Edaphobase 2.0: Advanced international data warehouse for collating and using soil biodiversity datasets

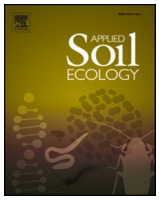
David J Russell, Evi Naudts, Nadia A Soudzilovskaia, Maria J I Briones, Meriç Çakır, Erminia Conti, Jérôme Cortet, Cristina Fiera, Hackenberger Kutuzovic Davorka, Mickael Hedde, et al.

Applied Soil Ecology 204 (2024) 105710

Contents lists available at ScienceDirect

Applied Soil Ecology

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



Edaphobase 2.0: Advanced international data warehouse for collating and using soil biodiversity datasets

D.J. Russell *, E. Naudts, N.A. Soudzilovskaia, M.J.I. Briones, M. Çakır, E. Conti, J. Cortet, C. Fiera, D. Hackenberger Kutuzovic, M. Hedde, K. Hohberg, D. Indjic, P.H. Krogh, R. Lehmitz, S. Lesch, Z. Marjanovic, C. Mulder, L. Mumladze, M. Murvanidze, S. Rick, M. Roß-Nickoll, J. Schlaghamerský, O. Schmidt, O. Shelef, M. Suhadolc, M. Tsiafouli, A. Winding, A. Zaytsev, A. Potapov

Senckenberg Museum fur Naturkunde Goerlitz, Goerlitz, Germany

ARTICLE INFO

Keywords:
Soil biodiversity
Data warehouse
Edaphobase
Environmental metadata
Data sharing

ABSTRACT

Soil and soil-biodiversity protection are increasingly important issues in environmental science and policies, requiring the availability of high-quality empirical data on soil biodiversity. Here we present a publicly available data warehouse for the soil-biodiversity domain, Edaphobase 2.0, which provides a comprehensive toolset for storing and re-using international soil-biodiversity data sets, following the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. A major strength is the possibility of annotating biodiversity data with exhaustive geographical, environmental and methodological metadata, allowing a wide range of applications and analyses. The system harmonises and integrates heterogeneous data from diverse sources into standardised formats, which can be searched together using numerous filter possibilities, and offers data exploration and analysis tools. Edaphobase features a strict data transparency policy, comprehensive quality control, and DOIs can be provided for individual data sets. The database currently contains >450,000 data records from >35,0000 sites and is accessed nearly 14,000 times/year. The data curated by Edaphobase 2.0 can greatly aid researchers, conservationists and decision makers in understanding and protecting soil biodiversity.

1. Introduction

Soil and soil-biodiversity protection are increasingly important issues in global environmental policy agendas (FAO and ITPS, 2015; Ronchi et al., 2019; FAO et al., 2020; EU Commission, 2021; FAO, 2022). Existing challenges include the sustainable use and preservation of the multi-functionality of soils (i.e., Vrebos et al., 2017), which relies on rich soil biodiversity (Loreau et al., 2001). Healthy functioning of the soil ecosystem enables the delivery of multiple ecosystem services, such as food provisioning for the growing human population, water purification, carbon sequestration to stabilise climate, and resilience against multiple natural and anthropogenic pressures (Fan et al., 2023), including soil pollution (Stolte et al., 2016), warming, drought, heavy metal contamination (Rillig et al., 2021), plastic residues (Rillig and Lehmann, 2020; Li et al., 2022) and pesticides (Zhou et al., 2020; Riedo et al., 2021). Most of these ecosystem services are supported and regulated by the activities of soil organisms, including fauna and microorganisms (Gardi and Jeffery, 2009; Turbé et al., 2010; Adhikari and Hartemink, 2016). Soil organisms are crucial for key ecological functions that underlie ecosystem services, such as soil aggregate formation (Scheu and Schulz, 1996; Frouz et al., 2006; Coleman et al., 2017; Shelef et al., 2019, 2020), plant litter and other detritus decomposition (Smith and Bradford, 2003; Castro-Huerta et al., 2015), and nutrient cycling (Coleman et al., 1993; Jeffery et al., 2010; Shelef et al., 2013).

Developing management practices for soil protection and conservation requires data- and knowledge-based evaluation of the composition and functioning of soil-organism communities (Dunbar et al., 2013). For that, high-quality, spatially explicit (georeferenced) empirical quantitative data on soil biodiversity are urgently needed for the evaluation of effects of, e.g., agricultural intensification or soil-management practices (Mathews et al., 2020) as well as for developing efficient soil management and protection policies. Yet, the effects of soil organisms on many

* Corresponding author.
E-mail address: david.russell@senckenberg.de (D.J. Russell).

https://doi.org/10.1016/j.apsoil.2024.105710
Received 23 April 2024; Received in revised form 6 October 2024; Accepted 16 October 2024

ecosystem functions, as well as the response of soil organisms to environmental changes, remain poorly understood. A number of recent analyses attempting to understand the role of soil biodiversity as mediated by a suite of environmental factors across large biogeographical regions have collated, synthesised and harmonised datasets assembled to analyse individual groups of soil organisms (e.g., van den Hoogen et al., 2019; Phillips et al., 2021; Potapov et al., 2023). However, assembling a complex ecologically multidimensional dataset from multiple individual datasets requires a huge work investment, which has essentially been limited to very few attempts. As such, establishing a sustainable and curated infrastructure for standardised, quality-controlled data is essential for generating and providing the knowledge needed for protecting soils and the biodiversity therein.

The last decade has shown an increased awareness throughout all science domains of the need to share and provide research data publicly (Van den Eynden and Corti, 2014; Alter and Vardigan, 2015; Gewin, 2015), following FAIR data principles (Wilkinson et al., 2016; Jacobsen et al., 2020; https://www.go-fair.org/fair-principles). However, in view of the challenges discussed above, applying the FAIR principles in soil-biodiversity data remains difficult due to the highly fragmented availability of data sets, lack of globally accepted data-generation methods and disparate understandings of the role of 'essential portions' of soil biodiversity and therefore of the necessary data to be collated. Nonetheless, protocols for assessing the current status of soil biodiversity must be science- and data-based (using quantitative data and standardised methods: Potapov et al., 2022), comprising broad-scale yet site-specific information on species taxonomic diversity, functional diversity and environmental metadata (Emmerling et al., 2002; Ramirez et al., 2015; Wall et al., 2015; Römbke et al., 2018; Sechi et al., 2018). However, to date, there have been more calls in this direction than concrete responses and actions. A recent call for collaboration presents an incentive for sharing data by co-constructing a European Atlas of Soil Fauna (Tsiafouli et al., 2022).

In recent decades advanced methodologies for mapping biodiversity and relating it to environmental drivers have been developed (Ferrier et al., 2007; Guisan et al., 2017). Some Proof-of-Concepts (PoC) have illustrated the viability of mapping soil biodiversity for establishing potential baselines (Rutgers et al., 2016; Salako et al., 2023). However, these methods are based on time-consuming hand collation of both biodiversity data and environmental data. International biodiversity databases such as GBIF, PREDICTS, TRY, BOLD or DiSSCo, although focusing on global diversity of plants and animals, still have limited data on soil organisms and often no data on species abundances, habitat types or soil types and other environmental characteristics of the sites of occurrence. More importantly, these databases are often not operational for soil ecological assessments or advancing decision-support tools, as they hardly contain information on functional traits of organisms (cf. TRY database of plant functional traits [Kattge et al., 2020] or BETSI database for soil animal traits [http://betsi.cesab.org/]), and in many cases they lack the environmental metadata needed to provide details about the ecosystems where soil biota were collected.

Therefore, a data warehouse specifically designed for the soil-biodiversity domain is an essential step forward. The major challenge in developing state-of-the-art data repositories and warehouses is to build a unified framework for integrating an inherently massively multidisciplinary field (in this case, species taxonomies, occurrences and abundances, including their functional traits and molecular data; but also information about land use, vegetation cover, and soil chemical and physical properties, etc.). Development of such a data-diverse multidisciplinary repository requires at least three steps. Firstly, researchers with different expertise (e.g. taxonomists, soil scientists, ecologists) must agree to collaborate, combine, and share their data, thereby establishing a culture of data visibility, management and re-use. Secondly, a data infrastructure must be established and maintained where shared data can be assimilated and disseminated in the long term. Thirdly, efficient data collection, standardisation and software application tools must be developed to meet the data workflow and visualisation needs of data users. Many scientific disciplines have already developed data resources that successfully facilitate data-intensive research such as molecular and biotechnological data. Due to the complexity and high variability of the methods used to record and assess soil-organism communities, integrating heterogeneous data sources - including the various levels of morphological and molecular taxonomic identification and the difficulties in combining numerical counts and molecular sequences - into complexly linked datasets needed for assessing soil biodiversity and functioning at broad spatio-temporal scales constitutes a further challenging task. Also, detailed methodological metadata are required to enable internal comparability of data used for common meta-analyses. Finally, quality control of uploaded data as well as applying internal data filters and quality-assessment tools are critically needed procedures for assembly of soil biota datasets.

Here, we present the first and largest international data warehouse devoted solely to the soil-biodiversity domain, Edaphobase 2.0. This database has evolved from an earlier version (Burkhardt et al., 2014), which focused on the entry of data for the major soil invertebrate-animal taxa occurring in Germany and neighbouring countries. Based on this earlier version, we developed an advanced data repository, which includes elements of a data warehouse and provides a comprehensive set of tools to (i) store soil-biodiversity data accompanied by detailed metadata, and (ii) re-use the data for new analyses, following the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Much of the current development was undertaken within the framework of the EU COST Action 'EUdaphobase' (https://www.eudaphobase.eu/), a consortium of ca. 100 soil scientists, biologists and software experts from over 30 countries from Europe and beyond, establishing a consensus on data collation and re-use goals as well as on standardisation, data structures and policies. In Edaphobase 2.0, heterogeneous data are harmonised and stored in standardised formats, which can be searched using numerous filter possibilities (e.g., by taxa, regions, habitat types, data sources, etc.) and offers basic analytical tools allowing data exploration aimed for a wide range of applications. Edaphobase 2.0 features a strict data transparency policy, quality control, and DOIs can be provided for individual data sets. The open-access data curated by Edaphobase 2.0 can greatly aid researchers, conservation scientists and decision makers to further progress in understanding and conserving soil biodiversity and protecting soils.

2. Methods

2.1. Edaphobase data management

The Edaphobase data warehouse is open and free for submission of soil-biodiversity data by any and all data providers, whereby recognition of data providers is a key priority. The data warehouse is structured in a manner that allows all soil-biodiversity data to be easily obtained and be appropriately linked to relevant metadata in order to answer scientific questions or to assess ecological and distributional information on soil organisms (see below, Section 2.1.3 and Fig. 3a). To this goal, the core focus is on taxonomic and observational data, whereby possibilities and guidelines for submitting exhaustive geographical, environmental and methodological metadata are also provided.

2.1.1. Data sources and intellectual rights

Data from various source types of data providers are integrated in Edaphobase, for instance raw data from research projects or monitoring programs; published literature; unpublished reports and theses, museum collections, etc. Detailed information on the data source helps to organise data for further re-uses and interpretations as well as allowing proper citation of data sets used in meta-analyses and data syntheses. Metadata for the source of a data set includes the project name and principal investigator (for project sources), the collection name, collection-object number and collection manager (for museum sources),

or the author, journal, article title, volume and page numbers, DOI (for literature sources), etc.

Data providers retain the ownership rights to their data while acknowledging that data uploaded to Edaphobase are open access according to CC-BY licences, although certain restrictions to data access can be superimposed. The provider of the original data set can specify a priori the public-access extent to which their data will be made available: (i) open access (ii) permanently restricted public access to sensitive parts of the data set ('anonymised'), or (iii) temporarily blocked for release ('embargoed' until the original study is published, etc.). While these safeguards have been installed according to data-providers concerns, in reality 95 % of all datasets are open access (see Section 3.3 below). Anonymised data sets are not hidden entirely, but only the sensitive aspects (primarily: location data or related methodological metadata) of individual data records within the data set (viable to the public as "available upon request"). The source metadata, non-sensitive data and similar remain open and publicly available. Anonymised sections or embargoed data can be made available to a specific data user upon request to the original data owner. Further information can be found in the Edaphobase data policy (Senckenberg, 2023).

2.1.2. Taxonomic system of soil biodiversity taxa

The taxonomic module is the core module of the system. To be unambiguous, a taxon is defined by a nomenclatural full name (including describing author and year) and is hierarchically classified within a systematic tree representing the 'taxonomic backbone' of Edaphobase. This backbone is based on available taxonomic databases and resources (i.e., Ghilarov and Krivolutsky, 1975; Bellinger et al., 1996ff; Csuzdi and Zicsi, 2003; Weigmann, 2006; Degma and Guidetti, 2007; Boyko et al., 2008; de Jong et al., 2014; DriloBase Project, 2014ff; Peter et al., 2019; GBIF, 2020; Ahyong et al., 2023; Sierwald and Spelda, 2023; etc.), revised and kept up-to-date by the European taxonomic expert(s) responsible for the respective taxon group in Edaphobase. For Fungi and Prokarya, the system is linked to the internationally renowned databases Mycobank (Robert et al., 2013), UNITE (Nilsson et al., 2018) and the Bac*Dive* system (Reimer et al., 2022), which provide curated, up-to-date taxonomies as data layers to Edaphobase. Based on the systematic tree, taxa at any hierarchical level can be queried and the system will find data on all subordinate taxa. Synonyms of taxon names are linked to their valid names so that older taxonomic designations can be found and appropriately handled.

2.1.3. Observational data and metadata

Edaphobase allows data to be submitted as individual data sets (containing multiple observation records) and metadata valid for an entire data set. Number of individuals (relative and/or absolute), numerical densities and biomass are examples of soil-biodiversity observation records, while taxon-identification and sampling methods, preparation and morphological or molecular methods often describe the entire data set and can therefore be considered as metadata. Nonetheless, the system allows different metadata to be linked to individual data records within a data set. The inclusion of metadata enhances data reuse and assists in filtering data appropriate for specific analyses. Three types of metadata are handled in Edaphobase based on information on (i) the dataset itself and its source, (ii) the sampling locations and (iii) the methods (field, laboratory, taxon identification).

2.1.3.1. Observational data. Besides qualitative information about species occurrences ('presence' in a specific sample, site, etc.), Edaphobase can host quantitative data linked to a specific taxon and sample/site. Quantitative data can be expressed in absolute counts per sample, density or biomass per area or volume, activity density (for traps), dominance (relative abundance), frequency, etc. The main harmonisation units within Edaphobase are individuals/m^2 and g/m^2 (for area-based sampling methods; data fields for sampling depth also exist) and individuals/trap/time period (for trap-based methods). Absolute counts are automatically harmonised to these units, provided that the number of samples and sampling effort are also given (i.e. number of samples per plot or site, sampled area or exposition time). To maintain computational efficiency with the database, raw records of 'absence' are generally not stored. However, for data sets at the community level, based on the 'scope' data field in the metadata (see below, Section 2.1.3.4) the system generates absences ('pseudo-absences'; species contained in the data set are assigned with zero occurrence values if not included in a sample or site listed in the dataset) for data-aggregation functions. This is done, e.g., to allow correct aggregation of sample-level data to 'site'-level summaries, which represents the main spatial harmonisation level of Edaphobase (see below, Section 2.1.3.3). Measurements and observations on individual specimens (i.e., morphological or ecological characteristics, as well as stabile-isotope, fatty-acid or gut-content analyses) as well as molecular data can be included in a submitted data set for individual taxa.

2.1.3.2. Metadata regarding the data set. Full metadata concerning a data set is required during submission and allows easy access of the individual data set as well assessment of re-use possibilities and correct citation of the data set by users. Mandatory metadata for each individual data set includes the dataset title, the source & source type, the author or PI, region/country of origin, locations & geo-coordinates, taxonomic groups(s) contained, DOI (when requested by the data provider), etc. This metadata can be accessed in the publicly available online Edaphobase Portal (https://portal.edaphobase.org), and is included in all downloaded data files, whereby this metadata is never limited. Even in download tables of collated data (which includes records from various data sets), all metadata for each data point (derived from the data set containing the point) is provided.

2.1.3.3. Environmental and geographical metadata. An important metadata category ('site description') includes data fields for information on the location (incl. geographical coordinates), sampling date(s), local weather (i.e. temperature, precipitation etc. during or, e.g., shortly before the sampling period), climate (long-term averages of sampling sites or areas), vegetation (species specific, including cover percentages), habitat (biotope) and microhabitat type, land use as well as specific soil properties (i.e., physical, chemical, hydrological parameters, etc.). Each field can be linked to the associated methods used for their generation as well as the dates of the measurement period, in order to associate such metadata with the soil-biological data. In contrast to other data repositories, environmental metadata can be linked to observational data at various (hierarchical) spatial and geographical scales. Both biodiversity data and metadata can be entered and linked from the subsample level to an individual sample, plot represented by several samples, site represented by several plots, or selected region/country (Fig. 1). This is established by specifically linking environmental and methodological data to spatio-temporal variables (sampling event [i.e. date or time period], individual sample [i.e., a soil core or trap], subsample [i.e., vertically divided soil core or exposition period of a trap] and - if the data are not at a sample level - linkage to a specific plot or site). These spatial hierarchical levels are specifically defined (https://service.edaphobase.org/Edaphobase_datafield_documentation). The 'site' (of occurrence) is the main spatial harmonisation level of Edaphobase for cross-data set analyses. Any data available at more specific spatial levels (plot, sample, sub-sample) are automatically aggregated (via sums or means, as appropriate) in Edaphobase to the site level. Nonetheless, all original data (including sample-level data or treatments in separate plots) is retained in the system, allowing further options for the user. For instance, data can be specifically filtered to obtain treatment comparisons by choosing to group (aggregate) data at the "plot" level, which will isolate treatment data in contingency tables, provided that such data is indeed given by data contributors.

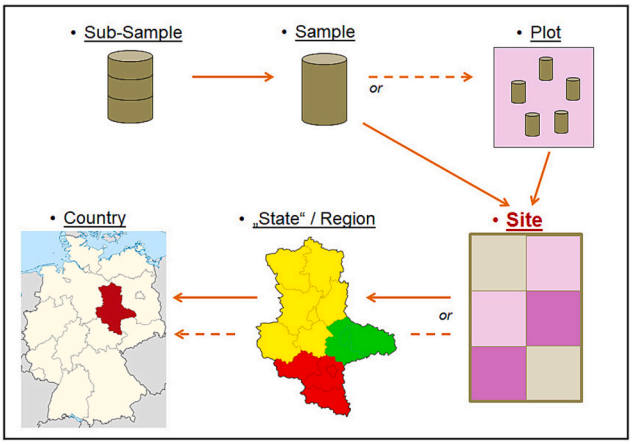


Fig. 1. Hierarchical spatial levels available in Edaphobase, from the most local ('Sub-Sample' in the upper left) clockwise to the broadest level ('Country' in the lower left). Data can be entered at any hierarchical level. Each spatial level is nested in the next higher level (if provided) and Edaphobase automatically aggregates lower-level data to higher spatial levels. 'Site' is the spatial harmonisation level for data analyses across all data sets (whereby data only at a regional or country level is not included).

Furthermore, sample-level data can be specifically filtered and downloaded for analyses of local-scale (site) spatial heterogeneity.

2.1.3.4. Methodological metadata. Submitted methodological metadata allows provision of information on the fieldwork during a sampling event and/or in the laboratory (i.e., sampling date, sampling methods, extraction methods, species identification, etc.). An important set of methodological metadata are data fields describing the 'data scope', which stores information on the goal/purpose of data collection and the type of biological data included in the dataset. They can be considered as describing the objectives and goals of a sampling event or survey. The information allows selective data filtering for improved data re-use functionality, since not every data set is appropriate for every type of analysis. The four data fields within 'data scope' are: sampling effort (single observations or systematic survey), quantification level (quantitative or qualitative [presence/absence]), selected species or community-composition level, and sample-level or (aggregated) site-level data. Again, these data fields are defined and explained in detail and available to the public (https://service.edaphobase.org/Edaphobase_datafields and https://service.edaphobase.org/Edaphobase_datafield_documentation).

2.2. Data model

To integrate all submitted data sets, Edaphobase implements a 'hybrid' relational data model in the database-management system PostgreSQL (Fig. 2). The model combines a conventional relational model and an entity-attribute-value (EAV) information system (Nadkarni et al., 1999). One major advantage of the EAV model is the flexibility to introduce new attributes (=datafields/variables) without requiring major changes in the underlying data model. The EAV system also creates space for complex hierarchies and supports efficient data exploration (since zero values are excluded).

This core database is stored only on internal servers to ensure security and prevent external manipulation ('hacking'). The database is transferred daily from this transaction-optimised schema to a read-only analysis-optimised schema in the publicly available Edaphobase Data-Query Portal (https://portal.edaphobase.org). Thereby, all data sets are pre-prepared in a manner that can be queried and combined as efficiently as possible in the Query Portal. This procedure also protects personal data of, i.e., data providers, which are only maintained in the internal database and not sent to the Data-Query Portal. The following technologies are used for the Data-Query Portal: Qooxdoo framework for graphical user interface (GUI) and OpenLayers for map presentation, MapServer for portions of the map management system, PHP for communication between database and front end, PostgreSQL for persistence, PostGIS for persistence and GIS calculations.

2.3. Sustainability

The Senckenberg Society is committed to sustainably maintaining the data infrastructure in the future, which is considered to be a "flagship product" of the entire institution. Currently, an Edaphobase manager and two software developers for the system have permanent positions; permanent scientific staff of the Department of Soil Zoology at Senckenberg have large portions of their daily tasks devoted to managing and further developing the Edaphobase system. This staff is augmented by further personnel from third-party financed projects, ensuring the sustainability and permanent availability of the Edaphobase data warehouse.

3. Results

3.1. Data warehouse

Edaphobase 2.0 is a sustainable, yet dynamic data repository, implementing structures of a data warehouse (Kimball and Ross, 2002; Inmon, 2005), that stores data sets containing occurrence and abundance information on soil organisms, their spatial and temporal distribution coupled with the habitat parameters of their sites of occurrence, and makes these data sets available to the public. Current developments include linking the system to trait databases to also allow representation of *functional* aspects of soil biodiversity (see Discussion). Heterogeneous data from various source types (such as literature, museum collections, unpublished or raw project data) are integrated, quality checked, and homogeneously structured within the database. An open-access Data-Query Portal (https://portal.edaphobase.org) allows exploration and download of data sets, as well as more complexly filtered queries throughout all data sets about species' communities in specific sites, habitat types or specific environmental conditions (soil, climate, etc.), creation of distribution maps for individual taxa, etc. It also provides basic data-exploration tools at both species and community levels. The following soil-organism groups are currently included: Crassiclitellata (earthworms), Enchytraeidae (potworms), Diplopoda (millipedes), Chilopoda (centipedes), Isopoda (woodlice), Collembola (springtails), Oribatida (moss mites or armoured mites), Mesostigmata (predatory mites), and Nematoda (roundworms). Recently, Tardigrada (water bears), Diplura (two-pronged bristletails), Protura (proturans) and microorganisms - including Fungi and fungi-like organisms as well as Prokaryota (Bacteria and Archaea) - have been included. Edaphobase can be easily expanded to include other soil-organism groups, whereby protists are the current next target.

3.2. Edaphobase structure

The structure of any database is guided by the goals and requirements of data usage (cf. Bray, 2002; Dick et al., 2017). The main focus of Edaphobase is providing a data infrastructure for understanding the distribution of soil biodiversity and the drivers thereof as well as its responses to environmental changes (Fig. 3a). The data-linkage between species-occurrence data and spatial, environmental and species' trait data is a major strength of the database and is specifically built to assist with exploring information on species' distributions and environmental

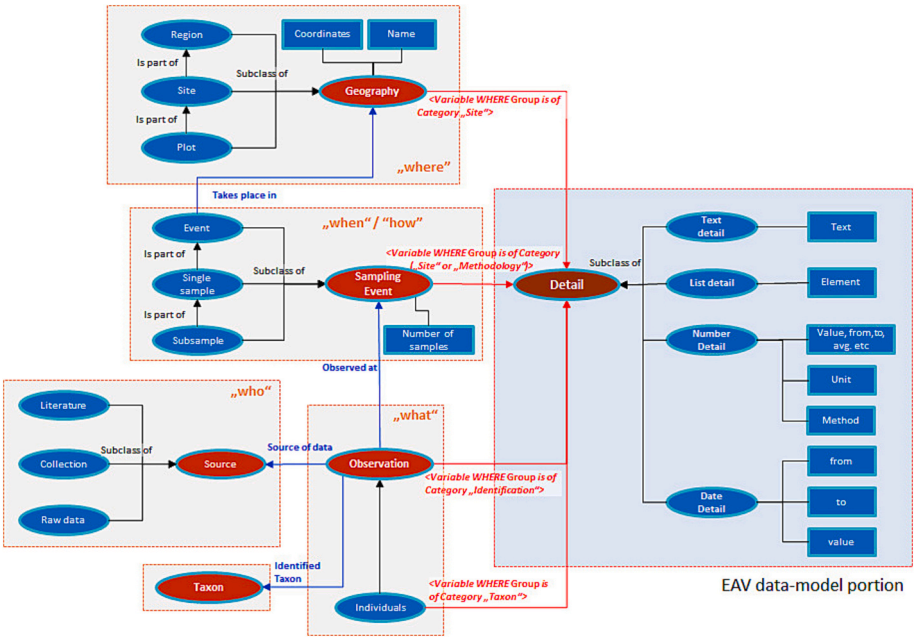


Fig. 2. Schematic overview of the 'hybrid' relational data model underlying the Edaphobase Data Warehouse. Most quantitative and categorical data entries are stored in the EAV part of the database (on the right of the figure, blue shading), with general data and metadata in the conventional relational database (left in the figure, grey shading). Major data-field classes (categories) in red ovals, and subclasses in blue ovals; examples of specific data fields in blue boxes. Red text denotes data relationships to the EAV tables; blue text relationships (properties) between relational tables. For more details of table relationships in the data model, see Supplementary Fig. 1).

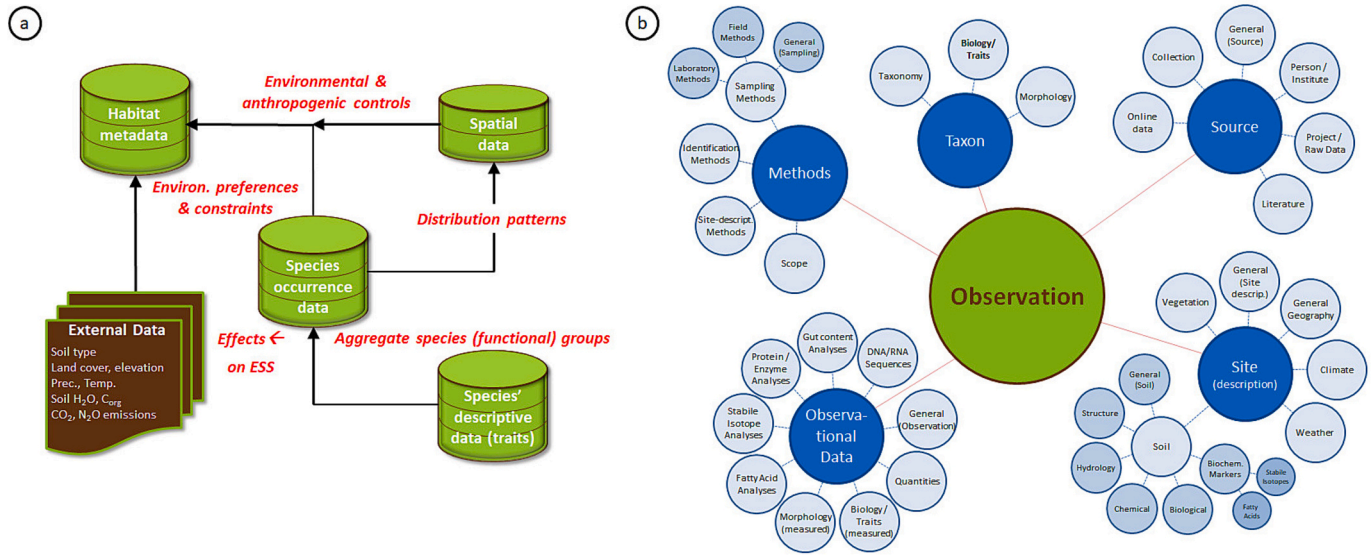


Fig. 3. Overview of Edaphobase data-field groups and their connections. a: Conceptual framework of how different data groups are combined in Edaphobase in order to answer different types of research and assessment questions. b: The main data-field classes (categories) and associated sub-classes in Edaphobase.

preferences (niche space). To manage all data sets in an understandable manner, the data fields[1] included in the database are organised into five major data categories (see Fig. 3b for data categories and their connections).

Data fields are hierarchically organised into categories (classes and subclasses, Fig. 3b and Supplementary Table 1), each category containing a set of data fields for entry of detailed biodiversity data (ca. 225 data fields) or - source, methodological or environmental - metadata (ca. 430 data fields; see below and Methods Section A.3 for a definition of Edaphobase metadata). Every data field has a unique identifier, name, definition, data format (number, string, element from a list, boolean) and measurement unit for numeric variables. For categorical variables, a selection list of possible standardised data entries ('authority' list) underlies each data field, which can be amended and expanded as needed (e.g., when suggested by a user and approved by the Edaphobase administrators). This information is publicly available at https://service.edaphobase.org/Edaphobase_datafields. For every data field, the data-field description also includes the data (sub)category to which it

[1] 'Data field', 'variable', 'attribute' are essentially synonyms referring to a specific data content. We use 'data field' here to represent a database 'field' or 'cell' into which a specific data ('variable' values) can be entered.

belongs, the lowest possible spatial hierarchical level to which the data can be linked, and if the data field is mandatory, recommended, automatically generated, etc. The minimum mandatory dataset requires full metadata describing the dataset itself, as well as data on the taxa, site of occurrence (including geo-coordinates), and sampling date. Recommended are furthermore quantitative data (densities, biomass, etc.) as well as metadata on the sites of occurrence (habitat type, soil properties, land use) and methods used (field sampling, laboratory, molecular, taxon identification). All data records in a data set are thereby linked to their source (e.g., literature citation, collection, online resource, project, as well as the data provider's and owners' names, etc.). An ontology of the Edaphobase data fields and the associated data structure is currently being developed, which will allow improved linking of Edaphobase to other relevant databases in the future.

The taxon module is the taxonomic backbone of Edaphobase. A complete hierarchical taxonomy is maintained for each soil-organism group contained in Edaphobase (see https://service.edaphobase.org/Edaphobase_taxonomy_ontology for an ontology of Edaphobase's taxonomic backbone; Aldana-Martín et al., submitted). Taxon names are linked to the describing author and year of description for unambiguous identification. Taxonomic synonyms are also indicated so that taxa can be found both by currently valid as well as by older names. The taxonomic backbone for each group is based on external expert sources (see Section 2.1.2 for examples) or - where available - directly linked to taxonomic databases (i.e., MycoBank, BacDive for fungal and bacterial taxonomies) in order to maintain curated, up-to-date taxonomies. A taxonomic expert for each group is responsible for its maintenance. A platform of soil-biodiversity taxonomy is currently being developed to reinforce this curation and standardisation process.

Additional modules manage information on descriptive data of the sites of occurrence, sampling and other methodological details, which are all connected to the soil-biodiversity observational data (more information provided in the Methods section). For each observation within a data set (i.e., a record of a taxon at a site using a specific method), an individual data record of multiple data fields is constructed. Together, this information provides specific answers to the questions 'what was found where, when, by whom, and how?'. In total, >650 possible data fields are included, allowing harmonised and highly linked data-entry *possibilities* for each data set. Of these, only 17–22 (depending on the data content) are mandatory (see above) to ensure minimum information is provided for each taxon occurrence. The hierarchical structure grouping data fields into categories (described above) facilitates finding the required data field for specific information. While the nomenclatures used for labels of the data fields have been established in European consensus (i.e. via online discussion fora and meetings within the EU COST Action 'EUdaphobase') to be intuitively understandable by soil-biodiversity researchers, where possible the DarwinCore equivalents are associated (see https://service.edaphobase.org/Edaphobase_datafields for details).

3.3. Edaphobase data content

For the 13 soil-organism higher-level taxonomic groups ('clades' sensu Hedde et al., 2022) presently included in Edaphobase, over 450,000 records have been compiled for over 14,000 taxa at >35,000 sites worldwide. Of these records, 95 % are currently open access and available for general re-use, while ca. 5 % contain restricted-access data (e.g., due to sensitive data on protected species, privately owned locations, etc.). At the time of this writing, no data set is embargoed (e.g. due to unpublished studies or thesis work). Data on the various soil-organism groups are being continuously added to Edaphobase. Presently, the majority of data records concern invertebrate animal groups such as Collembola, Oribatida and Diplopoda (24.4 %, 16.7 % and 19.3 %, respectively, Table 1). Many data sets, especially concerning recently included groups (e.g., Fungi, Prokaryota, Tardigrada, Protura, Diplura), have been uploaded by different users and are in the process of quality review and will be imported and available soon. Data from multiple source types are integrated, with currently 32.8 % originating from various museum collections, 44.2 % from literature (published articles, unpublished reports, theses, etc. uploaded by diverse users) and 22.9 % from unpublished external sources such as raw project data. Fig. 4 shows the global distribution of the occurrence locations currently registered in the database. Edaphobase presently contains data for 170 countries, representing all continents including Antarctica. Since the initial geographic focus in the first version of Edaphobase was Central Europe (Burkhardt et al., 2014), the extent of most data currently deposited in Edaphobase is centred in – but not limited to – this area (see Supplementary Table 2 for the distribution of all data records among global countries).

3.4. Data integration

3.4.1. Data standardisation

To maintain consistency between data sets for their comparability, Edaphobase uses standardised vocabularies for data entry in categorical variables ('authority lists'; see https://service.edaphobase.org/Edaphobase_datafields for details). Where possible, established standards are used (e.g., soil data fields based on FAO/WRB vocabularies, harmonised with international databases such as ISRIC and LUCAS and following INSPIRE metadata guidelines; land use and habitat types following CORINE and European EUNIS programs, ISO standards for country and regional names. See www.edaphobase.org for further details). However, in many cases (for instance, microhabitat types, field and laboratory methods, etc.), lists are created based on the expertise and consensus among soil biologists and their common usage in existing literature. The standardised data-entry possibilities are either simple lists or hierarchical where appropriate. New data fields or standardised data-entry nomenclatures not yet included in Edaphobase can be suggested by external data providers and added if determined appropriate and useful by the Edaphobase Steering Committee at Senckenberg. This ensures that data deriving from, e.g., new (future) techniques can also be integrated.

3.4.2. Data upload

All submitted data sets are uploaded to Edaphobase via specific software ('Upload Wizard') designed as a Java 8 application written in Eclipse iDE. This data-upload tool is publicly accessible (https://service.edaphobase.org/uploadWizard), does not require any advanced computer skill (not an "installed" program integrated in the operating system; but rather an app loaded to the home computer and therefore functioning cross-system [windows, mac or linux]), and is fully documented (https://service.edaphobase.org/Data_upload_information). The software is open source (© Senckenberg Society for Nature

Table 1
Distribution of data records according to taxonomic major groups and sources contained in Edaphobase (accessed on 11 December 2023). 'Other' includes groups such as Tardigrada, Protura, Diplura, Fungi and Prokaryota, which have been recently added to the database.

Taxonomic group	Data-source type			Total
	Collection	Literature	Raw data	
Diplopoda	55,691	26,879	4569	87,139
Chilopoda	30,082	18,053	2725	50,860
Isopoda	5229	5772	3234	14,235
Lumbricidae	3115	10,544	12,696	26,355
Enchytraeidae	1	4423	8051	12,475
Nematoda	9103	36,718		45,821
Collembola	11,918	75,983	22,170	110,071
Oribatida	17,592	10,227	47,523	75,342
Gamasina	10,195	10,201	98	20,494
Other	5465	1066	2565	9096
Total	148,391	199,866	103,631	451,888

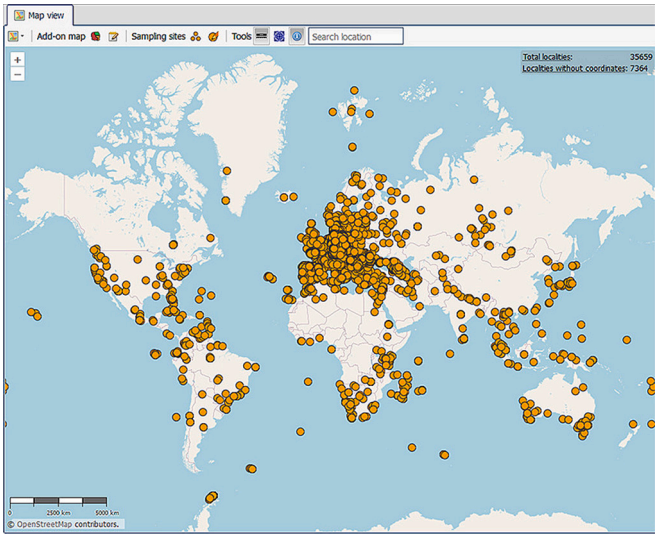


Fig. 4. Geographical distribution of the occurrence locations included in Edaphobase 2.0 (Screenshot from the Edaphobase Query Portal, accessed on 6 December 2023).

Research) and was specifically developed to assist in harmonisation and integration of heterogeneous data sets. A data-upload app was opted over an in-browser solution due to its technical advantages, e.g., higher security (i.e. checking and avoiding compromised files, trojans, etc.), much more and faster automated quality control (see below, Section 3.4.4), advanced control of the User Interfaces (GUIs) and User Experience (UX), etc. Existing data from a data provider (in CSV, Excel or Access format) can be uploaded to the Upload Wizard in its original format and nomenclatures ('as is'), whereby multiple files can be combined to create one data set, and the software guides the semantic annotation of the data fields and data entries of the files ('maps' them) to Edaphobase nomenclatures and formats. While the original files are maintained, for each uploaded observation or record within the entire data set (i.e. of one taxon in one site or sample, at one date), Edaphobase creates an individual data record consisting of multiple data fields and all records and files are combined into one coherent (citable, e.g., via a DOI or the internal data-set ID) data set. During this submission process and subsequently in the Portal database itself after data import, data provider-defined units are transformed to SI units (where possible), taxon quantities given in absolute numbers are transformed to standardised densities (provided that information on sample numbers/replicates, sample sizes/areas, etc. is given), or sample-level data are also aggregated to the site level to allow comparison across heterogeneous data sets. DOIs can be supplied - at the data provider's request (to avoid multiple DOIs in case the dataset has been submitted to other repositories) - for all uploaded data sets, following the DataCite scheme (DataCite Metadata Working Group, 2021). The DOI URL directs directly to the Edaphobase Portal, where a landing page is provided for the specific data set (e.g., https://doi.org/10.26129/491e-nv14, Suppl. Fig. 2). A user manual (https://service.edaphobase.org/Edaphobase_data-upload_manual) as well as an instructional video (https://service.edaphobase.org/Video_Edaphobase-data-upload) are available to guide the upload process and use of the software.

3.4.3. Data templates

While the data-upload software allows the import and integration of digitised data, data templates have also been developed to facilitate data digitisation by all providers and thus expedite comparable data sharing. As spreadsheets are the most widely used way to store data, the data templates have been designed in Excel®. These templates include mandatory data fields for both observational data (i.e. the taxon) and metadata (e.g., name of the dataholder/s, source of data, coordinates of the sampled areas, sampling date, methodologies, etc.) as well as additional recommended data fields for soil-biodiversity assessments. All fields are defined in the template and accompanied by instructions and an example data set is provided. The Soil Fauna Data Template (Fig. 5) was made publicly available in the recent Call for Collaboration (Tsiafouli et al., 2022), which can be found via the link https://drive.google.com/drive/folders/1Om94lMTiZP_Uu-ob1xQfJYcslCxnbkR0 and updated versions via the link https://service.edaphobase.org/Edaphobase_data_templates. This template is focussed on soil animals identified morphologically (file SFDT ['Soil Fauna Data Template']), but has been modified in a further template to allow the inclusion of molecularly identified organisms (invertebrates, fungi, prokaryotes), which is available at the same link (file SSDT ['Soil Sequence-Data Template']). The templates were specifically designed to ease the upload process in Edaphobase via the data-upload software (see the previous section; https://service.edaphobase.org/uploadWizard).

3.4.4. Quality control

Maintaining the highest possible quality standard of data is a key priority of the Edaphobase system, ensuring comparability and therefore advanced data re-use possibilities of all data sets. This is sustained by a multi-step quality control and review process (Fig. 6) during the data set-submission process.

3.4.4.1. Pre-import quality control. The first steps of the quality control are automatically performed by the Upload Software during the initial data-submission process. They check that all variable names as well as species names or vocabularies of categorical data-field entries match Edaphobase standards. This increases comparability and prevents, e.g., typographical errors and spelling mistakes in taxonomic assignments or specific categorical data entries. Numerical data are screened to check if they fit within their possible and plausible ranges. "Possible" ranges exclude numeric data entries that are impossible, i.e. animal abundances below 0, or soil pH values below pH 0 or above 14. "Plausible" ranges are based on available data (or literature) for the variable in question (i.e., Tóth et al., 2016, for heavy metal levels in European soils). Since biological data depend on the organism group and the geographic/climatic region, plausible ranges are currently evaluated according to outliers in available data (in Edaphobase itself or in a submitted data set), whereby the parameter-free Hampel-Test (Dietrich and Schulze, 2014) is used with outliers defined as beyond 5× the median distance from the median (equivalent to ca. 1.5 standard deviations above - for maximally plausible values - the median). Such feedback regarding "plausibility" is given to data providers during data upload, who retain the final decision if such data is a "mistake" or actually true (many organism groups can exceptionally occur in very large population numbers).

To allow data providers to check for data-entry errors before final data upload, the software further creates boxplots of the quantitative data and tests for outliers, visualises frequency of terms for categorical variables ('word clouds') and creates maps of the occurrence sites based on the provided geo-coordinates. After full confirmation of correctness by the data provider, data submission proceeds via automatic upload of the data set to Edaphobase servers.

3.4.4.2. Peri- and post-import quality review. Between initial data-set upload and final import to the database, a manual review of the data is also performed by an international review board of taxon experts ('peri-import quality review'), similar to the peer review process after submitting a paper to a journal. The manual quality review part is undertaken by an international Review Board (similar to an editorial board of a journal) and is thus a community peer effort, currently spread among many reviewers throughout Europe. They carry this out via a standardised checklist, which – as opposed to a journal manuscript review – does not evaluate the scientific content and generation of the

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Source type	Data owner(s)	Taxonomic Major group	*Project* title	*Parent* source	Publication year	Volume number	Page(s)	Upload to GBIF	Upload to other databases	Scope - Species composition level	Scope - Sampling Effort	Scope - Spatial level	Scope - Quantification level	Observation date (Sampling event)	Sampling event code	Sample code	Sampling person/observer	Sampling / observation method
2	Source										Methods								
3	Article	Michailidou, Danai-Eleni; Tsiafouli, Maria	Collembola	Collembolan diversity in olive groves of Greece	Agriculture & Environment	2022	322	455-478	Yes	Yes	Species composition	Sample series	Sample-level data	Quantitative	5/21/2021	C_Olive_20210525	C_Olive_20210525_01	Tsiafouli, Maria	soil corer

Fig. 5. Example of a data template for data digitisation and upload (https://service.edaphobase.org/Edaphobase_data_templates).

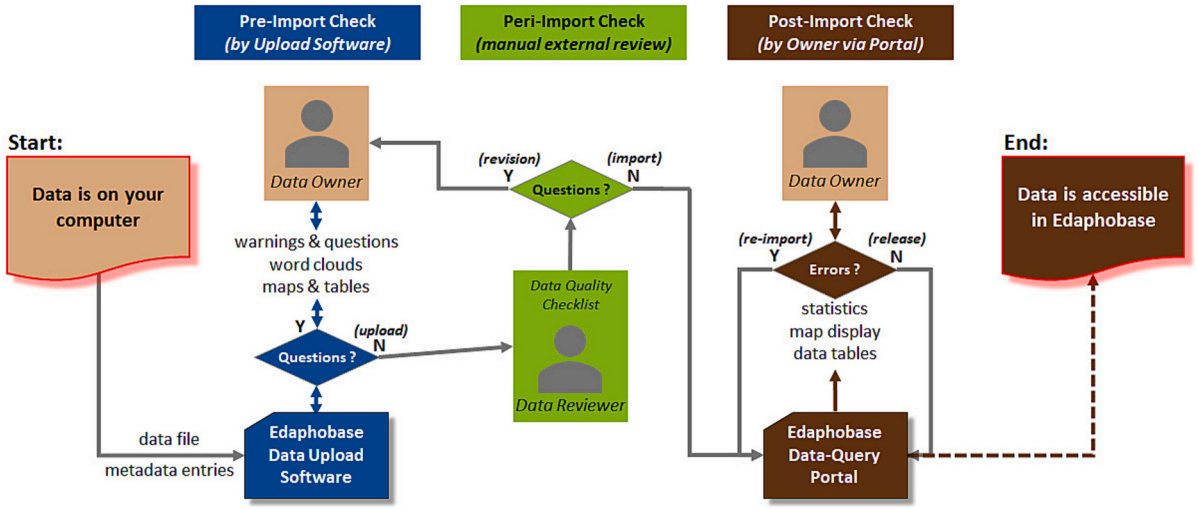


Fig. 6. Overview of the workflow in the quality-review process during data-set upload and final import into Edaphobase. Left: Data upload by the data provider (owner) and automated quality control by the upload software (primarily regarding data standardisation, valid data entries, etc.). Middle: Manual quality review by external experts (primarily regarding data content for integration in the overall Edaphobase dataset and re-use purposes). Right: Final import of a data set into Edaphobase and quality control (primarily regarding correct importation) by the data owner in the online Edaphobase Portal. Blue: procedures and workflow between Data Upload software (Wizard) and data owner; Grey and Green: procedures and workflow between data reviewer and data owner; Brown: procedures and workflow between Data-Query Portal and data owner.

data, but rather ensures that the data is complete (mandatory data), correct (e.g., geo-coordinates, avoiding non-atomized data), consistent (no data schizophrenia) and informative (i.e., can be re-used in further statistical analyses). Questions arising from this manual review are sent by e-mail to the data provider. After the exchange of information and data revision, the revised version of the data set is imported and integrated into the data warehouse. However, before the data is released and can be openly shared, the data provider receives a link to the data (temporary access link) and must confirm the correctness of the data (or request corrections) by going through an automated tutorial-like guide of the data set directly in the online Edaphobase Portal ('post-import quality review').

3.5. Data query and exploration

3.5.1. Data queries

Basic queries can be carried out by users without registration via the online open-access Edaphobase Data-Query Portal (https://portal.edaphobase.org). To hinder data misuse ("robots") and safeguard the rights of data providers, users must register to the system for full access, i.e. to access detailed data, to view specific results in maps, to create customisable tables (i.e. 'flat' data-record tables or contingency 'matrix' tables), to use data-exploration tools, and to download queried-data tables. The log-in is easy, non-restrictive and free of charge. Only one log-in account is needed for all components of the system (i.e., mandatory for the data-upload software, in order to log the specific data provider).

A primary function of the Portal is the query of individual data sets via its DOI, (internal) data-set ID, the data provider's or author's name, project name or article title, etc., and to view them in maps and tables within the online Portal and as well as download for further analyses by the user. An additional strength of the system (representing data warehouse aspects), however, is that data from all data sets can be queried together in a topic-specific manner. Such overall data queries are specified by the user by choosing filters for various topics of interest. Soil-biodiversity data from specific countries, regions or sites (location filter) can be chosen from hierarchical lists or a customised polygon on a map. Taxonomic names, including or excluding synonyms, can be searched at the level of species, genus, family or major soil-organism groups (i.e. Nematoda, Diplopoda, Fungi, etc.; taxon filter). Data can be filtered for one specific taxon or simultaneously for several taxa. Queries can be further refined so that only data from specific sites, specific habitat or land-use types, certain soil parameters, etc. are shown. Additionally, data can be filtered according to management practices (e.g. organic vs conventional, tillage vs no-till, pesticide or fertilizer treatments), provided data providers have included this information in their uploaded data sets (see Section 2.1.3.3 "Environmental and geographical metadata", above).

Furthermore, queries can be further refined for individual countries or be focused on specific time periods, such as sampled years or seasons.

These search options and more can be combined according to the user's choice, selected as displayed data fields (data columns) in output tables or in categorised maps. In the map or table views of the query results, further data fields can be selected and grouped to show biodiversity patterns according to species (groups), habitat types, soil parameters, time periods, etc.

3.5.2. *Data exploration (basic data analyses)*

As an expanded functionality beyond standard data repositories, Edaphobase also offers basic descriptive analyses and visualisations through the 'EdaphoStat' function (Hausen et al., 2017). The tool examines the data from all data sets for a chosen group or taxon and, from this, calculates the ecological niche width and optima for any set of selected habitat parameters. The results are displayed as bar charts of a species' occurrence frequency along a user-defined habitat gradient (e.g., habitat type, soil types, etc.), as regressions of a species' occurrence along a (user defined) quantitative habitat parameter (e.g., soil pH, soil organic matter content, average annual temperature, etc.), or as niche-space diagrams (2D scatterplots) of one or two species' occurrence(s) in relation to two quantitative habitat parameters, etc. (Fig. 7).

A further function within this data-set exploration tool is 'EdaphoClass', which determines the probabilities of site-specific species composition of soil communities based on specified habitat properties. After selecting a set of habitat conditions (such as habitat type and soil pH or organic matter content), this function searches the entire database for all species (of the selected taxonomic major group) recorded under these conditions and builds a histogram of the species occurrence frequencies based on the (selected) environmental conditions where the major group has been recorded (Fig. 8). 'EdaphoStat' and 'EdaphoClass' functions are useful tools for data exploration, also to assist with further statistical analysis of soil biodiversity data after being downloaded by a user.

3.6. *Data download, data use and data sharing (data policy)*

Data download is available for all registered users of Edaphobase. Data can be freely used following Edaphobase terms and conditions and any data-owners' restrictions on public availability (see Methods Section 2.1.1, above). For instance, data providers must be cited when any derived results/analyses are published or made publicly available. Complete data sets or – in the case of queries throughout all data sets – raw data 'flat' tables (i.e., a spreadsheet of observations in rows and all variables [e.g., sources, taxon, site, coordinates, …] in columns) or user-specified contingency ('matrix') tables (e.g., species x sites, whereby all cells contain the same unit [e.g., individuals/m^2]) can be downloaded as CSV or Excel files and saved locally to continue offline data exploration and analyses by the individual user.

The Edaphobase Data Policy (Senckenberg, 2023; https://service.edaphobase.org/Edaphobase_Data_Policy) regulates how data sets and their owners should be cited in publications using Edaphobase data. Besides providing a suggested citation for Edaphobase in general, it suggests recommendations for: (1) offering individual data providers co-authorship, (2) citation suggestions for references as well as (3) acknowledgement sections, depending on the scope and the degree of contribution of specific data sets to the overall analysis. Since data-providers' personal contact information is not made publically available, Edaphobase provides services to data users for contacting specific data providers.

Edaphobase transfers soil-biodiversity data to GBIF, the BonaRes Soil Data Centre and other general biodiversity data infrastructures such as national research-data infrastructures. During the upload procedure (see Results Section 3.4.2), the data provider has the option to allow or restrict such further data sharing, either to all cooperating data infrastructures or only to individual data repositories. This allows data providers to upload data once, alleviating the time-consuming need to upload to multiple repositories (i.e., Edaphobase and GBIF); but also avoids duplicate data uploads (i.e., if uploaded separately to both Edaphobase and GBIF).

4. Discussion

Scattered and non-systematically stored data is a major obstacle towards quantitative assessments of broad-scale geographical patterns of soil biodiversity, estimation of the local and regional drivers of this distribution, as well as the use of this information in land and soil management planning. Edaphobase is to our knowledge the first data repository that provides geographically referenced and quantitative taxonomic data, accompanied by environmental metadata to allow detailed assessments of soil biodiversity distribution and its relation to local environmental variables. By combining data on taxonomy, abundance and distribution of soil organisms with edaphic, climate and other environmental conditions of the sampled locations, it provides a unique platform for sharing, re-using and synthesising soil-biodiversity data for multiple scopes in research, education, management and policy. Edaphobase is currently the most exhaustive data source for soil-biodiversity, as it includes not only published data, but also data

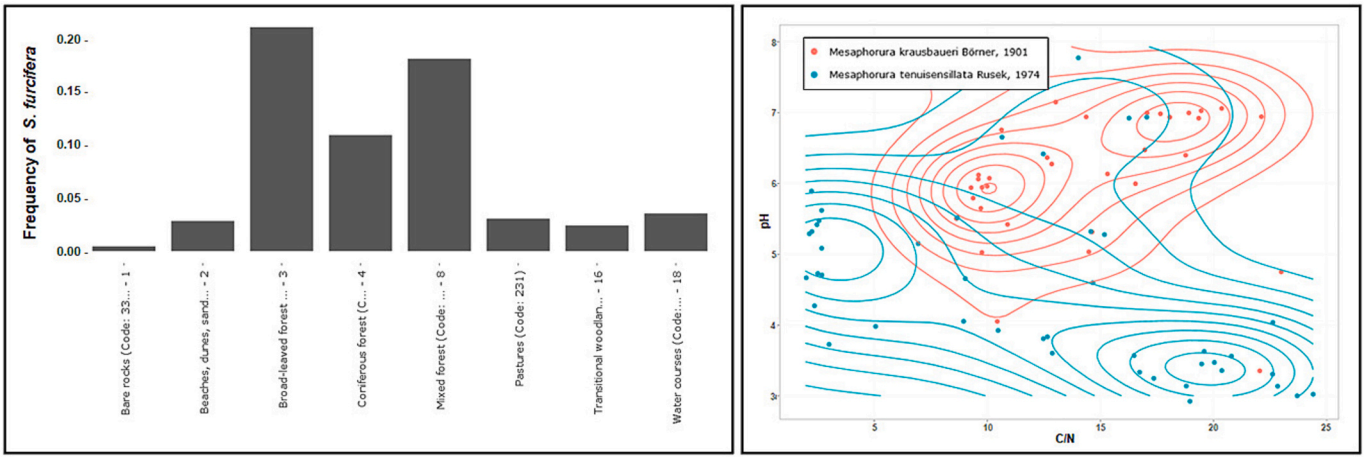


Fig. 7. Examples of visualisations that can be created by the EdaphoStat data-analysis tool available in the Edaphobase Data-Query Portal. Left: Column chart of the occurrence frequency of *Supraphorura furcifera* (Collembola) across EUNIS habitat-type classifications (showing its preference for woodland sites). Right: Niche-space diagram of the occurrences of *Mesophorura krausbaueri* and *M. tenuisensillata* (Collembola) in relation to soil C/N ratio and soil pH, showing potential niche partitioning between the two species.

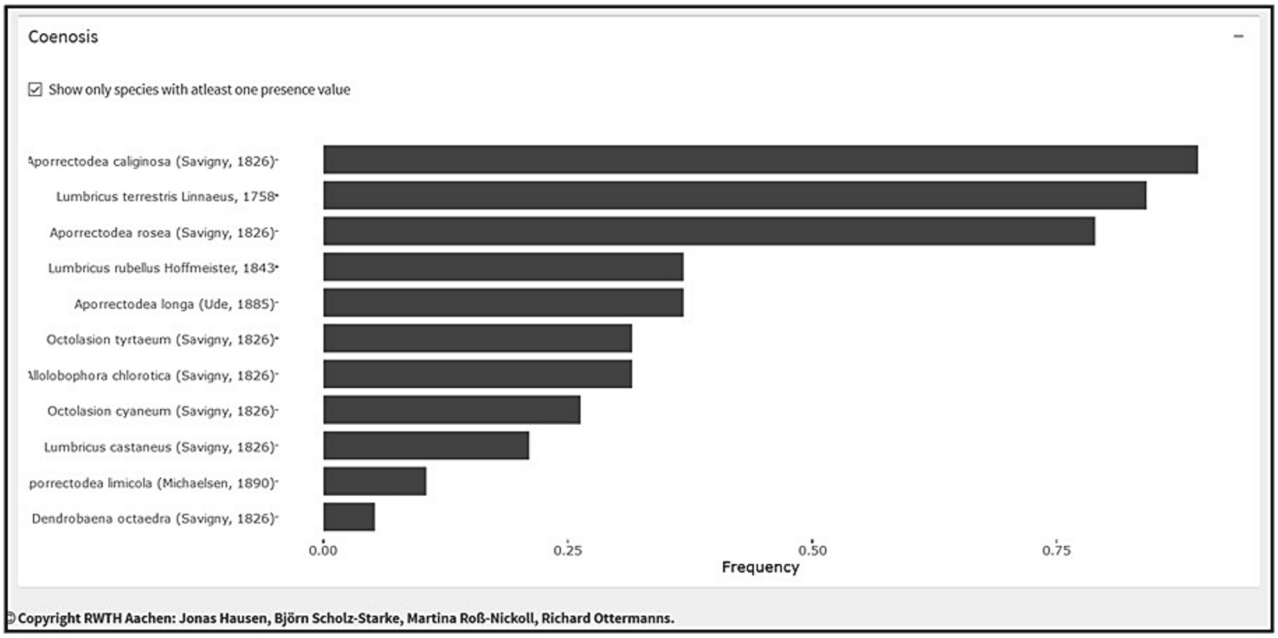


Fig. 8. Example of output from the EdaphoClass function of the Data-Query Portal. Bar chart showing the occurrence frequencies of earthworm species collected in arable-field sites with a soil C/N ratio ranging between 8 and 14 and a soil pH between 5.2 and 6.5.

contained in unpublished museum collections, technical and project reports, PhD and MSc theses, etc. of diverse data providers, which is often difficult to find and access.

Integrating and harmonising such heterogeneous data sets remains a major challenge. One approach to this is to achieve consensus among both data providers and data users. To this end, recent Edaphobase developments are based on a collective effort of a broad group of ca. 100 soil ecology experts from >30 countries in and around Europe working within the framework of the EU COST Action 'EUdaphobase' and building on an earlier version of Edaphobase (Burkhardt et al., 2014). This broad consortium enabled exhaustive expertise coverage, through reviewing, amending and expanding the available possibilities for data set entries, standardising vocabularies and their definitions to make them understandable for a wide range of soil ecologists. Thereby, many international standards are used (the inclusion of more global standards is still needed and currently planned). Edaphobase therefore strives to collate, integrate and make heterogeneous data sets on soil-biodiversity available to scientists, stakeholders and the public in a harmonised manner. The resultant >650 data fields allow the integration of soil biodiversity related data for the majority of repository and classification formats and needs. This complexity can, however, be overwhelming for most researchers and data users. Therefore, the user interfaces throughout the Edaphobase system organises the data according to hierarchical thematic categories to facilitate data filtering. Furthermore, the suggested templates and guidelines for data digitisation help to reduce this complexity. Finally, publicly available recorded webinars (https://service.edaphobase.org/Edaphobase_webinars) explain the data structures and how to use the Edaphobase system.

Another challenge in achieving data harmonisation and integration is technical. Despite the consensus in building integrative data sets, individual researchers and projects follow their own vocabularies and data structures during data collection. To transform these to standardised vocabularies and data structures is not only time consuming, but highly error-prone. Therefore, within the EUdaphobase COST Action framework, the data upload software was developed and tested, which annotates ('maps') individual, previously digitised data sets to Edaphobase standards and transforms these data sets to Edaphobase data structures to allow the best possible integration (see Results Section 3.4.2). Nonetheless, complexly linking soil-biodiversity data to precise spatial, geographical and environmental metadata is challenging for data providers. Therefore, the upload software is constantly being further developed, intending, e.g., to collect user-specific terminologies to create thesauri, where machine-learning techniques will help further automate such semantic-annotation procedures, further reducing the effort needed by data providers.

A third Edaphobase tier of data harmonisation and integration is the multi-step quality control procedures that all submitted data sets undergo in Edaphobase 2.0 (see Results Section 3.4.4). While the data-set upload software performs a control check for correct taxonomy, standardised data structures and vocabularies, and offers data providers procedures for correcting possible erroneous data, manual quality control after data upload confirms that the data set is appropriate for future re-use. As this is time consuming, an international review board, analogous to an editorial board of a scientific journal, has been established. Members are not only taxon specialists, but also understand statistical assessment procedures in order to ensure the highest possible data re-use. Finally, data providers review their uploaded data sets in Edaphobase itself, to ensure that all harmonisation and integration steps are correct. While the Edaphobase quality control does not score the *scientific* quality of a dataset, provision of methodological metadata (i.e., regarding fieldwork and laboratory methods, molecular pipelines, morphological identification procedures, or the content scope of the dataset) allows a *data user* to assess the suitability of individual datasets for his/her assessment needs.

Overall, Edaphobase provides valuable information for all interested in understanding and protecting the importance of soil biodiversity in terrestrial ecosystems. A large variety of individuals and organisations will benefit from such a domain-specific data repository. For instance, soil biologists, ecologists and researchers can use Edaphobase to access information on soil biota, their distribution, and ecological correlations. Environmental scientists and consultants can use the system to support environmental impact assessments and soil conservation/management planning. Agricultural and horticultural practitioners can gain support in understanding the role of soil biodiversity in soil health and fertility and to make informed decisions about their management practices. Land managers and policy makers can use Edaphobase to make informed decisions on land use and management practices, including conservation and restoration of soil biodiversity. Researchers and students can use the

system as a resource for their studies and research projects, including data collection, comparison and analysis.

Although the ready-to-use Edaphobase platform is publicly available, further effort is still required in multiple dimensions. To ensure a better representation of soil biodiversity at larger spatial and temporal scales, an international multidisciplinary cooperation network, constantly collating widespread data from multiple soil ecology disciplines is needed. For instance, there are large gaps in probably existing data from eastern Europe, while much existing data from western Europe has not yet been made publicly available. The COST Action EUdaphobase has worked intensively to fill these data gaps (cf. Tsiafouli et al., 2022; https://www.cost.eu/cost-events/the-european-atlas-of-soil-fauna/). A number of EU-funded research projects (i.e. MINOTAUR [https://ejpsoil.eu/soil-research/minotaur], eco2adapt [https://www.eco2adapt.eu/], MicroEco [https://www.biodiversa.eu/2023/04/19/microeco/], INTACT [https://intactproject.eu/]) on soil biodiversity as well as national soil monitoring programs (i.e., the developing German National Soil Monitoring Centre) are currently considering using Edaphobase as a data repository in their data management plans, which will significantly enlarge the database's spatial coverage at least within the European territory. Furthermore, the recent paradigm change that researchers' routine tasks also include public sharing of data generated in research and monitoring programs must continue (e.g., Osawa, 2019; Sim et al., 2020; Tenopir et al., 2020; Tedersoo et al., 2021). This is a non-trivial activity and requires not only effort, but a better understanding of data and data structures, including the provision of methodological and environmental metadata linked to biodiversity data. Since this had not been required in the past Edaphobase version, currently only approximately 40 % of Edaphobase data includes such metadata. Another future perspective for data collation is using artificial intelligence and text mining tools for screening earlier publications and enhancing Edaphobase with data on soil biodiversity from papers, reports and books. Finally, and related to data sharing, incentives for publicly providing research data are imperative. Motivation factors for sharing data have been identified as scientific progress, data sharing policies of funding agencies and publishers, safeguards against scientific fraud, enhanced collaboration and, in particular, increased authorships (e.g., Schmidt et al., 2016; Bierer et al., 2017; Chawinga and Zinn, 2019; Pasquetto et al., 2019). Edaphobase offers basic incentives for sharing data, e.g. making data owners and research projects publicly transparent, offering citable DOIs for individual data sets and the requirement that all publications using Edaphobase data acknowledge or cite the individual data owners, in order to increase the visibility of researchers and their activities. Nonetheless, data sharing remains an abstract concept. Recently, the call for collaboration in producing a European Atlas on Soil Fauna (Tsiafouli et al., 2022) intends to give a concrete example for the value of data sharing. Questionnaires are available for potential data providers (https://service.edaphobase.org/Data_provider_questionnaire) and stakeholders (https://service.edaphobase.org/Stakeholder_survey) to not only assess their potential interests and constraints in data-sharing possibilities, but especially to determine their data and information needs in order to further develop the Edaphobase soil biodiversity data warehouse.

To increase the usefulness of the data infrastructure for overall soil biodiversity assessments, the system has been expanded within the EUdaphobase COST Action framework from a ‘pure’ soil invertebrate animal database to include molecularly generated data as well as data sets on soil microorganisms. For instance, the system now includes necessary data-entry options for molecular data (i.e. molecular taxon-identification methods as opposed to morphological methods, including metadata on bioinformatic pipelines). While Edaphobase does not intend to host sequences themselves (which would duplicate successful sequence-data infrastructures, i.e. NCBI, BOLD), taxa identified via sequencing are annotated with sequence accession numbers deposited elsewhere. Most recently, Edaphobase has been also expanded to allow upload of data on fungi, fungi related organisms and prokaryotes. Permanent linkages (‘APIs’: Application Programming Interfaces) with more taxonomically oriented international databases (i.e. MycoBank [Robert et al., 2013] and Unite [Nilsson et al., 2018] for Fungi, the Bac*Dive* system [Reimer et al., 2022] for prokaryotes) ensures a sustainable taxonomic ‘backbone’ for these soil-organism groups. Furthermore, data entry possibilities for soil microbial summary parameters (i.e., soil respiration, microbial biomass, metabolic quotient, etc.) have been introduced in Edaphobase. Such data-set uploads as well as their data re-use within Edaphobase are currently being tested for large-scale regions within Europe.

A further ongoing development involves the representation of *functional* soil biodiversity within the data infrastructure. The current approach is to integrate morphological, behavioural and physiological trait data and to link these with corresponding taxa, in order to aggregate taxon-specific observational data to ecological or functional groups, and then relate these to site characteristics (as is currently done for taxa data). Also within the EUdaphobase COST Action framework, ongoing work includes collating trait data for, i.e., earthworms, collembola, nematodes and protozoa. Recent work developing a soil-faunal trophic trait ontology (Le Guillarme et al., 2023) as well as a tool for developing data knowledge graphs linking trophic traits to taxa (Le Guillarme and Thuiller, 2023) is a promising direction. Current Edaphobase activities are developing APIs to permanently connect soil-faunal trait databases such as BETSI (https://portail.betsi.cnrs.fr/) and EcoTaxonomy (http://ecotaxonomy.org/), as external data layers to Edaphobase, linking via soil fauna taxonomies. Considering that this functionality is estimated to require ca. 6–7 person-months to technically develop and ca. 2.5 full-time staff are available from own resources (see Section 2.3 “Sustainability” above), this will be available within the next 4–5 months. To include traits for further soil-organism groups, future work will intensify the linkage to the Bac*Dive* system, which includes prokaryotic functional characteristics; and the FungalTrait database (Põlme et al., 2020) is also a promising potential partner. The goal is for users to be able to query taxonomical and trait data associated with other biological data (e.g., which species and traits are associated with a particular edaphic community), or query trait-associated data linked to a combination of non-biological data (e.g., which ecological or functional groups are expected to be found under specific environmental conditions or be influenced by certain land- or soil-use methods).

These activities highlight the potential of various soil biodiversity-related data repositories as data layers in a widely distributed data network. The developed software linkages to the fungal and bacterial databases as well as the mentioned trait databases exemplify how multiple data infrastructures can be combined in a domain-specific manner to increase overall data usage and assessment for further and more complex research and assessment questions. Further possibilities for a networked data platform could include linkages between other external data sources to augment gaps in environmental metadata missing from submitted data sets (e.g., ISRIC or LUCAS for European soil data), or with national monitoring or specialised initiatives (e.g., Soil BON, NETSOB/GLOSOB) to increase overall soil biodiversity data coverage.

Linking Edaphobase with other databases requires high levels of interoperability (cf., Edwards et al., 2000; Marenco et al., 2007; Berendsohn et al., 2011). The controlled vocabularies mentioned above, which include numerous international standards, is a basic prerequisite. These are all well documented (https://service.edaphobase.org/Edaphobase_datafields) and resolvable via unique and persistent identifiers (as are all datasets, taxa and sites). As Edaphobase exchanges data to GBIF and elsewhere via ABCD and DarwinCore (Wieczorek et al., 2012; TDWG, 2014) procedures, web application wrappers (via XML formats) are already in place and Edaphobase data fields therefore annotated with the corresponding DarwinCore equivalents (where applicable). The programming interfaces mentioned above allow data exchange between database layers, and are being continually developed further by the Edaphobase team. Furthermore, and within the EUdaphobase COST

Action framework, formal ontologies for all data structures are currently being developed (Aldana-Martín et al., in prep.) to explicitly enable this connection with other databases. Based on these ontologies, semantic web services are being developed to link Edaphobase in research-data infrastructures via middleware software services.

While Edaphobase has several strengths as a data storage and sharing platform, it also faces some challenges that need to be addressed to improve its usefulness and impact., such as (1) limited coverage: activities are currently focused on Europe and has limited information from other regions, which limits its usefulness for global comparisons and studies; (2) data incompleteness: it still lacks data on some soil fauna groups (e.g., protozoa, prostigmatid mites) and regions (i.e. south-eastern Europe, Scandinavia), which limits its ability to provide a comprehensive picture of soil biodiversity as a whole for all regions; further, many data sets do not or only patchally include metadata on methods and environmental parameters.

Despite such challenges, Edaphobase data has been used numerous times (see Figs. 9 & 10 for Edaphobase access and usage) for fundamental and applied studies. Already in previous development stages, the system was used, i.e., for descriptions of soil-faunal distribution and environmental correlations at a national scale (Jänsch et al., 2013; Römbke et al., 2013) or in broad-scale analyses of anthropogenic impacts on biodiversity (Bowler et al., 2017). More recently, to name just a few examples, Edaphobase data has been used in analyses of the global distribution of soil biota (Phillips et al., 2019; van den Hoogen et al., 2019; Potapov et al., 2023), in studies of the effects of environmental or management changes on soil fauna (Russell and Gergocs, 2019; Pižl et al., 2023), to develop red lists of soil biota (Lehmitz et al., 2016; Reip et al., 2016; Phillips et al., 2017), to quantify reference conditions for monitoring programs (Jenssen et al., 2021; Salako et al., 2023), to develop methods of species abundance estimates (Gotelli et al., 2023), and tools for ecological network analyses (Marzidovsek et al., 2022), among many others.

In summary, the Edaphobase platform for soil-biodiversity data has great potential to be a valuable solution for anyone involved in soil biodiversity research, monitoring, and conservation, as well as decision- and policy-making and education. Both providing and gaining soil biodiversity data and information to and from EUdaphobase is straightforward, user-friendly, and beneficial. The existing community working within the framework of the EUdaphobase COST Action is large enough to maintain and further develop a sound data infrastructure. We hope to transform the EUdaphobase consortium into an overarching platform by inclusion of a worldwide stakeholder community. Edaphobase represents an exceptional tool that can still be discovered by further data holders and data users.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apsoil.2024.105710.

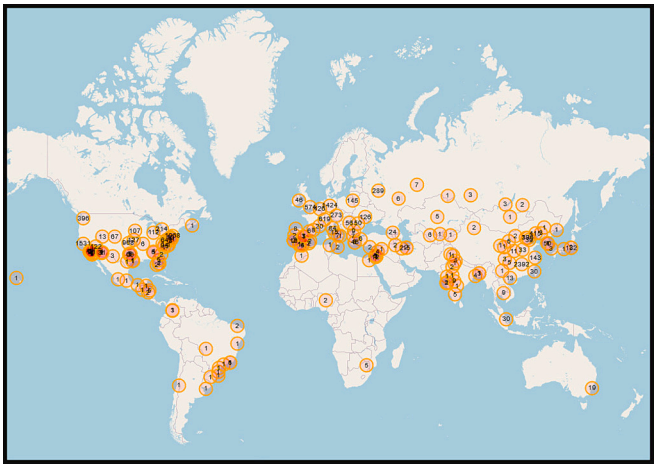


Fig. 9. Number of access 'hits' (=individual IP addresses and do not include multiple access by the same IP) to the Edaphobase Data Query Portal in 2022 (>14,000).

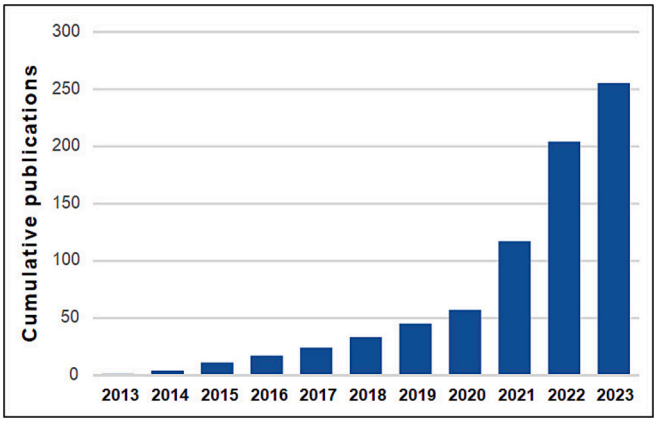


Fig. 10. Cumulative number of publications using or citing Edaphobase, showing exponential growth in the use of Edaphobase in international research. Information based on a Web of Science query using the filter 'Edaphobase'.

CRediT authorship contribution statement

D.J. Russell: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **E. Naudts:** Writing – original draft. **N.A. Soudzilovskaia:** Writing – review & editing, Writing – original draft. **M.J.I. Briones:** Writing – review & editing, Methodology. **M. Çakır:** Writing – review & editing. **E. Conti:** Writing – review & editing. **J. Cortet:** Writing – review & editing, Conceptualization. **C. Fiera:** Writing – review & editing. **D. Hackenberger Kutuzovic:** Writing – review & editing, Conceptualization. **M. Hedde:** Writing – review & editing, Conceptualization. **K. Hohberg:** Writing – review & editing. **D. Indjic:** Writing – review & editing. **P.H. Krogh:** Writing – review & editing. **R. Lehmitz:** Writing – review & editing, Conceptualization. **S. Lesch:** Writing – review & editing, Software. **Z. Marjanovic:** Writing – review & editing. **C. Mulder:** Writing – review & editing. **L. Mumladze:** Writing – review & editing. **M. Murvanidze:** Writing – review & editing. **S. Rick:** Writing – review & editing, Software. **M. Roß-Nickoll:** Writing – review & editing, Methodology, Conceptualization. **J. Schlaghamerský:** Writing – review & editing, Conceptualization. **O. Schmidt:** Writing – review & editing. **O. Shelef:** Writing – review & editing. **M. Suhadolc:** Writing – review & editing. **M. Tsiafouli:** Writing – review & editing, Conceptualization. **A. Winding:** Writing – review & editing. **A. Zaytsev:** Writing – review & editing. **A. Potapov:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Edaphobase was developed with the help of multiple partners throughout its lifetime (see https://www.edaphobase.org for details). Initial development of Edaphobase was funded by the German Ministry for Education and Research (BMBF) under the support codes 01LI0901A (2009–2013) and 01LI1301A (2014–2018). From October 2019 to March 2024, further development of Edaphobase 2.0 took place within the framework of the EU COST Action 'EUdaphobase - European Soil-Biology Data Warehouse for Soil Protection' (CA18237), supported by

the COST Association (European Cooperation in Science and Technology and funded by the European Union. We are immensely grateful to all data providers and especially the members of the COST Action EUdaphobase for developmental input to, and testing of, all software components, nomenclatural standardisation, the Edaphobase Data Policy, data re-use procedures and much more.

Data availability

The data presented in the article is publically available via https://portal.edaphobase.org

References

Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services — a global review. Geoderma 262, 101–111. https://doi.org/10.1016/j.geoderma.2015.08.009.

Ahyong, S., et al., . World Register of Marine Species. https://www.marinespecies.org. https://doi.org/10.14284/170.

Alter, G.C., Vardigan, M., 2015. Addressing global data sharing challenges. J. Empir. Res. Hum. Res. Ethics 10 (3), 317–323. https://doi.org/10.1177/1556264615591561.

Bellinger, P.F., Christiansen, K.A., Janssens, F., 1996ff. Checklist of the Collembola of the world. http://www.collembola.org.

Berendsohn, W.G., Güntsch, A., Hoffmann, N., Kohlbecker, A., Luther, K., Müller, A., 2011. Biodiversity information platforms: from standards to interoperability. ZooKeys 150, 71–87. https://doi.org/10.3897/zookeys.150.2166.

Bierer, B.E., Crosas, M., Pierce, H.H., 2017. Data authorship as an incentive to data sharing. New Engl. J. Med. 376, 1684–1687. https://doi.org/10.1056/NEJMsb1616595.

Bowler, D., Hof, C., Haase, P., Krönke, I., Schweiger, O., Aderian, E.R., Baeert, L., Bauer, H.-G., Blick, T., et al., 2017. Cross-realm assessment of climate change impacts on species' abundance trends. Nat. Ecol. Evol. 1, 0067. https://doi.org/10.1038/s41559-016-0067.

Boyko, C.B., et al., . World Marine, Freshwater and Terrestrial Isopod Crustaceans Database. https://www.marinespecies.org/isopoda. https://doi.org/10.14284/365 (Accessed on 2022-06-27).

Bray, I.K., 2002. An Introduction to Requirements Engineering. Addison-Wesley, Boston (408 pp.).

Burkhardt, U., Russell, D.J., Decker, P., Döhler, M., Höfer, H., Lesch, S., Rick, S., Römbke, J., Trog, C., Vorwald, J., Wurst, E., Xylander, W.E.R., 2014. The Edaphobase project of GBIF-Germany — a new online soil-zoological data warehouse. Appl. Soil Ecol. 83, 3–12. https://doi.org/10.1016/j.apsoil.2014.03.021.

Castro-Huerta, R.A., Falco, L.B., Sandler, R.V., Coviella, C.E., 2015. Differential contribution of soil biota groups to plant litter decomposition as mediated by soil use. PeerJ 3, e826. https://doi.org/10.7717/peerj.826.

Chawinga, W.D., Zinn, S., 2019. Global perspectives of research data sharing: a systematic literature review. Libr. Inform. Sci. Res. 41 (2), 109–122. https://doi.org/10.1016/j.lisr.2019.04.004.

Coleman, D.C., Foissner, W., Paoletti, M.G., 1993. Soil Biota, Nutrient Cycling and Farming Systems. Lewis Publishers, Boca Raton.

Coleman, D., Callaham, M., Crossley Jr., D., 2017. Fundamentals of Soil Ecology, 3rd ed. Academic Press, London.

Csuzdi, C., Zicsi, A., 2003. Earthworms of Hungary (Annelida: Oligochaeta, Lumbricidae). https://doi.org/10.5281/zenodo.4309820.

DataCite Metadata Working Group, 2021. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V. https://doi.org/10.14454/3w3z-sa82

de Jong, Y., Verbeek, M., Michelsen, V., de Place Bjorn, P., Los, W., Steeman, F., Bailly, N., Vasiere, C., et al., 2014. Fauna Europaea - all European animal species on the web. Biodivers. Data. J. 2, e4034. https://doi.org/10.3897/BDJ.2.e4034.

Degma, P., Guidetti, R., 2007. Notes to the current checklist of Tardigrada. Zootaxa 1579, 41–53. https://doi.org/10.11646/zootaxa.1579.1.2.

Dick, J., Hull, E., Jackson, K., 2017. Requirements Engineering. Springer, Cham, Switzerland (239 pp.).

Dietrich, E., Schulze, A., 2014. Statistische Verfahren zur Maschinen- und Prozessqualifikation, 7th ed. Hanser Verlag, Munich Vienna. (779 pp. ISBN: 978-3-446-44055-5).

DriloBase Project, 2014ff. The world earthworm database. http://drilobase.org.

Dunbar, M.B., Panagos, P., Montanarella, L., 2013. European perspective of ecosystem services and related policies. Integr. Environ. Assess. Manag. 9, 231–236. https://doi.org/10.1002/ieam.1400.

Edwards, J.L., Lane, M.A., Nielsen, E.S., 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289, 2312–2314. https://doi.org/10.1126/science.289.5488.2312.

Emmerling, C., Schloter, M., Hartmann, A., Kandeler, E., 2002. Functional diversity of soil organisms — a review of recent research activities in Germany. J. Plant Nutr. Soil Sci. 165 (4), 408–420. https://doi.org/10.1002/1522-2624(200208)165:4%3C408::AID-JPLN408%3E3. 0.CO;2-3.

EU Commission, 2021. EU Soil Strategy for 2030- Reaping the benefits of healthy soils for people, food, nature and climate. In: Com(2021)699. European Commission, Brussels (25 pp.).

Fan, K., Chu, H., Eldridge, D.J., Gaitan, J.J., Liu, Y.-R., Sokoya, B., Wang, J.T., Hu, H.-W., He, J.-Z., Sun, W., Cui, H., Alfaro, F.D., Abades, S., Bastida, F., Díaz-López, M., Bamigboye, A.R., Berdugo, M., Blanco-Pastor, J.L., Grebenc, T., Duran, J., Illán, J.G., Makhalanyane, T.P., Mukherjee, A., Nahberger, T.U., Peñaloza-Bojacá, G.F., Plaza, C., Verma, J.P., Rey, A., Rodríguez, A., Siebe, C., Teixido, A.l., Trivedi, P., Wang, L., Wang, J., Yang, T., Zhou, X.-Q., Zhou, X., Zaady, E., Tedersoo, L., Delgado-Baquerizo, M., 2023. Soil biodiversity supports the delivery of multiple ecosystem functions in urban greenspaces. Nat. Ecol. Evol. 7, 113–126. https://doi.org/10.1038/s41559-022-01935-4.

FAO, 2022. Global Status of Black Soils. Rome. https://doi.org/10.4060/cc3124en (200 pp.).

FAO, ITPS, 2015. Status of the World's Soil Resources (SWSR) – Main Report. Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils, Rome, Italy.

FAO, ITPS, GSBI, CBD, EC, 2020. State of knowledge of soil biodiversity - status, challenges and potentialities. In: Report 2020. FAO, Rome. https://doi.org/10.4060/cb1928en.

Ferrier, S., Manion, G., Elith, J., Richardson, K., 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Divers. Distrib. 13, 252–264. https://doi.org/10.1111/j.1472-4642.2007.00341.x.

Frouz, J., Elhottová, D., Kuráž, V., Šourková, M., 2006. Effects of soil macrofauna on other soil biota and soil formation in reclaimed and unreclaimed post mining sites: results of a field microcosm experiment. Appl. Soil Ecol. 33 (3), 308–320. https://doi.org/10.1016/j.apsoil.2005.11.001.

Gardi, C., Jeffery, S., 2009. Soil biodiversity. In: JRC Scientific and Technical Report 50304, Luxenbourg. https://doi.org/10.2788/7831.

GBIF, 2020. The Global Biodiversity Information Facility. https://www.gbif.org/what-is-gbif.

Gewin, V., 2015. An open mind on open data. Nature 529, 117–118. https://doi.org/10.1038/nj7584-117a.

Ghilarov, M.S., Krivolutsky, D.A., 1975. Identification Keys of Soil Inhabiting Mites, Sarcoptiformes. Nauka, Moscow (491 pp., in Russian).

Gotelli, N.J., Booher, D.B., Urban, M.C., Ulrich, W., Suarez, A.Y., Skelly, D.K., Russell, D. J., Rowe, R.J., Rothendler, M., Rios, N., et al., 2023. Estimating species relative abundances from museum records: quantitative validation with field data of animal and plant assemblages. Methods Ecol. Evol. 14, 431–443. https://doi.org/10.1111/2041-210X.13705.

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: With Applications in R. Cambridge Univ. Press, Cambridge (478 pp.).

Hausen, J., Scholz-starke, B., Burkhardt, U., Lesch, S., Rick, S., Russell, D., Roß-Nickoll, M., Ottermanns, R., 2017. Edaphostat: interactive ecological analysis of soil organism occurrences and preferences from the Edaphobase data warehouse. Database 2017, bax080. https://doi.org/10.1093/database/bax080.

Hedde, M., Blight, O., Briones, M.J.I., Bonfonti, J., Braumean, A., Brondani, M., Sanau, I. C., Clause, J., Conti, E., et al., 2022. A common framework for developing robust soil fauna classifications. Geoderma 426, 116073. https://doi.org/10.1016/j.geoderma.2022.116073.

Inmon, W., 2005. Building the Data Warehouse, 4th ed. John Wiley and Sons, Indianapolis. (576 pp.).

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., et al., 2020. FAIR principles: interpretations and implementation considerations. Data Intelligence 2 (1–2), 10–29. https://doi.org/10.1162/dint_r_00024.

Jänsch, S., Steffens, L., Höfer, F., Roß-Nickoll, M., Russell, D., Burkhardt, U., Toschki, A., Römbke, J., 2013. State of knowledge of earthworm communities in German soils as a basis for biological soil quality assessment. Soil Org. 85 (3), 215–233. https://soil-organisms.org/index.php/SO/article/view/369.

Jeffery, S., Gardi, C., Jones, A., Montanarella, L., Marmo, L., Miko, L., Ritz, K., Peres, G., Römbke, J., van der Putten, W.H. (Eds.), 2010. European Atlas of Soil Biodiversity. European Commission, Publications Office of the European Union, Luxembourg.

Jenssen, M., Nickel, S., Schütze, G., Schröder, W., 2021. Reference states of forest ecosystem types and feasibility of biocenotic indication of ecological soil condition as part of ecosystem integrity and services assessment. Environ. Sci. Eur. 33, 18. https://doi.org/10.1186/s12302-021-00458-2.

Kattge, J., Bönisch, G., Diaz, S., Lavorel, S., Prentice, I.C., Leadley, P., Tautenhahn, S., Werner, G.D.A., Aakala, T., Abedi, M., et al., 2020. TRY plant trait database–enhanced coverage and open access. Glob. Change Biol. 26 (1), 119–188. https://doi.org/10.1111/gcb.14904.

Kimball, R., Ross, M., 2002. The Data Warehouse Toolkit, 2nd ed. John Wiley and Sons, New York. (464 pp.).

Le Guillarme, N., Thuiller, W., 2023. A practical approach to constructing a knowledge graph for soil ecological research. Eur. J. Soil Biol. 117, 103497. https://doi.org/10.1016/j.ejsobi.2023.103497.

Le Guillarme, N., Hedde, M., Potapov, A.M., Martizen-Munez, C.A., Berg, M.P., Briones, M.I.J., Calderon-Sanau, I., Degrune, F., Hohberg, K., Almoyna, C.M., Martínez-Muños, C., Pey, B., Russell, D.J., Whuillar, W., 2023. The Soil Food Web Ontology: aligning trophic groups, processes, and resources to harmonise and automatise soil food web reconstructions. Ecol. Inform. 78, 102360. https://doi.org/10.1016/j.ecoinf.2023.102360.

Lehmitz, R., Römbke, J., Graefe, U., Beylich, A., Krück, S., 2016. Rote Liste und Gesamtartenliste der Regenwürmer (Lumbricidae et Coriodrilidae) Deutschlands. In: Gruttke, H., et al. (Eds.), Rote Liste gefährdeter Tiere, Pflanzen und Pilze Deutschlands, Band 4: Wirbellose Tiere (Teil 2). (Bundesamt für Naturschutz, Bonn) Naturschutz und Biologische Vielfalt, vol. 70(4), pp. 565–590.

Li, S., Ding, F., Flury, M., Wang, Z., Xu, L., Li, S., Jones, D.L., Wang, J., 2022. Macro- and microplastic accumulation in soil after 32 years of plastic film mulching. Environ. Pollut. 300, 118945. https://doi.org/10.1016/j.envpol.2022.118945.
Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J.P., Hector, A., Hooper, D.U., Huston, M.A., Raffaelli, D., Schmid, B., Tilman, D., Wardle, D.A., 2001. Biodiversity and ecosystem functioning: current knowledge and future challenges. Science 294 (5543), 804–808. https://doi.org/10.1126/science.1064088.
Marenco, L., Nadkarni, P., Martone, M., Gupta, A., 2007. Interoperability across neuroscience databases. Methods Mol. Biol. 401, 23–36. https://doi.org/10.1007/978-1-59745-520-6_2.
Marzidovsek, M., Podpecan, V., Conti, E., Debeljak, M., Mulder, C., 2022. BEFANA: a tool for biodiversity-ecosystem functioning assessment by network analysis. Ecol. Model. 471, 110065. https://doi.org/10.1016/j.ecolmodel.2022.110065.
Mathews, J., Glante, F., Berger, M., Broll, G., Eser, U., Faensen-Thiebes, A., Feldwisch, N., König, W., Patzel, N., Sommer, R., Xylander, W.E.R., 2020. Soil and biodiversity - demands on politics. Soil Org. 92 (2), 95–98. https://doi.org/10.25674/so92iss2pp95.
Nadkarni, P.M., Marenco, L., Chen, R., Skoufos, E., Shepherd, G., Miller, P., 1999. Organization of heterogeneous scientific data using the EAV/CR representation. J. Am. Med. Inform. Assn. 6 (6), 478–493. https://doi.org/10.1136/jamia.1999.0060478.
Nilsson, R.H., Larsson, K.H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., et al., 2018. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res. 47 (D1), D259–D264. https://doi.org/10.1093/nar/gky1022.
Osawa, T., 2019. Perspectives on biodiversity informatics for ecology. Ecol. Res. 34, 446–456. https://doi.org/10.1111/1440-1703.12023.
Pasquetto, I.V., Borgman, C.L., Wofford, M.F., 2019. Uses and reuses of scientific data: the data creators' advantage. Harvard Data Science Review 1 (2). https://doi.org/10.1162/9960 8f92.fc14bf2d.
Peter, D., Bertolani, R., Guidetti, R., 2019. Actual Checklist of Tardigrada Species. https://doi.org/10.25431/11380_1178608 Accessed on 2022-05-20.
Phillips, H.R.P., Camerton, E.K., Ferlian, O., Türke, M., Winter, M., Eisenhauer, N., 2017. Red list of a black box. Nat. Ecol. Evol. 1, 0103. https://doi.org/10.1038/s41559-017-0103.
Phillips, H.R.P., Guerra, C.A., Bartz, M.L.C., Briones, M.J.I., Brown, G., Crowther, T.W., Ferlian, O., Gongalsky, K.B., van den Hoogen, J., et al., 2019. Global distribution of earthworm diversity. Science 366, 480–485. https://doi.org/10.1126/science.aax4851.
Phillips, H.R.P., Bach, E.M., Baertz, M.L.C., Bennett, J.M., Beugnon, R., Briones, M.J.I., Brown, G.G., Ferlian, O., et al., 2021. Global data on earthworm abundance, biomass, diversity and corresponding environmental properties. Sci. Data 8, 136. https://doi.org/10.1038/s41597-021-00912-z.
Pižl, V., Sterzynska, M., Tajovsky, K., Stary, J., Nicia, P., Zaderozny, P., Bejger, R., 2023. Effects of hydrologic regime changes on a taxonomic and functional trait structure of earthworm communities in mountain wetlands. Biology 12, 482. https://doi.org/10.3390/biology 12030482.
Põlme, S., Abarenkocv, K., Nilsson, R.H., Lindahl, B.D., Clemmensen, K.E., Kauserud, H., Nguyen, N., Kjoller, R., et al., 2020. FungalTraits: a user-friendly traits database of fungi and fungus-like stramenopiles. Fungal Divers. 105, 1–16. https://doi.org/10.1007/s13225-020-00466-2.
Potapov, A.M., Sun, X., Barnes, A.D., Briones, M.J., Brown, G.G., Cameron, E.K., Chang, C.-H., Cortet, J., Eisenhauer, N., Franco, A.L., Fujii, S., Geisen, S., Guerra, C., Gongalsky, K., Haimi, J., Handa, I.T., Janion-Sheepers, C., Karaban, K., Lindo, Z., Mathieu, J., Moreno, M.L., Murvanidze, M., Nielsen, U., Scheu, S., Schmidt, O., Schneider, C., Seeber, J., Tsiafouli, M., Tuma, J., Tiunov, A., Zaytsev, A.S., Ashwood, F., Callaham, M., Wall, D., 2022. Global monitoring of soil animal communities using a common methodology. Soil Org. 94, 55–68. https://doi.org/10.25674/so94iss1id178.
Potapov, A.M., Guerra, C.A., van den Hoogen, J., Babenko, A., Bellini, B.C., Berg, M.P., Chown, S.L., Deharveng, L., et al., 2023. Globally invariant metabolism but density-diversity mismatch in springtails. Nat. Commun. 14, 674. https://doi.org/10.1038/s41467-023-36216-6.
Ramirez, K., Döriing, M., Eisenhauer, N., Carrdi, C., Ladau, J., Leff, J.W., Lentendu, G., Lindo, Z., Rillig, M.C., Russell, D., Scheu, S., St. John, M.G., de Vries, F.T., Wubet, T., van der Putten, W.H., Wall, D.H., 2015. Toward a global platform for linking soil biodiversity data. Frontiers Ecol. Evol. 3, 91. https://doi.org/10.3389/fevo.2015.00091.
Reimer, L.C., Carbasse, J.S., Koblitz, J., Ebeling, C., Adam Podstawka, A., Jörg Overmann, J., 2022. Bac*Dive* in 2022: the knowledge base for standardized bacterial and archaeal data. Nucleic Acids Res. 50 (D1), D741–D746. https://doi.org/10.1093/nar/gkab961.
Reip, H.S., Spelda, J., Voigtlaender, K., Decker, P., Lindner, E.N., 2016. Rote Liste und Gesamtartenliste der Doppelfüßer (Myriapoda: Diplopoda) Deutschlands. In: Gruttke, H., et al. (Eds.), Rote Liste gefährdeter Tiere, Pflanzen und Pilze Deutschlands, Band 4: Wirbellose Tiere (Teil 2). (Bundesamt für Naturschutz, Bonn) Naturschutz und Biologische Vielfalt, vol. 70(4), pp. 301–324.
Riedo, J., Wettstein, F.E., Rösch, A., Herzog, C., Banerjee, S., Büchi, L., Charles, R., Wächter, D., Martin-Laurent, F., Bucheli, T.D., Walder, F., van der Heijden, M.J.A., 2021. Widespread occurrence of pesticides in organically managed agricultural soils — the ghost of a conventional agricultural past? Environ. Sci. Technol. 55 (5), 2919–2928. https://doi.org/10.1021/acs.est.0c06405.
Rillig, M., Lehmann, A., 2020. Microplastic in terrestrial ecosystems. Science 368 (6498), 1430–1431. https://doi.org/10.1126/science.abb5979.
Rillig, M.C., Ryo, M., Lehmann, A., 2021. Classifying human influences on terrestrial ecosystems. Glob. Change Biol. 27, 2273–2278. https://doi.org/10.1111/gcb.15577.
Robert, V., Vu, D., Amor, A.B.H., van de Wiele, N., Brouwer, C., Jabas, B., Szoke, S., Dridi, A., Triki, M., ben Dauod, S., et al., 2013. MycoBank gearing up for new horizons. IMA Fungus 4 (2), 371–379. https://doi.org/10.5598/imafungus.2013.04.02.16.
Römbke, J., Jänsch, S., Höfer, H., Horak, F., Roß-Nickoll, M., Russell, D., Burkhardt, U., Toschki, A., 2013. State of knowledge of enchytraeid communities in German soils as a basis for biological soil quality assessment. Soil Org. 85 (2), 123–146. https://soil-organisms.org/index.php/SO/article/view/378.
Römbke, J., Bernard, J., Martin-Laurent, F., 2018. Standard methods for the assessment of structural and functional diversity of soil organisms: a review. Integr. Environ. Assess. Manage. 14 (4), 463–479. https://doi.org/10.1002/ieam.4046.
Ronchi, S., Salata, S., Arcidiacono, A., Piroli, E., Montanarella, L., 2019. Policy instruments for soil protection among the EU member states: a comparative analysis. Land Use Policy 82, 763–780. https://doi.org/10.1016/j.landusepol.2019.01.017.
Russell, D.J., Gergocs, V., 2019. Forest-management types similarly influence soil collembolan communities throughout regions in Germany – a data bank analysis. Forest Ecol. Manag. 434, 49–62. https://doi.org/10.1016/j.foreco.2018.11.050.
Rutgers, M., Orgiazzi, A., Gardi, C., Römbke, J., Jänsch, S., Keith, A.M., Neilson, R., Boag, B., Schmidt, O., et al., 2016. Mapping earthworm communities in Europe. Appl. Soil Biol. 97, 98–111. https://doi.org/10.1016/j.apsoil.2015.08.015.
Salako, G., Russell, D.J., Stucke, A., Eberhardt, E., 2023. Assessment of multiple model algorithms to predict earthworm geographic distribution range and biodiversity in Germany: implications for soil-monitoring and species-conservation needs. Biodivers. Conserv. 32, 2365–2394. https://doi.org/10.1007/s10531-023-02608-9.
Scheu, S., Schulz, E., 1996. Secondary succession, soil formation and development of a diverse community of oribatids and saprophagous soil macro-invertebrates. Biodivers. Conserv. 5 (2), 235–250.
Schmidt, B., Gemeinholzer, B., Treloar, A., 2016. Open data in global environmental research: the Belmont Forum's open data survey. PloS One 11 (1), e0146695. https://doi.org/10.1371/journal.pone.0146695.
Sechi, V., De Goede, R.G., Rutgers, M., Brussaard, L., Mulder, C., 2018. Functional diversity in nematode communities across terrestrial ecosystems. Basic Appl. Ecol. 30, 76–86. https://doi.org/10.1016/j.baae.2018.05.004.
Senckenberg, 2023. Edaphobase Data Policy: Balancing Intellectual Property Rights and Data (Re-)Usage. https://doi.org/10.26129/bp3m-ra79 (17 pp.).
Shelef, O., Helman, Y., Friedman, A.L.L., Behar, A., Rachmilevitch, S., 2013. Tri-party underground symbiosis between a weevil, bacteria and a desert plant. PLoS One 8 (11), e76588. https://doi.org/10.1371/journal.pone.0076588.
Shelef, O., Hahn, P.G., Pineda, A., Tejesvi, M.V., Martinez-Medina, A., 2019. Progress in understanding the role of belowground interactions in ecological processes. Front. Ecol. Evol. 7, 318. https://doi.org/10.3389/fevo.2019.00318.
Shelef, O., Hahn, P.G., Pineda, A., Tejesvi, M.V., Martinez-Medina, A. (Eds.), 2020. As Above So Below? Below-Ground Interactions in Ecological Processes. Frontiers Media SA. https://doi.org/10.3389/978-2-88963-258-9 (235 pp.).
Sierwald, P., Spelda, J., . MilliBase. https://www.millibase.org. https://doi.org/10.14284/370.
Sim, I., Stebbins, M., Bierer, B.E., Buttre, A.J., Drazen, J., Dzau, V., Hernandes, A.F., Krumholz, H.M., Lo, B., Munos, B., et al., 2020. Time for NIH to lead on data sharing. Science 367, 6484. https://doi.org/10.1126/science.aba4456.
Smith, V.C., Bradford, M.A., 2003. Litter quality impacts on grassland litter decomposition are differently dependent on soil fauna across time. Appl. Soil Ecol. 24 (2), 197–203. https://doi.org/10.1016/S0929-1393(03)00094-5.
Soil threats in Europe. In: Stolte, J., Tesfai, M., Øygarden, L., Kværnø, S., Keizer, J., Verheijen, F., Panagos, P., Ballabio, C., Hessel, R. (Eds.), 2016. Technical Report No. EUR 27607. Joint Research Centre, European Commission. https://doi.org/10.2788/828742.
TDWG (Darwin Core Maintenance Interest Group, Biodiversity Information Standards), 2014. Darwin Core. Zenodo. https://doi.org/10.5281/zenodo.592792.
Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedastze, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., Swepp, T., 2021. Data sharing practices and data availability upon request differ across scientific disciplines. Sci. Data 8, 192. https://doi.org/10.1038/s41597-021-00981-0.
Tenopir, C., Rice, N.M., Allard, S., Baird, L., Borycz, J., Christrian, L., Grant, B., Olendorf, R., Sandusky, R.J., 2020. Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. PloS One 15 (3), e0229003. https://doi.org/10.1371/journal. pone.0229003.
Tóth, G., Herman, T., Da Silva, M.R., Mananarella, L., 2016. Heavy metals in agricultural soils of the European Union with implications for food safety. Environm. Intern. 88, 299–309. https://doi.org/10.1016/j.envint.2015.12.017.
Tsiafouli, M., Cortet, J., Russell, D., 2022. A call for collaboration to create the European Atlas of Soil Fauna. Soil Org. 94 (3), 175–181. https://doi.org/10.25674/so94iss3id307.
Turbé, A., de Toni, A., Benito, P., Lavelle, P., Lavelle, P., Ruiz Camacho, N., van Der Putten, W.H., Labouze, E., Mudgal, S., 2010. Soil biodiversity: functions, threats and tools for policy makers. In: Report No. 07.0307/2008/517444/ETU/B1. Bio Intelligence Service, IRD, and NIOO, Report for European Commission (DG Environment).
Van den Eynden, V., Corti, L., 2014. The importance of managing and sharing research data. In: Corti, L., Van den Eynden, V., Bishop, L., Woollard, M. (Eds.), Managing and Sharing Research Data. Sage, Los Angeles London New Delhi Singapore, pp. 1–32.
van den Hoogen, J., Gesisen, S., Routh, D., Ferris, H., Trtaunspurger, W., Wardle, D.A., de Goede, R.G.M., et al., 2019. Soil nematode abundance and functional group composition at a global scale. Nature 572, 194–198. https://doi.org/10.1038/s41586-019-1418-6.
Vrebos, D., Bampa, F., Creamer, R.E., Gardi, C., Ghaley, B.B., Jones, A., Rutgers, M., Sandén, T., Staes, J., Meire, P., 2017. The impact of policy instruments on soil

multifunctionality in the European Union. Sustainability 9, 407. https://doi.org/10.3390/su9030407.
Wall, D.H., Nielsen, U.N., Six, J., 2015. Soil biodiversity and human health. Nature 528, 69–76. https://doi.org/10.1038/nature15744.
Weigmann, G., 2006. Hornmilben (Oribatida). In: Dahl, F. (Ed.), Die Tierwelt Deutschlands, vol. 76. Goecke & Evers, Keltern (520 pp.).
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin Core: an evolving community-developed biodiversity data standard. PloS One 7 (1), e29715. https://doi.org/10.1371/journal.pone.0029715.
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axtron, M., Baak, A., Blomberg, N., Boiten, J.-W., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3 (1), 160018. https://doi.org/10.1038/sdata.2016. 18.
Zhou, Z., Wang, C., Luo, Y., 2020. Meta-analysis of the impacts of global change factors on soil microbial diversity and functionality. Nat. Commun. 11 (1), 3072. https://doi.org/10.1038/s41467-020-16881-7.