C., Westoby, M., Díaz, S., et al., 2017. Global climatic drivers of leaf size. Science 357 (6354), 917–921.