Methods in Ecology and Evolution

DR FLORIAN D. SCHNEIDER (Orcid ID : 0000-0002-1494-5684)

DAVID FICHTMUELLER (Orcid ID : 0000-0002-0829-5849)

DR MARTIN M. MAXIMILIAN GOSSNER (Orcid ID : 0000-0003-1516-6364)

DR MALTE JOCHUM (Orcid ID : 0000-0002-8728-1145)

PROFESSOR BIRGITTA KÖNIG-RIES (Orcid ID : 0000-0002-2382-9722)

DR CATERINA PENONE (Orcid ID : 0000-0002-8170-6659)

DR PETER MANNING (Orcid ID : 0000-0002-7940-2023)

MR ANDREAS OSTROWSKI (Orcid ID : 0000-0002-2033-779X)

DR NADJA K. SIMONS (Orcid ID : 0000-0002-2718-7050)

Article type : Review

Handling editor: David Orme

Towards an Ecological Trait-data Standard

Florian D. Schneider[*, †], David Fichtmueller[‡], Martin M. Gossner[§],
Anton Güntsch[‡], Malte Jochum[**, ††], Birgitta König-Ries[‡‡],

[*] Corresponding author: *florian.dirk.schneider@gmail.com*

[†] unaffiliated, c/o Birgitta König-Ries, Department of Mathematics and Computer Science, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

[‡] Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin, Berlin, Germany

[§] Forest Entomology, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

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

[††] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany & Leipzig University, Institute of Biology, Leipzig, Germany & Leipzig University, Institute of Biology, Leipzig, Germany

Gaëtane Le Provost[§§], Pete Manning[§§], Andreas Ostrowski[‡‡],
Caterina Penone[**], Nadja K. Simons[***†††]

running title: Towards an Ecological Trait-data Standard

Abstract

1. Trait-based approaches are widespread throughout ecological research as they offer great potential to achieve a general understanding of a wide range of ecological and evolutionary mechanisms. Accordingly, a wealth of trait data is available for many organism groups, but this data is underexploited due to a lack of standardisation and heterogeneity in data formats and definitions.

2. We review current initiatives and structures developed for standardising trait data and discuss the importance of standardisation for trait data hosted in distributed open-access repositories.

3. In order to facilitate the standardisation and harmonisation of distributed trait datasets by data providers and data users, we propose a standardised vocabulary that can be used for storing and sharing ecological trait data. We discuss potential incentives and challenges for the wide adoption of such a standard by data providers.

4. The use of a standard vocabulary allows for trait datasets from heterogeneous sources to be aggregated more easily into compilations and facilitates the creation of interfaces between software tools for trait-data handling and analysis. By aiding decentralised trait-data standardisation, our vocabulary may ease data integration and use of trait data for a broader ecological research community and enable global syntheses across a wide range of taxa and ecosystems.

[‡‡] Department of Mathematics and Computer Science, Friedrich-Schiller-Universität Jena, Jena, Germany

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

[***] Department of Ecology and Ecosystem Management, Technische Universität München, Freising, Germany

[†††] Ecological Networks, Department of Biology, Technische Universität Darmstadt, Darmstadt, Germany

Key-words:

functional ecology, species traits, semantic web, ontologies, data standardisation, ecoinformatics

Introduction

Functional traits are phenotypic (i.e., morphological, physiological, behavioural) characteristics that are related to the fitness and performance of an organism (McGill et al., 2006; Violle et al., 2007). Recent years have seen a proliferation of trait-based research in a wide range of fields: trait data have been used to understand the evolutionary basis of individual-level properties (Salguero-Gómez et al., 2016), global patterns of biodiversity (Díaz et al., 2016), and the relationship between ecosystem functions and the functional composition of species assemblages (Bello et al., 2010; Mouillot et al., 2013). This research provides the mechanistic framework for linking climate change or anthropogenic land use to biodiversity and its related functions (Díaz et al., 2011; Lavorel & Grigulis, 2012; Allan et al., 2015). Species traits have been suggested as indicator variables for monitoring ecosystem health at the individual level, like for instance changes in average body sizes in a population of fish (Kissling et al., 2018). Because functional traits allow us to infer the ecological role of organisms from their apparent features, regardless of their taxonomic identity (Grime, 2001; Moretti et al., 2017; Villéger et al., 2017), their measurement is also a promising means of bypassing taxonomic impediment, i.e., the fact that most species are yet undescribed, and little is known of their interactions with other organisms and their environment.

Despite the importance of trait-based approaches, fully exploiting their potential relies heavily on the broad availability and compatibility of trait data to achieve sufficient taxonomic and regional coverage, both of present-day taxa as well as in evolutionary deep-time. However, the heterogeneity of data arising from different research contexts render trait data extremely heterogeneous and make the task of data compilation time-consuming and error-prone. To date, trait data have traditionally been harmonised and compiled into centralised databases only for specific organism groups and regional scope, often centred around particular research questions (e.g., PanTHERIA, Jones et al., 2009; TRY, Kattge, Díaz, et al., 2011; AmphiBio, Oliveira et al., 2017). Less well-studied taxa and specialised research questions lack the resources for such an endeavour. Besides initiatives aiming at assembling data, tools to enable the compatibility of data across databases are being developed. These include software to access trait data from the Internet (e.g., Chamberlain et al., 2017; Ankenbrand et al., 2018), semantic-web standards (Page, 2008; Wieczorek et al., 2012) and thesauri of consensus terms (Walls et al., 2012; Garnier et al., 2017).

Meanwhile, open and reproducible science has become mainstream: publication of research data without access restrictions, with structured metadata and in accordance to data standards to enable their reuse, has become the declared goal of an open biodiversity knowledge management (*http://www.bouchoutdeclaration.org/*) and is increasingly demanded by journals and public research funding agencies (Alliance of German Science Organisations, 2010; Royal Society Science Policy Centre, 2012). As a result, an increasing number of individual research projects publish their primary data on general-purpose file hosting services, where no data standards are enforced upon the uploaded material (Wilkinson et al., 2016). It is thus likely that trait data will become increasingly available, but a lack of data and metadata standardisation will hamper the efficient reuse and synthesis of published datasets.

In this paper, we review existing initiatives for trait-data collection and standardisation from the pragmatic view of data providers, data curators and data users, as well as data managers. We discuss current efforts to make trait data visible, accessible, interoperable and re-useable in downstream data analysis, as demanded by the FAIR guiding principles for scientific data (Wilkinson et al., 2016). Furthermore, we show how the current deficit in the standardisation of primary data hampers the implementation of interoperability and reuse of trait data. Based on these considerations, we propose a versatile vocabulary for describing ecological trait datasets, which builds upon, and is compatible with, existing terminology standards for biodiversity data, in particular the Darwin Core Standard for biodiversity data (DwC; Wieczorek et al., 2012). Since a standard vocabulary relies on the adoption by a broad research community, we discuss incentives for its use and lay out mechanisms for future consensus-building and community development towards an accessible and easy-to-use ecological trait-data standard vocabulary.

Initiatives for trait-data standardisation

The need for standardising trait data arises from the prospective gain of compiling heterogeneous trait datasets for data synthesis. Often, the scientific scope and focus differs between data providers measuring and assessing the trait data in the first place and data users who re-use published data for a broader synthesis application. Furthermore, data curators and data managers are taking up the task of providing compiled and harmonised data and prepare them for future use and long-term preservation. Data managers are concerned with the development of complex digital infrastructures for handling and analysing large amounts of data. These are idealised roles of researchers that are dealing with trait-data standardisation throughout the data life-cycle. In this chapter, we review four types of initiatives that are of relevance for trait-data standardisation (see Glossary in Table 1 for italicised terms):

1. Initiatives that provide trait *datasets* which have been assembled out of a particular research interest, either by measurement or collated from the literature.

2. Initiatives that aim to harmonise trait data from the literature or from direct measurements into data compilations or database infrastructures and make those data widely available on the Internet.

3. Initiatives that aim at the standardisation and development of consensus measurement methods and definitions for traits and provide standard *terminologies*.

4. Initiatives that aim to combine data (1 & 2) and terminologies (3) into formalized structures for knowledge representation to link trait data to a wider set of biodiversity data.

We consider these initiatives separately although they are often developed in conjunction to serve a particular database project, such as the TRY plant database (Kattge, Díaz, et al., 2011; Kattge, Ogle, et al., 2011) and the Thesaurus of Plant characteristics (TOP; Garnier et al., 2017). We show how the degree of trait-data standardisation in existing datasets is highly variable, and which tools and standards are currently applied to achieve harmonisation of data from multiple, distributed sources. The objective of this review is to raise awareness of the generic structure of trait data and aid researchers in how to share and publish their own datasets in an appropriate form.

Trait datasets

In the field of comparative biology, morphological traits, such as traits related to flower shape, leaf and stem structures for plants or wing and beak measurements for birds, as well as life-history traits such as Ellenberg values for plants or physiological and reproductive traits for animals (e.g., feeding biology, dispersal, metabolic rate and body size) have been assessed for decades and have been published in regular journal articles or books. With the rise of ecological trait-based research, measurements and information available from species descriptions have been compiled into project-specific datasets that typically comprise a local set of taxa and a focal set of traits. A plethora of such static datasets has been published alongside scientific articles, or as standalone data publications (see Kleyer et al., 2008 for a review on plant data; for animal data, e.g., Gossner et al., 2015 and Supplementary Material A, Table A1).

Today, the online publication of such data is greatly facilitated by file hosting services (e.g., Figshare, Zenodo, Researchgate, Data Dryad), which warrant long-term accessibility, and citeability via DOIs, and govern data sharing via license statements. These platforms offer the hosting of publicly accessible *file repositories* at low-cost or for free, which makes them attractive for small and intermediate-sized research projects that cannot dedicate extra resources for data management. Most importantly, these platforms enable public hosting of data with very low quality-thresholds regarding *metadata* documentation and data standardisation. Thus, although open for download, the trait datasets on such data repositories might be stored in variable tabular structures and labelled following self-defined terms which makes extraction and further use unnecessarily tedious.

For trait data, there are common issues arising from the variability of data structures and metadata quality. In terms of structure, trait data usually are reported in a species×traits wide-table format. In this intuitive data table, each row represents a species (or taxon) for which multiple traits are reported in columns. Similarly, when reporting raw data, researchers place observations on individual organisms in rows with multiple trait measurements applied to the same individual across multiple columns. Co-variates on the taxon, the individual specimen (e.g., sex or life-stage) or context of observation (e.g., time and place of sampling) would be placed in additional columns and would further expand the two-dimensional data table. The resolution or scope of these co-variates varies greatly depending on the research question and observation context. The column descriptions and terminology applied to taxa and traits are mostly project-specific and rarely chosen for compatibility with larger database initiatives. Variability in the number and meaning of columns in these data tables requires tedious manual adjustments when merging multiple datasets (Wickham, 2014). Furthermore, metadata provided along with the primary data vary in their level of detail, e.g., for documenting descriptions of variables, measurement procedures or sampling context (Kattge, Ogle, et al., 2011). While, in some datasets, information like geolocation or sampling date and time might be dataset-level information, thus qualifying as metadata, in other datasets they might be collected on a level of individual observations (see section on data compilations below). More importantly, clear statements on ownership and authorship, terms of use, or internationalisation (e.g., separators and delimiters), are often still neglected in primary trait-data publications. The task of harmonizing trait data is taken up by data-curating initiatives, who compile heterogeneous data into comprehensive databases (see next section).

Data compilation initiatives

In the past two decades, many distributed trait datasets have been aggregated and harmonised into greater collections with particular taxonomic or regional focus (e.g., Kleyer et al., 2008; Oliveira et al., 2017, see Supplementary Material A, Table A1). While these initiatives successfully address issues of heterogeneity in units or categorical variables, or achieve high taxonomic or geographic coverage, few of these compilations apply a standardised terminology for taxa or trait definitions. Additionally, in the process of data aggregation, rich metadata content might be lost, as the detail in the original files differs, while the reference to the original dataset becomes obscured, as only aggregated values are reported (e.g., means or medians). Such trait-data compilations are often labelled ‘database’, although they do not formally provide data in a database structure in the strict data-management sense. Instead, the data are released as static data tables of raw measurements or aggregate trait values on journal websites or open-access file hosting platforms, which may be updated irregularly.

As they deal with much larger amounts of data, initiatives that compile data from natural history museum collections are traditionally more concerned with standardisation. The amount of morphological measurements data extracted from museum collections and herbaria is likely to skyrocket in the near future due to digitisation efforts supported by new technology for scanning and pattern recognition (Smith & Blagoderov, 2012, and references therein; Ströbel et al., 2018) and citizen science initiatives (e.g., www.markmybird.org). For example, the VertNet database compiled and harmonised large quantities of vertebrate trait data from collections; the resulting data are published as versioned data tables which are updated as new data sources become available (*http://vertnet.org*, Guralnick et al., 2016).

Specialised online portals have been created to attract data submissions from a defined research field and take care of data harmonisation, thereby greatly facilitating data synthesis. For example, by aiming for a universal framework for plant traits, the TRY database (Kattge, Díaz, et al., 2011) attracted more data submissions and downloads than any other trait-data platform. The online portal enables selective data download and management of user permissions. For animal trait data, however, a single unified platform and harmonising scheme is still lacking. Nonetheless, initiatives for particular groups of animals do exist. Examples are the BETSI database on soil invertebrate traits (*http://betsi.cesab.org/*; Pey et al., 2014), the Carabids.org web portal (*http://www.carabids.org/*), the Coral Trait Database (Madin et al., 2016), or the Global Ants Database (Parr et al., 2017, see Supplementary Material A, Table A1). The role of online portals and database initiatives in standardising data and making them more accessible is paramount. Trait-data portals incentivise data submissions by offering increased data visibility and usage, while providing data-use policies that secure author attribution and, potentially, co-authorship of associated articles. However, maintaining centralised database infrastructures is costly and requires long-term funding (Bach et al., 2012).

Terminology standards for traits

A major challenge in trait-data standardisation is the lack of widely accepted and unambiguous trait definitions (Kissling et al., 2018). Previous standard definitions of trait *concepts* range from listings of selected definitions in vocabularies, over well-defined method handbooks and comprehensive *thesauri*, to formalized definitions of trait concepts in *ontologies*. The initiatives behind method handbooks, thesauri and ontologies are essential for building community consensus for trait definitions.

Very general classes of traits are defined within the list of GeoBON Essential Biodiversity Variables (Kissling et al., 2018) aiming for a list of functional indicators for ecosystem health.

Assigning a detailed and unambiguous methodological protocol for a trait, including the units to use or the ordinal or factor levels to be assigned, is essential for standardising its measurement process. Efforts to develop handbooks for measurement protocols provide such a methodological standardisation for plants (Cornelissen et al., 2003; Perez-Harguindeguy et al., 2013) or invertebrates (Moretti et al., 2017), but are of limited use in harmonising trait data that pre-date or ignore this standard (Kattge, Ogle, et al., 2011).

A *thesaurus* provides a "controlled vocabulary designed to clarify the definition and structuring of key terms and associated concepts in a specific discipline" (Laporte et al., 2013; Garnier et al., 2017). To provide a logic structure for trait terms, Garnier et al. (2017) suggest the Entity-Quality model (EQ), where a trait is defined as 'an entity having a quality' (for instance for trait 'femur length', 'femur' is the entity and 'length' the quality). In thesauri, hierarchies of concepts can be formalized by linking each term to broader or narrower terms, or to synonyms. For example, the definition of 'femur length of first leg, left side' is narrower than 'femur length' which is narrower than 'leg trait' which is narrower than 'locomotion trait'. Being publicly available, it is also possible to refer to these defined terms via globally unique *Uniform Resource Identifiers (URIs)*. For example, a measurement of fruit mass could be linked to the definition of the term within the Thesaurus of Plant characteristics (TOP, Garnier et al., 2017) via its URI '*http://top-thesaurus.org/annotationInfo?viz=1&&trait=Fruit_mass*'.

In addition to defining terms for human interpretation, *ontologies* define terms by their relationship to other defined terms, thereby providing a semantic model of the concepts used within a domain of research, with the objective of enabling the computational interpretation of data (Walls et al., 2012, 2014; Kissling et al., 2018). The Plant Trait Ontology (TO) definition of the concept 'seed size' contains references to other globally defined terms: "A seed morphology trait (TO:0000184) which is the size of a seed (PO:0009010)". Thus, trait definitions may refer to related terms or synonyms defined in other trait ontologies or other scientific ontologies, like units as defined by the Units of Measurement Ontology (Gkoutos et al., 2012). By providing ontologies in a formalized syntax, like Web Ontology Language (OWL), a machine-readable web of definitions is spun across the Internet allowing researchers and search engines to relate independent trait measurements with each other and connect them to the wider *semantic web* of online data (Gruber, 1995; Berners-Lee et al., 2001; Page, 2008; Walls et al., 2012).

Comprehensive trait thesauri have been developed in TOP (which is employed in the TRY database, Garnier et al., 2017) and in the Thesaurus for Soil Invertebrate Trait-based Approaches (T-SITA, *http://t-sita.cesab.org/*, Pey et al., 2014). Ontologies of trait definitions have been developed for plants (e.g., the Plant Ontology, Jaiswal et al., 2005; Walls et al., 2012; the Flora Phenotype Ontology, Hoehndorf et al., 2016), and for specific animal taxa (e.g., the Hymenoptera Anatomy Ontology, Yoder et al., 2010; the Vertebrate Trait Ontology, Park et al., 2013). The UBERON ontology is an integrated cross-species anatomy ontology for all animals, which combines concepts from different existing ontologies, with wide application in biomedical or physiological research (Mungall et al., 2012).

To conclude, there is already a suite of globally available thesauri and ontologies for traits. However, definitions in some domains are better covered than others (Kissling et al., 2018), and different curation strategies and measures for peer-review and community building are employed. To this end, the OBO Foundry is providing a development platform for (biological) ontologies and offers review and quality control (Smith et al., 2007, *http://www.obofoundry.org/*). While defined vocabularies are increasingly used in biodiversity data management, distributed trait data of smaller projects published in general-purpose file servers rarely refer to standard terminologies. Finding and applying the most suited and highest quality ontology from the range of available ontologies is not an easy task for ecological researchers. To mitigate this effort, meta-ontology initiatives, like Ontobee (*http://www.ontobee.org/*), Bioportal (*https://bioportal.bioontology.org/*, Whetzel et al. (2011)), or the GFBio Terminology Service (Karam et al., 2016, *https://terminologies.gfbio.org/*), provide centralised hosting for trait ontologies, structured browsing, and harmonized web services for computational access.

Trait-data structures

While trait thesauri and trait ontologies typically define concepts of measurements and observations for focal groups of organisms, they do not specify the format or structure in which trait data should be stored and labelled.

A trait dataset typically contains multiple data entries, where each entry describes a trait value observed on an instance of a scientific taxon. The item on which the value has been observed can be very variable, ranging from an *occurrence* of an individual at a specific place and time in its natural environment or a preserved specimen in a collection (Fig. 1*a*), a group of individuals of a specific taxon (Fig. 1*b*), or an entire population of a species (Fig. 1*c*-*d*). The reported trait values may be quantitative measurements or qualitative facts. Quantitative measurements are values obtained either by direct morphological, physiological or behavioural observations on single specimens (Fig. 1*a*), by aggregating replicated measurements on multiple entities (Fig. 1*b*) or by estimating the means or ranges for the respective taxon as reported in the literature or other published sources (e.g., databases, Fig. 1*c*). This encompasses a wide range of numeric data types, including continuous, binary, integer, intervals or ratios, as well as categorical (ordinal or nominal) values. Qualitative facts are assignments of categorical information, often on entire taxa, e.g., of a behavioural or life-history trait (Fig. 1*d*).

Beyond these core observations, further information might be available that specify the taxon concept applied, provide detail on the measurement method, or that place the reported measurement in a broader observation context (including geolocation as well as date and time of sampling). As such data may be useful for future analysis of the causal reasons of trait variation or to explain noise in measurement data, it should always be published along with the core data. In most cases, information on place and time apply to the entire dataset, and thus would be included in the metadata accompanying a data publication (potentially applying Ecological Metadata Language, EML, KNB, 2011 as a formal structure). In the case of trait data and depending on the research scope, the information may also have been collected on a level of measurement, occurrence or taxon level. Geolocation or date and time would then not be provided as metadata, but as covariate data in additional columns of the primary dataset. When compiling datasets, it is a key task of data curators to deal with dataset-level information and maintain it for downstream analysis by incorporating it into the compiled data table.

Standard terms for the formal description of the common concepts of biodiversity knowledge have been provided in the schema for biological collection records (Access to Biological Collection Data, ABCD; Holetschek et al., 2012) or the *Darwin Core Standard* for biodiversity data (DwC; Wieczorek et al., 2012). Both DwC and ABCD are ratified standards of the Biodiversity Information Standards (TDWG, www.tdwg.org) which is a global network to support the development and wide adoption of exchange standards for biodiversity data. These terms may be used for defining columns in data tables that contain measurement values, units and categorical levels, taxon names, variables such as sex or life stage, information of time and date of observation, and methodological details (Robertson et al., 2009). A suite of terminology extensions links to and expands the capacities of DwC (Wieczorek et al., 2012). Of particular importance for trait data is the 'MeasurementOrFact' extension, which typically would be used in database management and bioinformatics to structure trait observations (Parr et al., 2016).

While the abovementioned standards provide terms and concept definitions, and the logic relationships of those, they do not prescribe explicit structure for trait data. Based on the terms of DwC, the Extensible Observation Ontology (OBOE, Madin et al., 2007; Schildhauer et al., 2016) formalizes observations and measurements into a machine-readable ontology, thus being easily integrated into larger database management systems. By applying this scheme for plant traits, Kattge, Ogle, et al. (2011) propose a generic database structure that covers most potential use cases of trait-based ecology. This data structure is built around a central data table that contains observations of individual plants linked to several measurements of traits via identifiers. The observations are also linked to a taxonomy and metadata descriptors of the observation context, like location or experimental treatment. Kissling et al. (2018) discuss different ontologies (including OBOE) that formalize the structure of observation data and attest that for the use cases of trait data these ontologies are still difficult to integrate.

The Encyclopedia of Life (EOL) has proposed TraitBank (Parr et al., 2016) as a standard structure for uploading data on physiological and life-history traits of all kingdoms of life. It is to date the most general approach of an integrated structure for trait data. The framework employs established terms provided by the DwC and the DwC MeasurementOrFact extension (Parr et al., 2016). Additional layers of information cover bibliographic references, multimedia archives and ecological interactions. TraitBank invites data submissions to the EOL database in a structured Darwin Core Archive (DwC-A, GBIF, 2017), which is a set of simple text files (csv), a file to specify relationships between these text files (called meta.xml), and a file for metadata descriptions using EML (called, EML.xml, see GBIF, 2017 for specifications, archives can be validated before upload on *https://tools.gbif.org/dwca-validator/*).

All of these structures suggest the use of stable URIs to refer to taxon concepts. The difficulties with keeping taxonomic references intact along with continuous changes in taxonomy consensus are a central challenge of biodiversity data management and are beyond the scope of this review (Franz et al., 2016). Initiatives that aim at providing a stable reference while tracking the changing taxon concepts are for instance the Catalogue of Life (*https://www.catalogueoflife.org/*) or the EDIT Platform for Cybertaxonomy (*https://cybertaxonomy.eu/*). The GBIF Backbone Taxonomy (GBIF Secretariat, 2017) collects and bundles existing terminologies into a single reference framework.

Closing gaps to improve trait-data re-use

In sum, we attest to a gap between the trait-data structures developed for data curators and data managers and the data input produced by data providers. Hardly any of the aforementioned standalone or aggregated trait datasets for birds, amphibians, mammals or invertebrates employs the described standard terminologies, ontologies or data standards. As it stands, re-using these data in larger compilations or integrating them into structured database initiatives is error-prone and labour-intensive and the potential for a broad synthesis is diminished.

One likely reason for this lack of standardisation is the complexity of the task: the proposed data structures are designed for multi-layered, relational databases rather than for standalone datasets for which a two-dimensional data table may suffice. In the eyes of the data-provider, in most cases, any co-variates can be appended as extra columns to the dataset. The other reason is lack of awareness of the need for trait-data standardisation among data providers, who are not trained in the demands of biodiversity data-management. In addition, complying with what may be non-intuitive data structures is an investment without clear incentive or immediate pay-off, and hardly affordable for small and intermediate-size research projects, especially since funders often do not require these efforts to be included into proposals.

By filling this gap, data-brokering services (the German Federation for Biological Data; gfbio.org, Diepenbroek et al., 2014; e.g., Data Observation Network for Earth, DataONE, Michener et al., 2011) or data management systems for scientific projects (e.g., KNB and its open-source database back-end Metacat, *https://knb.ecoinformatics.org/*; Diversity Workbench, *http://diversityworkbench.net*; BEXIS2, *http://bexis2.uni-jena.de/*) are likely to gain importance. These services simplify and direct the standardised upload of research data and descriptive metadata into reliable and interlinked data infrastructures. The goal of such initiatives is to facilitate data re-use by providing standardisation of data, for instance by mapping to unambiguous terminologies and ontologies for biodiversity data and clarifying conditions of data re-use.

Another solution for data-users to access trait data in a structured way is offered by decentralised tools and toolchains to facilitate the use and analysis of trait data. For instance, the R-package ‘traits’ (Chamberlain et al., 2017) contains functions to extract trait data directly from their source, including Birdlife, EOL TraitBank or BetyDB. The package ‘TR8’ provides similar access to plant traits from a list of databases (including LEDA, BiolFlor and Ellenberg values; Bocci, 2015) and aggregates them into a species×traits wide-table. FENNEC (Ankenbrand et al., 2018) is an online tool or self-hosted service capable of extracting trait information from multiple sources for a target species community.

A more widespread implementation of ontologies would advance the possibilities to integrate datasets and reduce noise and uncertainty when aggregating data. First, groups of trait researchers must take up the task of developing consensus definitions into semantically defined ontologies that are useful for their use case. Platforms like OBO Foundry can help structuring this process. Second, the reference to ontologies and thesauri must be incentivised and facilitated for individual data providers by the development of tools for matching concepts from the available ontologies to their data. Third, frameworks for providing trait data in an unambiguous and machine-readable structure must be simplified to match the limited resources of small and intermediate research projects. This can be achieved by extending documentation or providing tools for the application of existing ontology frameworks and database structures (e.g. data validator services), and by defining easy-to-use standard vocabularies that enable the interoperability of data at minimal effort.

However, no unified and widely adopted terminology for primary trait-data publications has emerged across the multiple sub-disciplines of trait-based research. In the following chapter, we propose a unified vocabulary for trait data that can serve as a minimal consensus for describing and labelling trait data. The simplicity of this standard terminology will lower the thresholds and offer high pay-off in the visibility and reuse of published data. By establishing this as a "best-practice" in trait-based research, trait data will eventually fulfil the FAIR guiding principles for scientific data (Wilkinson et al., 2016).

Introducing the Ecological Trait-data Standard Vocabulary

As a response to the challenges outlined above, we propose a versatile standard vocabulary for trait-based ecological research. The Ecological Trait-data Standard Vocabulary (ETS) is accessible at *https://terminologies.gfbio.org/terms/ets/pages/* and combines terms of DwC with newly defined terms to cover the variety of trait-based approaches and their different needs to report measurement detail. Rather than prescribing a data structure or exchange format, the vocabulary is intended as a more inclusive terminology that can be used in three major use cases:

1. by data providers: for publication of standardised primary data on open-access data repositories, or for labelling project-specific data for local use and exchange with collaborators, e.g., in two-dimensional data tables or project databases,
2. by data users and data curators: as a consensus vocabulary when compiling data from distributed sources into aggregate datasets, e.g., to map standardised columns and refer to taxa and trait definitions in a uniform way, and
3. by data managers: in developing data exchange formats between online resources, web services and software tools, e.g., when providing database queries via a web service or defining input and output formats of software packages.

All terms may be applied to describe columns of a data table (Fig. 2; see Supplementary Material B for best-practice principles and examples for publishing primary data). By applying these standard terms, data providers can ensure that the description of trait measurements uploaded into public data repositories will be unambiguous. It will facilitate interoperability of published data and enable their re-use for future data aggregation initiatives and data synthesis, while warranting long-term accessibility.

The definitions of terms are hosted on the GFBio Terminology Service (Karam et al., 2016, *https://terminologies.gfbio.org/*), providing permanent and redirectable individual URIs and URLs for each term. The service can be accessed programmatically (i.e., via the API; *https://terminologies.gfbio.org/api/terminologies/*).

Our vocabulary offers three extensions to contain additional information on the context of the observation along with the core data in analogy to DwC extensions (“Taxon”, “Measurement or Fact”, and “Occurrence”; see section on extensions below). Further terms are provided for dealing with typical dataset-level information on authorship and rights of re-use of the data (based on terms of Dublin Core Metadata Initiative, DCMI), as well as for defining own trait concepts (see section on metadata below). Aspects not covered by the vocabulary may draw from terms provided by other existing terminologies (in particular DCMI and DwC and its extensions), or be added as user-defined columns (which should then be clearly specified in the metadata-information accompanying the dataset).

Building community consensus

In designing this vocabulary, we drew on the combined expertise of empirical biodiversity researchers (data providers), biodiversity synthesis researchers (data users), and biodiversity informatics researchers (data managers). The aim was to develop a simple, easy-to-use template for standalone trait-data publications or data compilations, to facilitate their re-use for synthesis and integration into larger database structures. Earlier proposals for trait-data standards (e.g., Kattge, Ogle, et al., 2011; Parr et al., 2016) have been designed for relational database structures from a data manager perspective, which may be the reason why they have so far hardly been adopted for primary data publications. We paid particular attention to these existing data standards (e.g., Madin et al., 2007; Kattge, Díaz, et al., 2011; Kattge, Ogle, et al., 2011; Parr et al., 2016; Garnier et al., 2017) to maximize compatibility.

Nonetheless, we are aware of the diverse use-cases of trait data that might not yet be covered by the current version of the vocabulary. The version presented here is a mere starting point of a community effort towards a consolidated and comprehensive Ecological Trait-data Standard Vocabulary, as a key resource for trait-data standardisation in ecological research. For future development of the vocabulary, we will engage with a broader community of trait researchers, in particular via the Open Traits Network (*http://opentraits.org*), and work towards full compatibility with other initiatives of biodiversity data standardisation by collaborating with Biodiversity Information Standards TDWG (Taxonomic Databases Working Group, *http://www.tdwg.org*). This will also link our initiative to other trait-based research fields, like biomedical and agricultural research. We invite communities of all trait-based research fields to discuss, revise and submit terms and extensions of the vocabulary (coordinated via Github Issues at *https://github.com/EcologicalTraitData/ETS/issues*). The standard vocabulary will be released in subsequent versions and published as a stable reference on the GFBio Terminology Service.

Specification of core terms

To qualify as trait data complying with the ETS, the following content is required at minimum (Fig. 2 *b*):

1. a value (column `traitValue`) and – for numeric values – a standard unit

(*traitUnit*);

2. a descriptive trait name (*traitName*) that links the observation to a standardised definition (i.e., a concept);
3. the scientific taxon name (*scientificName*) for which the measurement or fact was obtained that links the observation to an accepted taxon concept.

The *traitName* and *scientificName* would use unambiguous terms assigning both to clearly defined concepts. Eventually, disambiguation can be warranted by adding globally valid Uniform Resource Identifiers (URIs) for taxon (*taxonID*) and trait definitions (*traitID*). For example, referring to GBIF Backbone Terminology, for *Bellis perennis,* the *taxonID* would be '*https://www.gbif.org/species/3117424*'; the *traitID* for 'fruit mass' according to Flora Phenotype Ontology would be '*http://purl.obolibrary.org/obo/FLOPO_0005265*'. Wherever possible, the field *traitID* should point to an unambiguous trait definition in a published ontology. If no suitable reference exists, trait data should always be accompanied by a dataset-specific listing of trait concepts. Such a controlled vocabulary would, in its simplest form, assign trait names with an unambiguous definition of the trait and an expected format of measured values or reported facts (e.g., units or legit factor levels). Ideally, this definition refers to or refines terms from published trait ontologies. By providing a minimal vocabulary for trait lists within the ETS, we hope to facilitate the unambiguous definition of traits for trait datasets. This vocabulary might also prove useful for the future publication of trait ontologies.

To ensure compatibility with project-specific databases or analytical code, it might be in the interest of the data author to keep user-specific identifiers for those terms, for which we are suggesting the use of *verbatimScientificName* and *verbatimTraitName* (Fig. 2 *c*). By allowing user-side entries along with consensus terms we acknowledge the fact that most authors have their own schemes for standardisation which may refer to different scientific community standards (as also practiced in TRY, Kattge, Ogle, et al., 2011; Kattge, Díaz, et al., 2011). The redundancy of labelling allows for continuity for data providers while also enabling quality checks and comparability for data curators.

Similarly, standardisation of units can be achieved by relying on SI base units or by relating units to unambiguous concepts via URIs provided by ontologies (Madin et al., 2007; Gkoutos et al., 2012; Keil & Schindler, 2018). For categorical or binary traits, the categories should conform to expected levels as defined in the trait concept or be unambiguously defined in the metadata of the dataset. The vocabulary offers terms for keeping the user-defined values in dataset-specific units and factor levels along with standardised entries (*verbatimTraitValue* and *verbatimTraitUnit*, Fig. 2 *c*).

Extensions for additional data layers

Beyond measurement units or higher taxon information, further information might complement the core data which are related to the individual specimen, the reported fact, measurement or sampling event. We propose three extensions of the vocabulary that should be used to describe this information (Fig. 2*d*), in line with the existing DwC extension structure:

1. The *Taxon* extension provides further terms for specifying the taxonomic resolution of the observation and to ensure the correct reference in case of synonyms and homonyms.

2. The *MeasurementOrFact* extension provides terms to describe information at the level of single measurements or reported facts, such as the original literature reference for the reported value, the method of measurement or statistical method of aggregation. It provides important information that allows for the tracking of potential sources of noise or bias in measured data (e.g., variation in measurement method) or aggregated values (e.g., statistical method), as well as the source of reported facts (e.g., literature source or expert reference).

3. The *Occurrence* extension contains vocabulary to describe information on the observation context of individual organisms, such as sex, life stage or age. This also includes the method of sampling and preservation, as well as the date and geographical location, which provide an important resource to analyse trait variation due to differences in space and time.

These additional layers of information can either be added as extra columns to the core dataset or kept in separate data sheets, thus avoiding redundancy and duplication of content. A unique identifier links to these other datasheets, encoding single measurements or reported facts (*measurementID*) or individual organisms of a species (*occurrenceID*).

The concept of ‘occurrence’ is prone to cause confusion. By definition of DwC it is “An existence of an Organism at a particular place at a particular time”. Thus, any individual observed twice would have two distinct ‘occurrences’. If sampling of an individual is only performed once, this results in any occurrence being semantically identical with the individual organism (i.e., the DwC term ‘organism’). Some data types directly refer to existing global identifiers for occurrence IDs, e.g., a GBIF URI or a stable identifier references the precise specimen at a particular place and time from which the measurement was taken (Groom et al., 2017; Güntsch et al., 2017). Also, as ‘occurrence’ is strictly defined by a date-time event, it may be identical to the common-sense concept of ‘observation’. As such, data entries for location of sampling (provided in column *LocationID*) and sampling campaigns (*eventID*), which are often recorded and published along with trait data, are tightly linked to the concept of ‘occurrence’. As occurrence is the narrower term and the key concept for linking an individual organism to a location and sampling event in DwC, and since it is indeed relevant to distinguish between multiple ‘occurrences’ of the same organism in some trait-based research applications, the ETS sticks to this terminology.

Identifiers can also be used to provide a structure within the measurement data table, e.g., to link rows of measurements on the same individual (by having entries share the same ID in column *occurrenceID*). Similarly, the values of multivariate measurements can be linked by using the same *measurementID* for several rows.

The terms of the extensions draw from terms of the DwC extensions of particular relevance for trait data. See the documentation of the ETS for further detail on the use of extensions.

Specification of metadata

Dataset-level information about structure, provenance of data, authorship and data ownership, as well as terms of use should be considered when sharing and working with trait datasets (Michener, 2006; Kissling et al., 2018). In the case of primary measurement data, this information usually applies to the entire trait dataset, and would be stored along with the published data as metadata entered in a template provided by the file hosting service. To facilitate interoperability and computational evaluation of metadata, specific standards for metadata may be provided, e.g., by applying Ecological Metadata Language (EML, KNB, 2011). Whenever data from different sources are compiled into a single dataset, metadata information would become part of the resulting data table, as each data entry would have to maintain reference to the original data provider and conditions of re-use of these data. This can be achieved by appending the metadata terms as columns to the core dataset, or by linking to a secondary data table via an unambiguous *datasetID* (e.g., a URI pointing to the source DOI) and a descriptive *datasetName* (e.g., a descriptive name for the source). The ETS metadata vocabulary provides terms for a minimal set of information that should be provided along with trait data. The suggested terms originate from Dublin Core Metadata Initiative (DCMI), and are widely compatible with terms provided by the DataCite Metadata Schema (DataCite Metadata Working Group, 2019). The terms can be extended and complemented by using terms from these resources.

In order to ensure traceability, the metadata of any dataset that employs the ETS should refer to the specific online version that was used to build the dataset , e.g., by entering "Schneider, F.D., Jochum, M., Le Provost, G., Penone, C., Ostrowski, A. and Simons, N.K., 2019, Ecological Traitdata Standard Vocabulary v0.10, DOI: 10.5281/zenodo.2605377, URL: *https://terminologies.gfbio.org/terms/ets/pages/*" in the metadata field *conformsTo*. Wherever referring to individual terms of the vocabulary in publications or metadata, this should be done via their individual URIs.

Discussion

To serve the demand for the standardisation and harmonisation of ecological trait data which has arisen from a growing number of distributed datasets of different research contexts, we propose a versatile vocabulary for the publication of new datasets, for the creation of data compilations, and for the exchange and handling of trait data in the context of the semantic web.

Consensus building on how traits are to be used and evaluated is currently under way in several fields of ecological research with their taxonomic focus and project-specific questions (Pey et al., 2014; Garnier et al., 2017; Moretti et al., 2017; Kissling et al., 2018). Such community discussions on trait definitions and measurement practices are leading to a better quality of data, naturally. However, they still require a stronger linkage into the global biodiversity data initiatives. With our proposal of an Ecological Trait-data Standard Vocabulary (ETS), we aim to capture the common core concept of trait data in a single resource terminology and provide a starting point for the development of a joint language and terminology around traits as a cross-sectoral topic of ecological and evolutionary research. To enable the ETS to capture the different approaches in trait-based research across fields, we invite researchers to contribute to future versions of the standard vocabulary and develop their own applications and ontologies that interact with it. Development will also aim at linking the initiative to the joint efforts for biodiversity data terminologies, in particular within Biodiversity Information Standards (TDWG).

Data released according to consensus standards, especially if published under open-access licenses, are more easily re-used in compilations and synthesis studies. By providing the ETS, an easy-to-use vocabulary for trait-based research, the investment of time and resources in trait-data standardisation before publication will be mitigated for individual researchers and small research projects. A well-defined minimal vocabulary for metadata will also ensure that authorship and terms of use are appropriately documented along the data life cycle. However, for these incentives to take effect, data publications and data citations must become viewed as a valid scientific contribution to the community and recognized in the professional evaluation of individual researchers (Costello, 2009; Roche et al., 2015).

At the community level, shifting the task of standardisation from the data-user side to the data-owner side yields great gain in accuracy and reduces the risk of misinterpretation. For instance, measurement results depend very much on the precise methodology used and often systematic biases could be corrected for when providing an unambiguous definition. On the other hand, plausibility checks and evaluation of statistical methods, e.g., for aggregating trait values to the species level, can only be done in comparison across a wide array of datasets. Currently, these "big data" volumes are only available in centralised databases. However, to establish a best practice of data aggregation, an exploration and evaluation of different methods for quality assessment and quality control should be subject to a community discussion. This is only possible with large quantities of distributed data being available in a harmonised way. The ETS facilitates such a community-driven comparison.

Without clearly defined terms and concepts, handling of large amounts of trait data by computational assistance systems for scientific analysis ('e-Science') will be massively hampered (Wilkinson et al., 2016). The ETS represents an important building block for a unified mode to ease data exchange between web services and software packages and thus facilitates the development of a software toolchain for the trait-data lifecycle. Having well-defined terms is also a key precondition for developing exchange formats between large database initiatives and biodiversity data archives. Even further downstream, readying the primary data for the semantic web via references to ontologies and data standards will ease the application of automatized big-data mining and machine-learning techniques.

Conclusion

To date, there is a rich distributed body of independently published trait datasets, each with a specific focus on particular organism groups, ecosystem types or regions. These distributed data are heterogeneous in form and description, hampering endeavours to harmonise, compile and analyse these data.

Using a standard vocabulary with globally accessible definitions of terms would allow distributed trait data to be more easily re-used and harmonised into aggregated datasets. The biggest challenge in future standardisation of trait data may be consensus building for standard terms, the establishment of incentives and the development of tools for a user-side standardisation before the publication of data. This requires significant effort, but it returns great scientific benefit by enabling data-heavy synthesis for a general understanding of biodiversity and ecosystem functioning.

Acknowledgements

Thanks to all respondents to an internal online survey on trait data for the Biodiversity Exploratories project and to Diana Bowler, Klaus Birkhofer, Runa Boeddinghaus, Markus Fischer, Jens Kattge (and the TRY Steering Commitee), Felicitas Löffler, Catrin Westphal and two anonymous reviewers for comments on the manuscript drafts and pre-print, as well as the Ecological Trait-data Standard vocabulary. We are grateful to the organisers and participants of the Open Traits workshop in New Orleans, USA, in August 2018.

We thank the past and present scientific coordinators, local managers and data managers of the Biodiversity Exploratories program for their work, and Markus Fischer, Eduard Linsenmair, Dominik Hessenmöller, Daniel Prati, Ingo Schöning, François Buscot, Ernst-Detlef Schulze, Wolfgang W. Weisser and the late Elisabeth Kalko for their role in setting up the Biodiversity Exploratories program. The work has been partly funded by the DFG Priority Program 1374 "Infrastructure-Biodiversity-Exploratories" (DFG-Refno. Po362/18-3, MA7144/1-1, WE3081/21-1, KO2209/12-2); MMG obtained funding from Swiss National Science Foundation (SNF 310030E-173542/1); MJ was supported by the German Research Foundation within the framework of the Jena Experiment (FOR 1451) and by the Swiss National Science Foundation.

Authors' contributions

FDS, AO, CP, and NKS conceived the idea and developed the vocabulary for the trait-data standard with significant contributions of MJ and GLP; CP and FDS curated the living spreadsheet; AG and DF implemented the vocabulary in the GFBio terminology service; all authors contributed critically to the structure and content of the manuscript and gave final approval for publication.

Data Accessibility

The online reference for the Ecological Trait-data Standard Vocabulary described in this paper is *https://terminologies.gfbio.org/terms/ets/pages/*, with a stable DOI representing all versions: *https://doi.org/10.5281/zenodo.1041732*. Any future development of the vocabulary is coordinated via *https://github.com/EcologicalTraitData/ETS/*.

Supplementary Material A - an exemplary list of published trait datasets and data compilations, their regional and taxonomic focus, the number and scope of traits covered, their location on the Internet and the terms of use.

Supplementary Material B - contains best-practice guidelines and three worked examples of how to 1.) apply the core ETS vocabulary to species×traits wide-trables, 2.) convert data to long-table format using identifiers 3.) apply ETS extensions and identifiers to include co-variate information on different layers of the dataset.

References

Allan, E., Manning, P., Alt, F., Binkenstein, J., Blaser, S., Blüthgen, N., … Fischer, M. (2015). Land use intensification alters ecosystem multifunctionality via loss of biodiversity and changes to functional composition. *Ecology Letters*, *18*(8), 834–843. doi:*10.1111/ele.12469*

Alliance of German Science Organisations. (2010). Principles for the Handling of Research Data. Retrieved 9 November 2017, from *https://www.wissenschaftsrat.de/download/archiv/Allianz-Principles_Research_Data_2010.pdf*

Ankenbrand, M. J., Hohlfeld, S. C. Y., Weber, L., Foerster, F., & Keller, A. (2018). FENNEC - Functional Exploration of Natural Networks and Ecological Communities. *bioRxiv*, 194308. doi:*10.1101/194308*

Bach, K., Schäfer, D., Enke, N., Seeger, B., Gemeinholzer, B., & Bendix, J. (2012). A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. *Ecological Informatics*, *11*, 16–24.

Bello, F. de, Lavorel, S., Díaz, S., Harrington, R., Cornelissen, J. H. C., Bardgett, R. D., … Harrison, P. A. (2010). Towards an assessment of multiple ecosystem processes and services via functional traits. *Biodiversity and Conservation*, *19*(10), 2873–2893. doi:*10.1007/s10531-010-9850-9*

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 28–37.

Bocci, G. (2015). TR8: An R package for easily retrieving plant species traits. *Methods in Ecology and Evolution*, *6*(3), 347–350. doi:*10.1111/2041-210X.12327*

Chamberlain, S., Foster, Z., Bartomeus, I., LeBauer, D., & Harris, D. (2017). Traits: Species Trait Data from Around the Web (Version 0.3.0). Retrieved from *https://cran.r-project.org/web/packages/traits/index.html*

Cornelissen, J. H. C., Lavorel, S., Garnier, E., Diaz, S., Buchmann, N., Gurvich, D. E., … Van Der Heijden, M. G. A. (2003). A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. *Australian Journal of Botany*, *51*(4), 335–380.

Costello, M. J. (2009). Motivating Online Publication of Data. *BioScience*, *59*(5), 418–427. doi:*10.1525/bio.2009.59.5.9*

DataCite Metadata Working Group. (2019, March 20). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.2. DataCite e.V. doi:*10.5438/bmjt-bx77*

Diepenbroek, M., Glöckner, F. O., Grobe, P., Güntsch, A., Huber, R., König-Ries, B., … Tolksdorf, R. (2014). Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio). In *GI-Jahrestagung* (pp. 1711–1721).

Díaz, S., Kattge, J., Cornelissen, J. H. C., Wright, I. J., Lavorel, S., Dray, S., … Gorné, L. D. (2016). The global spectrum of plant form and function. *Nature*, *529*(7585), 167–171. doi:*10.1038/nature16489*

Díaz, S., Quétier, F., Cáceres, D. M., Trainor, S. F., Pérez-Harguindeguy, N., Bret-Harte, M. S., … Poorter, L. (2011). Linking functional diversity and social actor strategies in a framework for interdisciplinary analysis of nature's benefits to society. *Proceedings of the National Academy of Sciences*, *108*(3), 895–902.

Franz, N. M., Chen, M., Kianmajd, P., Yu, S., Bowers, S., Weakley, A. S., & Ludäscher, B. (2016). Names are not good enough: Reasoning over taxonomic change in the Andropogon complex1. *Semantic Web*, *7*(6), 645–667. doi:*10.3233/SW-160220*

Garnier, E., Stahl, U., Laporte, M.-A., Kattge, J., Mougenot, I., Kühn, I., … Klotz, S. (2017). Towards a thesaurus of plant characteristics: An ecological contribution. *Journal of Ecology*, *105*(2), 298–309. doi:*10.1111/1365-2745.12698*

GBIF. (2017, May 9). Darwin Core Archives - How-to Guide. Retrieved 8 November 2017, from *http://eol.org/info/structured_data_archives*

GBIF Secretariat. (2017). GBIF Backbone Taxonomy. doi:*10.15468/39omei*

Gkoutos, G. V., Schofield, P. N., & Hoehndorf, R. (2012). The Units Ontology: A tool for integrating units of measurement in science. *Database*, *2012*. doi:*10.1093/database/bas033*

Gossner, M. M., Simons, N. K., Achtziger, R., Blick, T., Dorow, W. H. O., Dziock, F., … Weisser, W. W. (2015). A summary of eight traits of Coleoptera, Hemiptera, Orthoptera and Araneae, occurring in grasslands in Germany. *Scientific Data*, *2*, 150013. doi:*10.1038/sdata.2015.13*

Grime, J. P. (2001). *Plant Strategies, Vegetation Processes, and Ecosystem Properties*. John Wiley & Sons.

Groom, Q., Hyam, R., & Güntsch, A. (2017). Data management: Stable identifiers for collection specimens. *Nature*, *546*(7656), 33.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5), 907–928. doi:*10.1006/ijhc.1995.1081*

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*(0). doi:*10.1093/database/baw158*

Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Röpert, D., Casino, A., … Triebel, D. (2017). Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database*, *2017*. doi:*10.1093/database/bax003*

Hoehndorf, R., Alshahrani, M., Gkoutos, G. V., Gosline, G., Groom, Q., Hamann, T., … Weiland, C. (2016). The flora phenotype ontology (FLOPO): Tool for integrating morphological traits and phenotypes of vascular plants. *Journal of Biomedical Semantics*, *7,* 65. doi:*10.1186/s13326-016-0107-8*

Holetschek, J., Dröge, G., Güntsch, A., & Berendsohn, W. G. (2012). The ABCD of primary biodiversity data access. *Plant Biosystems-an International Journal Dealing with All Aspects of Plant Biology*, *146*(4), 771–779. doi:*10.1080/11263504.2012.740085*

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., … Zapata, F. (2005). Plant Ontology (PO): A Controlled Vocabulary of Plant Structures and Growth Stages [Research article]. doi:*10.1002/cfg.496*

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., … Purvis, A. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, *90*(9), 2648–2648. doi:*10.1890/08-1494.1*

Karam, N., Müller-Birn, C., Gleisberg, M., Fichtmüller, D., Tolksdorf, R., & Güntsch, A. (2016). A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data. *Datenbank-Spektrum*, *16*(3), 195–205.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., … Wirth, C. (2011). TRY – a global database of plant traits. *Global Change Biology*, *17*(9), 2905–2935. doi:*10.1111/j.1365-2486.2011.02451.x*

Kattge, J., Ogle, K., Bönisch, G., Díaz, S., Lavorel, S., Madin, J., … Wirth, C. (2011). A generic structure for plant trait databases. *Methods in Ecology and Evolution*, *2*(2), 202–213. doi:*10.1111/j.2041-210X.2010.00067.x*

Keil, J. M., & Schindler, S. (2018). Comparison and evaluation of ontologies for units of measurement. *Semantic Web*, (Preprint), 1–19.

Kissling, W. D., Walls, R., Bowser, A., Jones, M. O., Kattge, J., Agosti, D., … Guralnick, R. P. (2018). Towards global data products of Essential Biodiversity Variables on species traits. *Nature Ecology & Evolution*, *2*(10), 1531–1540. doi:*10.1038/s41559-018-0667-3*

Kleyer, M., Bekker, R., Knevel, I., Bakker, J., Thompson, K., Sonnenschein, M., … Peco, B. (2008). The LEDA Traitbase: A database of life-history traits of the Northwest European flora. *Journal of Ecology*, *96*(6), 1266–1274. doi:*10.1111/j.1365-2745.2008.01430.x*

KNB. (2011). Ecological Metadata Language (EML) Specification. Retrieved 10 November 2017, from *https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/index.html*

Laporte, M.-A., Garnier, E., & Mougenot, I. (2013). A faceted search system for facilitating discovery-driven scientific activities: A use case from functional ecology. *Semantics for Biodiversity (S4BioDiv 2013), 25*. Retrieved from *https://hal-lirmm.ccsd.cnrs.fr/docs/00/83/17/57/PDF/Proceedings_S4BioDiv-2013.pdf#page=27*

Lavorel, S., & Grigulis, K. (2012). How fundamental plant functional trait relationships scale-up to trade-offs and synergies in ecosystem services. *Journal of Ecology*, *100*(1), 128–140. doi:*10.1111/j.1365-2745.2011.01914.x*

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., & Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, *2*(3), 279–296. doi:*10.1016/j.ecoinf.2007.05.004*

Madin, J. S., Anderson, K. D., Andreasen, M. H., Bridge, T. C., Cairns, S. D., Connolly, S. R., … Franklin, E. C. (2016). The Coral Trait Database, a curated database of trait information for coral species from the global oceans. *Scientific Data, 3,* 160017. doi:*10.1038/sdata.2016.17*

McGill, B. J., Enquist, B. J., Weiher, E., & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, *21*(4), 178–185. doi:*10.1016/j.tree.2006.02.002*

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, *1*(1), 3–7.

Michener, W., Vieglais, D., Vision, T., Kunze, J., Cruse, P., & Janée, G. (2011). DataONE: Data Observation Network for Earth - Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. *D-Lib Magazine*, *17*. doi:*10.1045/january2011-michener*

Moretti, M., Dias, A. T., Bello, F., Altermatt, F., Chown, S. L., Azcárate, F. M., … others. (2017). Handbook of protocols for standardized measurement of terrestrial invertebrate functional traits. *Functional Ecology*, *31*(3), 558–567. doi:*10.1111/1365-2435.12776*

Mouillot, D., Graham, N. A. J., Villéger, S., Mason, N. W. H., & Bellwood, D. R. (2013). A functional approach reveals community responses to disturbances. *Trends in Ecology & Evolution*, *28*(3), 167–177. doi:*10.1016/j.tree.2012.10.004*

Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., & Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, *13*(1), R5. doi:*10.1186/gb-2012-13-1-r5*

Oliveira, B. F., São-Pedro, V. A., Santos-Barrera, G., Penone, C., & Costa, G. C. (2017). AmphiBIO, a global database for amphibian ecological traits. *Scientific Data*, *4*, sdata2017123. doi:*10.1038/sdata.2017.123*

Page, R. D. M. (2008). Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, *9*(5), 345–354. doi:*10.1093/bib/bbn022*

Park, C. A., Bello, S. M., Smith, C. L., Hu, Z.-L., Munzenmaier, D. H., Nigam, R., … Reecy, J. M. (2013). The Vertebrate Trait Ontology: A controlled vocabulary for the annotation of trait data across species. *Journal of Biomedical Semantics*, *4*(1), 13. doi:*10.1186/2041-1480-4-13*

Parr, C. L., Dunn, R. R., Sanders, N. J., Weiser, M. D., Photakis, M., Bishop, T. R., … Gibb, H. (2017). GlobalAnts: A new database on the geography of ant traits (Hymenoptera: Formicidae). *Insect Conservation and Diversity*, *10*(1), 5–20. doi:*10.1111/icad.12211*

Parr, C. S., Schulz, K. S., Hammock, J., Wilson, N., Leary, P., Rice, J., … J, R. (2016). TraitBank: Practical semantics for organism attribute data. *Semantic Web*, *7*(6), 577–588. doi:*10.3233/SW-150190*

Perez-Harguindeguy, N., Diaz, S., Garnier, E., Lavorel, S., Poorter, H., Jaureguiberry, P., … Gurvich, D. E. (2013). New handbook for standardised measurement of plant functional traits worldwide. *Australian Journal of Botany*, *61*(3), 167–234.

Pey, B., Laporte, M.-A., Nahmani, J., Auclerc, A., Capowiez, Y., Caro, G., … Hedde, M. (2014). A Thesaurus for Soil Invertebrate Trait-Based Approaches. *PLOS ONE*, *9*(10), e108985. doi:*10.1371/journal.pone.0108985*

Robertson, T., Döring, M., Wieczorek, J., De Giovanni, Renato, & Vieglais, D. (2009, February 12). Darwin Core Text Guide. Retrieved 30 October 2017, from *http://rs.tdwg.org/dwc/terms/guides/text/index.htm*

Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology*, *13*(11), e1002295. doi:*10.1371/journal.pbio.1002295*

Royal Society Science Policy Centre. (2012). *Science as an open enterprise*. London, UK: The Royal Society. Retrieved from *https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/*

Salguero-Gómez, R., Jones, O. R., Jongejans, E., Blomberg, S. P., Hodgson, D. J., Mbeau-Ache, C., … Buckley, Y. M. (2016). Fast–slow continuum and reproductive strategies structure plant life-history variation worldwide. *Proceedings of the National Academy of Sciences*, *113*(1), 230–235. doi:*10.1073/pnas.1506215112*

Schildhauer, M., Jones, M. B., Bowers, S., Madin, J., Krivov, S., Pennington, D., … O'Brien, M. (2016). OBOE: Extensible Observation Ontology. *OBOE: The Extensible Observation Ontology, Version*

1.1. KNB Data Repository. doi:*10.5063/F11C1TTM*

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–1255. doi:*10.1038/nbt1346*

Smith, V. S., & Blagoderov, V. (2012). Bringing collections out of the dark. *ZooKeys*, (209), 1–6. doi:*10.3897/zookeys.209.3699*

Ströbel, B., Schmelzle, S., Blüthgen, N., & Heethoff, M. (2018). An automated device for the digitization and 3D modelling of insects, combining extended-depth-of-field and all-side multi-view imaging. *ZooKeys*, (759), 1–27. doi:*10.3897/zookeys.759.24584*

Villéger, S., Brosse, S., Mouchet, M., Mouillot, D., & Vanni, M. J. (2017). Functional ecology of fish: Current approaches and future challenges. *Aquatic Sciences*, *79*(4), 783–801.

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. (2007). Let the concept of trait be functional! *Oikos*, *116*(5), 882–892. doi:*10.1111/j.0030-1299.2007.15559.x*

Walls, R. L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M. A., Jaiswal, P., … Stevenson, D. W. (2012). Ontologies as integrative tools for plant science. *American Journal of Botany*, *99*(8), 1263–1275. doi:*10.3732/ajb.1200222*

Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., … Endresen, D. (2014). Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS One*, *9*(3), e89606.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., & Musen, M. A. (2011). BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, *39*(suppl_2), W541–W545.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. doi:*10.18637/jss.v059.i10*

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PloS One*, *7*(1), e29715. Retrieved from *http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029715*

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. doi:*10.1038/sdata.2016.18*

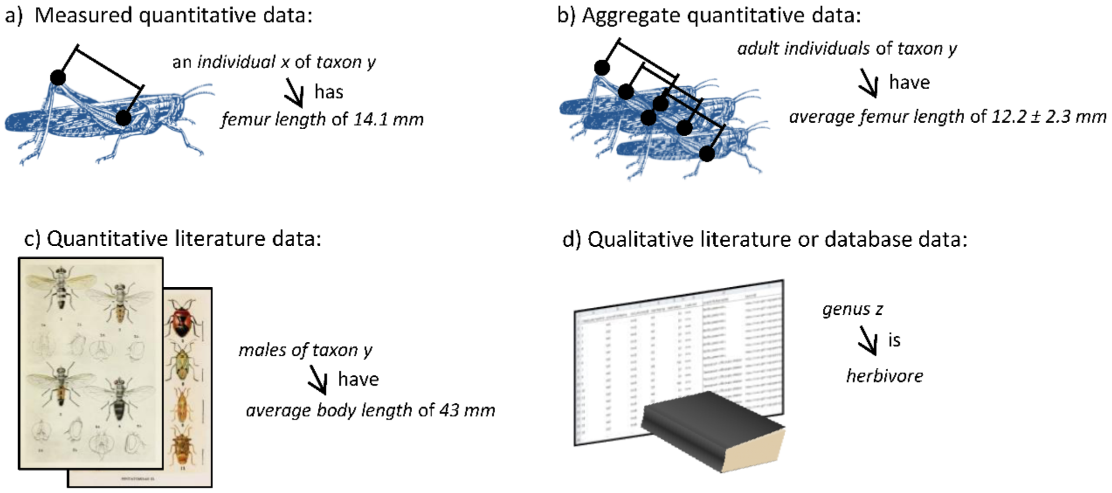
Yoder, M. J., Miko, I., Seltmann, K. C., Bertone, M. A., & Deans, A. R. (2010). A gross anatomy ontology for Hymenoptera. *PloS One*, *5*(12), e15991.

Table 1 | Glossary of terms from the biodiversity data-management context as they are used in this paper; draws from Garnier et al. (2017).

Term	Definition
Concept	An idea, notion or object that is made explicit in an information context by a *term* definition, and referenced to a *URI* or other accessible reference.
Controlled vocabulary	A list of *terms* that gives all valid consensus terms for a particular context, while no unlisted entries are accepted.
Darwin Core Standard (DwC)	Body of *terms* intended to facilitate the sharing of information about biological diversity; maintained by the Biodiversity Information Standards TDWG (*http://rs.tdwg.org/dwc/*).
Dataset	A set of measurements and observations, often stored in a data-table and originating from a single experimental set-up or study context; can be considered as being internally homogeneous across all data entries.
Database	A structured collection of data, usually organised as multiple data tables linked via *identifiers* into relational databases; usually constructed using a specific database management system, i.e., a software to provide a (offline or online) user interface.
File repository	A storage or archiving of datasets on file-hosting services like Figshare.com, Dryad (datadryad.org), Researchgate.net, or Zenodo.org; online repositories make data available for public access, provide *metadata*, state conditions of re-use, and (not always) facilitate citations via persistent identifiers, e.g., DOIs (Digital Object Identifiers).
Identifier (ID)	A unique label that relates data entries to information within and across *datasets* or external items of information; may be used to connect multiple *data tables* into a *database*; can be user-specific or, in form of a *URI*, point to a globally valid *ontology* or *thesaurus*.
Metadata	Data documentation of the higher-level information or instructions; describe the content, context, quality, structure, provenance and accessibility of a data object (Michener, 2006). In the context of trait data, such additional information can move to the body of the primary data table when data are compiled from different sources.
Occurrence	The observation context of a single individual, i.e., the existence of an organism at a particular place and time; Sometimes used as synonym of 'observation' in data management context.
Ontology	A semantic model of the objects and their relationships in a domain of interest (Gruber, 1995); defines *terms* and *concepts* in a formal language that provides cross-references and semantic meaning; commonly published in OWL format for machine readability.
Semantic web	An extension of the World Wide Web that aims for machine-readable meaning of information via well-defined *data standards*, *ontologies* and exchange protocols (Berners-Lee et al., 2001); the World Wide Web Consortium (W3C) defines standards, i.e., specifications of protocols and technologies for the semantic web (*http://www.w3.org/standards/semanticweb/*).
Term	A word that names or labels a particular *concept* as part of the specialised vocabulary of a field.
Terminology	The body of *terms* and *concepts* used with a particular application in a subject of study, usually formalised in a *thesaurus* or *ontology*.
Thesaurus	*Controlled vocabulary* that provides key *terms* with their associated *concepts* and relations for a specific field or domain of interest (Laporte et al., 2013); e.g., may define a hierarchy of broader or narrower terms.
Uniform Resource Identifier (URI)	An unambiguous pointer to a unique resource on the Internet; used to refer to single terms of a *thesaurus* or *ontology*; Example: '*http://purl.obolibrary.org/obo/TO_0000391*'.

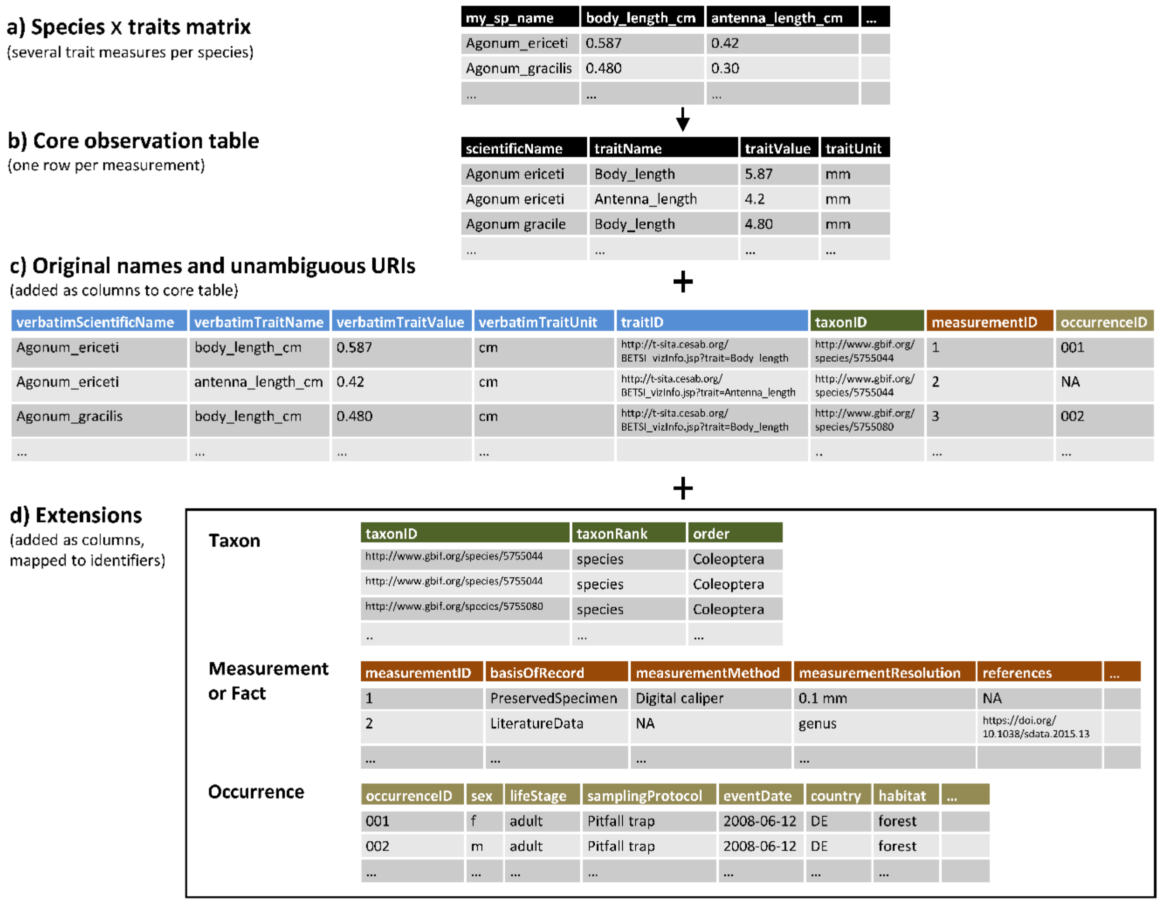
Figures

Figure 1


a) Measured quantitative data:
an individual x of taxon y
has
femur length of 14.1 mm
b) Aggregate quantitative data:
adult individuals of taxon y
have
average femur length of 12.2 ± 2.3 mm
c) Quantitative literature data:
males of taxon y
have
average body length of 43 mm
d) Qualitative literature or database data:
genus z
is
herbivore

Types of ecological trait data assume different entities or reported qualities: a) morphometric or morphological measurements of individual body features (lengths, areas, volumes, weights) or other quantities related to life history (e.g., reproductive rates, life spans); b) aggregated trait values are reported as means taken on multiple measures of organisms of a taxon; c) quantitative traits may be extracted from literature or existing databases, referring to the entire taxon (or a subset, e.g., a sex) as the subject of description; d) qualitative traits are categorical, ordinal or binary descriptors of the entire species or higher taxonomic level (also called 'facts').

Figure 2



a) Species x traits matrix
(several trait measures per species)

my_sp_name	body_length_cm	antenna_length_cm	...
Agonum_ericeti	0.587	0.42	
Agonum_gracilis	0.480	0.30	
...	...	...	

b) Core observation table
(one row per measurement)

scientificName	traitName	traitValue	traitUnit
Agonum ericeti	Body_length	5.87	mm
Agonum ericeti	Antenna_length	4.2	mm
Agonum gracile	Body_length	4.80	mm
...	...	...	...

+

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

verbatimScientificName	verbatimTraitName	verbatimTraitValue	verbatimTraitUnit	traitID	taxonID	measurementID	occurrenceID
Agonum_ericeti	body_length_cm	0.587	cm	http://t-sita.cesab.org/ BETSI_vizInfo.jsp?trait=Body_length	http://www.gbif.org/ species/5755044	1	001
Agonum_ericeti	antenna_length_cm	0.42	cm	http://t-sita.cesab.org/ BETSI_vizInfo.jsp?trait=Antenna_length	http://www.gbif.org/ species/5755044	2	NA
Agonum_gracilis	body_length_cm	0.480	cm	http://t-sita.cesab.org/ BETSI_vizInfo.jsp?trait=Body_length	http://www.gbif.org/ species/5755080	3	002
...	...	...	...		..	...	...

+

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

Taxon

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

Measurement or Fact

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

Occurrence

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

Formats used for trait datasets. a) taxon-level trait data compiled from literature or aggregated from measurements are often published as a compiled species × traits wide-table; b) observation long-tables are a well-defined and tidy data format, reporting one single measurement per row and relating it to a standard trait definition and accepted taxon name; c) additional columns may provide original names for maintaining author-side continuity, identifiers reference to taxa and trait concepts via unambiguous URI pointers. Additional identifiers relate each row to other layers of information on d) the taxon resolution, the individual organism (i.e. occurrence), or the origin or confidence on the reported measurement or fact.