Applied Soil Ecology 83 (2014) 3–12

Contents lists available at ScienceDirect

Applied Soil Ecology

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

The Edaphobase project of GBIF-Germany—A new online soil-zoological data warehouse

U. Burkhardt[a], D.J. Russell[a,*], P. Decker[a], M. Döhler[a], H. Höfer[b], S. Lesch[a], S. Rick[a], J. Römbke[c], C. Trog[a], J. Vorwald[a], E. Wurst[a], W.E.R. Xylander[a]

[a] Senckenberg Museum of Natural History Görlitz, P.O. Box 300154, 02806 Görlitz, Germany
[b] State Museum of Natural History Karlsruhe, Erbprinzenstraße 13, 76133 Karlsruhe, Germany
[c] ECT Oekotoxikologie GmbH, Böttgerstraße 2-14, 65439 Flörsheim am Main, Germany

ARTICLE INFO

Article history:
Received 31 January 2013
Received in revised form 5 March 2014
Accepted 31 March 2014
Available online 14 May 2014

Keywords:
Soil invertebrates
Multi-source datasets
Habitat parameters
Data standardization
Metadata analyses
Distribution patterns

ABSTRACT

Edaphobase is a non-commercial data warehouse on soil organisms (up to now including Chilopoda, Collembola, Diplopoda, Enchytraeidae, Gamasina, Lumbricidae, Nematoda, Oribatida) integrated in the Global Biodiversity Information Facility (GBIF) network. Edaphobase combines data on taxonomy, zoogeography and ecology of these organisms in a comprehensive manner. Data are derived from publications, unpublished results of field studies (theses, reports) and collection data from German museums and research institutions. Data types (=entities) comprise up-to-date taxonomic thesauri, geographical references, ecological indices, soil composition, vegetation, meteorological data, sampling and extraction methods, quantity of collected organisms, identification methods, preparation techniques and behavioural data. At present, the focus is on Germany and neighbouring countries, but data from other European countries can be incorporated in the future. Edaphobase offers a wide range of tools for data inclusion (data-input client, GIS-tool, semi-automatic literature analysis) and data exploration. Simple queries are possible as well as more sophisticated analyses of different data groups. Specific examples of exploration of multi-source datasets are presented to illustrate the potential of the system for detailed analyses (i.e., for the elucidation of species-specific habitat preferences, distribution patterns or environmental influences on population densities) with different soil invertebrate groups (i.e., Diplopoda, Nematoda and Collembola).

1. Introduction

Although much soil-zoological knowledge has been accumulated in the past, detailed information on the distributional area and ecological preferences of the majority of known soil-faunal species is lacking. Acquisition of this information is best achieved through meta-analyses of supraregional datasets containing taxonomical data as well as detailed collection and locality data such as geographical coordinates, quantity of sampled organisms, sampling and extraction methods, preparation techniques and identification methods, soil parameters, vegetation and meteorological data. Such detailed databases hardly exist. The most comprehensive collection data on soil organisms are usually stored locally at scientific collections (i.e., museums or universities) or by collectors themselves. They thus represent rather local and regional inventories of collection material and background habitat data and lack the interconnection to larger databases. More importantly, following completion of publications and reports, detailed data material is at best usually kept locally, i.e., remains unreleased, and with time often becomes lost. As a consequence of this lack of data availability, comparable soil-zoological data from different studies are unavailable or require cumbersome compilation by hand, verification of details, etc. The lack of information at a supraregional level concerning, e.g., ranges of environmental parameter tolerance or habitat preferences hinders an evaluation of species' distributional patterns and mechanisms, and comprehensive analyses of multiple datasets remains laborious, if not impossible.

Some researchers throughout the world compile comprehensive datasets for their respective countries. However, these are often limited to specific taxonomic groups and are usually not readily accessible to or even known by other scientists. In the last decade several national and international initiatives have begun to centralize and standardize data from biodiversity research and make them publically accessible. A prominent

* Corresponding author. Tel.: +49 358147605502.
E-mail address: david.russell@senckenberg.de (D.J. Russell).

http://dx.doi.org/10.1016/j.apsoil.2014.03.021

initiative is the Global Biodiversity Information Facility (GBIF; Edwards et al., 2000). The participating institutions (i.e., research museums, universities and NGOs) contribute collection and observational data to central national-level server systems (="Nodes"). In Europe, Fauna Europaea is an international database of "scientific names and distributions of all . . . European land and fresh-water animals" (http://www.faunaeur.org) coordinated by taxonomic experts. Databases of specific soil-animal groups, e.g., Chilobase (http://chilobase.bio.unipd.it/) or Collembola.org, offer complete taxonomical lists of the worldwide known species, partly with bibliographic references to the original descriptions, determination keys, etc. Other soil animal-oriented databases, such as the German GloMyrIS or French EcoBioSoil, focus on taxonomy or distribution of a limited number of taxa or provide only a small subset of site parameters (Cluzeau et al., 2010; Melzer et al., 2011). Unfortunately, none of these databases collects or offers comprehensive background habitat information and distributional data is often not georeferenced or only available at very large and rough spatial scales. Therefore, they do not provide the data necessary for meta-analyses of detailed species' spatiotemporal distributions or of species' ecological niche spaces. One exception is the USA-based Data Observation Network for Earth (DataONE), which aims to "ensure the preservation, access, use and reuse of multi-scale, multi-discipline, and multi-national science data" (http://www.dataone.org). However, it contains virtually no data on soil organisms.

At a more national level, in Germany, a few GBIF Nodes (e.g., Evertebrata I maintained by the Museum of Natural History Berlin, Evertebrata II by the Bavarian State Collection of Zoology, or Evertebrata III by Senckenberg) have brought together a comprehensive pool of data from several institutions at a central location. This data is made available online via systems like SeSam (http://sesam.senckenberg.de) or SYSTAX (http://www.biologie.uni-ulm.de/systax/) and is regularly harvested by GBIF International. However, many functionally important groups of soil organisms are not included at a broader level. Those German GBIF Nodes collecting soil-animal data (e.g., Collembola, Acari and Myriapoda) mainly include taxonomic information, storage location of type material or presence in museum collections. Information suitable for the characterization of individual species' zoogeography or niche space is rarely included. Thus ecological background information on soil organisms is sparse and a complex evaluation of biodiversity information stored in these GBIF Nodes is hardly possible.

A comprehensive compilation of all data on important soil-animal groups, including species' specific occurrences as well as habitat parameters and allowing meta-analyses at a species or community level, is thus wanting. Here we present Edaphobase as a data warehouse for the exploration of the distribution and ecology of soil organisms. Edaphobase combines existing data on soil organisms as well as the habitat data of sites of species' occurrences provided by different institutions and persons researching soil fauna. Importantly, these datasets represent compilations of German data acquired and extracted from the literature of the last 60 years, museum collections and raw project data.

The development of the Edaphobase data warehouse on soil animals as presented here pursued the following key objectives: (i) to combine collection data from different research institutions for a variety of taxonomic groups and trophic levels, (ii) to link information concerning species' distributions and the corresponding habitat parameters that is widely scattered in the literature, museum collections and project data, currently with a focus on Germany, (iii) to provide research tools for access and analysis of data for detailed ecological and biogeographical investigations, and (iv) to prepare a data basis for the development of tools for predictions of future changes in soil biological communities as a result of, e.g., land use and climate change.

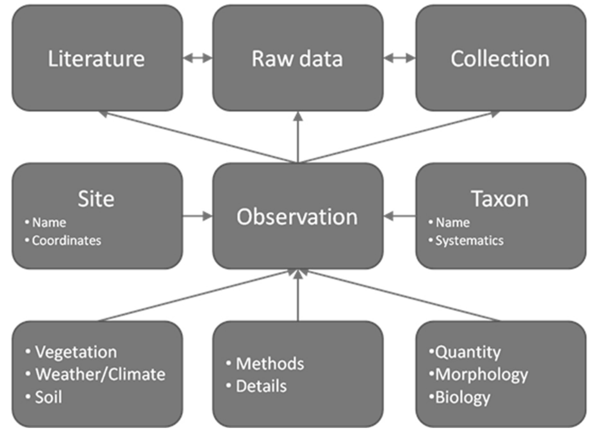

Fig. 1. Simplified data model of the Edaphobase information system.

Within Edaphobase, data are administered and publically accessible for analysis via its web application (http://portal.edaphobase.org). By this focus on soil animals, integration and standardization of data from various sources, time-variance of the data, durability of the system and the web-based availability for meta-analyses, Edaphobase fulfils the criteria of a data warehouse as defined by Kimball and Ross (2002), Inmon (2005), Jones et al. (2006). Here we introduce the structure and contents of the system. Edaphobase offers a wide range of opportunities for arranging and analysing the information contained therein. Specific examples of the exploration of multi-source datasets are presented to illustrate the potential of this data warehouse for detailed analyses on soil organisms when combining taxonomic with ecological data.

2. Methods

2.1. The soil-zoological data warehouse Edaphobase

The data model used for Edaphobase is a combination of a conventional relational model and a modern entity-attribute-value (EAV) concept-based information system (e.g., Hagedorn, 2007), implemented in the database-management system PostgreSQL. The information system is highly modularized. The use of the EAV concept renders the model further flexibility, so that future additions or applications can be programmed without the necessity of major changes in the programme code. A desktop client is used for data input and management, while queries and analyses are performed via a web portal; both modules use the same data basis.

Edaphobase regards an individual dataset as a record of a taxon found in one place at one time by an observer using one method, i.e., a record containing data in different information fields (attributes) giving species-specific "answers" to the question "what, where, when, by whom, how?". Information on the taxon, locality of occurrence, quantity (i.e., presence, or preferably number of individuals or densities of the taxon) and sampling date are regarded as the minimal dataset. Further information fields store data concerning data source (=metadata), soil parameters, vegetation and biotope type of the sites of occurrence, sampling methods, extraction, determination etc. In total, Edaphobase provides about 250 information fields (lists and definitions given at www.edaphobase.org), all of which are highly linked within one dataset (observation or record).

For central management and linkage of these information attributes, the database system is organized into different modules (Fig. 1). The taxon module is the taxonomic core of Edaphobase