VegVault database: linking global paleo-, and neo-vegetation data with functional traits and abiotic drivers

Authors

Ondřej Mottl[1,2], Franka Gaiser[3], Irena Šímová[1,4], Suzette G. A. Flantua[5,6]

Affiliations

1. Center for Theoretical Study, Charles University, Jilská 1, CZ-11000 Prague, Czech Republic
2. Department of Botany, Faculty of Science, Charles University, Benátská 2, CZ-12801 Prague, Czech Republic
3. BayCEER, Bayreuth Centre of Ecology and Environmental Research, University of Bayreuth, Bayreuth, Germany
4. Department of Ecology, Faculty of Science, Charles University, Viničná 7, CZ-12800 Prague, Czech Republic
5. Department of Biological Sciences, University of Bergen, PO Box 7803, N-5020 Bergen, Norway
6. Bjerknes Centre for Climate Research, Bergen, Norway

corresponding author(s): Ondřej Mottl (ondrej.mottl@gmail.com)

Abstract

Understanding the dynamics and persistence of biodiversity patterns over short (present) and long (thousands of years) time scales is crucial for predicting ecosystem changes under global climate and land-use changes. A key challenge is integrating currently scattered ecological data to assess complex vegetation dynamics over time. Here, we present VegVault, an interdisciplinary SQLite database that uniquely integrates paleo- and neo-ecological plot-based vegetation data on a global and millennial scale, directly linking them with functional traits, soil, and climate information. VegVault currently comprises data from *BIEN*, *sPlotOpen*, *TRY*, *Neotoma*, *CHELSA*, and *WoSIS*, providing a comprehensive and ready-to-use resource for researchers across various fields to address questions about past and contemporary biodiversity patterns and their abiotic drivers. To further support usability of the data, VegVault is complemented by the {*vaultkeepr*} R package, enabling streamlined data access, extraction, and manipulation. This study introduces the structure,

content, and diverse applications of VegVault, emphasizing its role in advancing ecological research to improve predictions of biodiversity responses to global climate change.

Background & Summary

Biodiversity and ecosystem dynamics are currently profoundly influenced by global climate change, creating an urgent need for robust predictive models capable of anticipating these impacts. Such models are crucial in guiding conservation efforts during this global crisis. However, many existing models rely on contemporary (modern) ecological data alone, which limits their ability to capture long-term responses of species and ecosystems to environmental change. Incorporating paleoecological data—"ecology of the past" as opposed to neo-ecological, or "ecology of the present"—offers a vital temporal perspective that can enhance our understanding of biodiversity dynamics under shifting climate[1–3].

In recent years, an increasing number of large-scale ecological datasets have become available, providing unprecedented opportunities to investigate ecosystem dynamics across spatial and temporal scales. These datasets encompass contemporary vegetation data (e.g., *Global Biodiversity Information Facility*[4] ('*GBIF*' hereafter), *sPlot*[5], *Botanical Information and Ecology Network*[6] ('*BIEN*' hereafter), paleo-vegetation data (e.g., *Neotoma Paleoecology Database*[7] ('*Neotoma*' hereafter)), plant functional traits (e.g., *TRY plant trait database*[8] ('*TRY*' hereafter)), and abiotic drivers (e.g, *CHELSA*[9], *WorldClim*[10]), which are all critical for understanding the complex interactions that shape ecosystems. Yet, despite the growing availability of such datasets, efforts to integrate these diverse sources into a unified framework remain limited. An integrated database, built with reproducible workflows, is a crucial step toward improving predictive models that estimate biodiversity dynamics under changing climatic conditions. Addressing this gap requires interdisciplinary solutions[11,12] and a commitment to compiling, processing, and linking data in a systematic and transparent manner. Currently, no comprehensive framework exists to link global vegetation data across temporal scales with functional traits and abiotic drivers.

There have been previous attempts to integrate vegetation data across extensive time scales and link them with other data types, such as functional traits and abiotic data. For example, *BioDeepTime*[13] compiles both contemporary and fossil biodiversity time series. However, it lacks coverage of contemporary vegetation plots and, more importantly, omits direct links to functional traits and abiotic data. GBIF[14] includes vegetation biodiversity linked with functional traits but only

holds taxonomic lists for specific regions, without data on relative abundances. Additionally, the latter dataset lacks the long-term dimension of vegetation dynamics over time, derived from the currently existing large databases containing paleoecological records, such as the Neotoma database.

Our database—VegVault addresses these limitations by offering a unique resource that compiles extensive paleo- and neo-ecological vegetation data into a single, accessible SQLite database, accompanied by functional traits and abiotic information (i.e., climate and soil). Its design bridges paleo and contemporary data, enabling the study of long-term processes , and linking them to today's patterns. Thereby, VegVault offers the possibility to specifically analyse the mechanisms of ecosystem response to environmental change which ultimately enhances our ability to predict future biodiversity trajectories. . Specifically, VegVault advances analyses across spatial, temporal, and taxonomic scales in several ways:

- **Contemporary vegetation data**: VegVault incorporates plot-based data from *BIEN*[6] and *sPlotOpen*[15] (open-access version of sPlot), including essential "true negative" information about species' absence. These datasets are pre-cleaned for taxonomic and georeferencing errors, ensuring high reliability for subsequent analyses. The focus on plot-based data rather than occurrence data (e.g., *GBIF*) provides a closer match to paleoecological data formats, facilitating robust comparisons across temporal scales.
- **Paleo-vegetation data**: VegVault employs standardised fossil pollen data processed using the *FOSSILPOL* workflow[16] to integrate records globally from Neotoma. This workflow ensures quality control and creates "ready-for-analysis" data through a reproducible and detailed process for handling large fossil pollen dataset compilations (see **Methods** for details).
- **Interdisciplinary integration**: VegVault combines past and present vegetation data with functional traits and abiotic information to support comprehensive ecological analyses. The abiotic information includes soil data from *World Soil Information Service*[17] ('*WoSIS*' hereafter) and global temperature and precipitation datasets from *CHELSA* and *CHELSA-TRACE21*[18,19], providing essential environmental context. Functional trait data from the *TRY* database and *BIEN* enrich the database by linking species characteristics to ecological roles and responses. This allows researchers to move beyond taxonomic diversity and explore functional dimensions, offering deeper insights into ecosystem processes.

- **Taxonomic classification**: VegVault ensures consistent and standardised taxonomic information by aligning all taxa present in vegetation data—both paleo and contemporary—with the GBIF backbone taxonomy[20]. This process, implemented via the {*taxospace*} R package[21], automates taxonomic classification by matching names to an existing taxonomy based on similarity (e.g., "Ulmus Type" will be classified as genus *Ulmus*). While original taxonomic names are preserved within the database, users can access data at varying taxonomic resolutions (e.g., species, genus, family; if available) to suit specific analytical needs. The same standardisation approach is applied to vegetation data and trait data, facilitating seamless integration (e.g., enabling to calculate genus-level average trait values, when species-level identification is unavailable).
- **Accessibility**: VegVault is designed with user accessibility in mind. The database is supported by the {*vaultkeepr*} R package[22], which simplifies the process of data extraction from the database for project-specific needs. It ensures that researchers can efficiently retrieve and subsequently manipulate data in a reproducible and transparent manner.

The comprehensive nature of VegVault enables diverse applications across scientific disciplines. Its integration of fossil pollen data and functional traits allows researchers to investigate long-term vegetation dynamics and their responses to past climate variations. Such integration provides critical context for contemporary ecological trends[e.g.,23,24], including trait information, which is currently underrepresented in paleoecological analyses of vegetation dynamics[25]. VegVault supports advanced modelling approaches, such as comparing species distributions across temporal scales[e.g.,26], and analysing biodiversity patterns using Joint Species Distribution Models[27] ('JSDM' hereafter). The database also facilitates studies of functional ecology and biogeography by enabling assessments of shifts in ecosystem functioning and climatic drivers over time, and by identifying the taxonomic and functional novelty of communities undergoing climate-driven changes[e.g.,28,29]. Furthermore, the detailed climate data, linked with vegetation records, provide robust foundation for examining the impacts of past and present climate change on ecosystems[e.g,30]. These capabilities make VegVault a valuable interdisciplinary resource for predicting biodiversity patterns and informing conservation strategies and policy decisions aimed at mitigating climate change effects. While this is not an exhaustive list of potential applications, VegVault offers a rich and accessible resource that can significantly advance ecological research and global biodiversity conservation.

Methods

VegVault workflow and reproducibility

VegVault integrates data from multiple well-established sources to provide a comprehensive view of vegetation dynamics. The current sources of data are summarised in **Table 1**.

The data processing and integration into VegVault follow a meticulous workflow to ensure consistency and reliability. Each data source is first accessed and processed individually using a specific pipeline. See **Figure 1** for the workflow process of accessing and processing all data sources.

In order to ensure the traceability of data, VegVault is designed as versioned project. Each version (including the current v1.0.0) is created using git Tags system[31]. Each Tag is a snapshot of code at certain points in time. Each of the individual data pipelines, coded in R programming language, is publicly available and stored in a separate GitHub repository with a specific git Tag and published with DOI (see legend of **Figure 1** for individual DOIs).

Finally, all processing pipelines with their corresponding Tags are then migrated into an SQLite database using another dedicated GitHub repository, ensuring its reproducibility and transparency of the development of the database.

In addition to the consolidation of all processed data into a unified SQLite database, the final VegVault 'migration' repository performs three additional procedures: (i) taxa classification, (ii) grouping of traits into *Trait Domains*, and (iii) creation of *gridpoints* for abiotic data.

(i) Taxa classification

As VegVault consist of data on taxa from various sources, the {*taxospace*} R package[21] is used to classify the diverse taxa into a unifying backbone in the VegVault database. The {*taxospace*} tool automatically aligns taxa names with the taxonomical backbone of the *GBIF*[20]. Specifically, we find the best match of the raw names of taxa using *Global Names Resolver*[32] that is then aligned with *GBIF*. Taxonomic information, detailed up to the family level, is stored for each taxon, ensuring consistency and facilitating comparative analyses across different datasets. However, note that taxonomic information down to species level is not available for each taxon (e.g., some fossil pollen types can be identified to genus or family).

(ii) Grouping of *Traits* into *Trait Domains*

As there are differences in trait names across sources of data (e.g., "*Leaf nitrogen (N) content per leaf dry mass*" and "*leaf nitrogen content per leaf dry mass*"), we added a new variable *Trait Domain* that groups traits together following the trait selection of Diaz et al. (2016)[33]. For example, trait "*Plant height vegetative*" from *TRY* and trait "*whole plant height*" from *BIEN* are both grouped under "*Plant height*" *Trait Domain*. See **Table 2** for full list of traits and their *Trait Domain*. Such grouping is aimed to serve as efficient extraction of traits across various data sources.

(iii) Creation of *gridpoints* of abiotic data

The abiotic data in the VegVault database provide essential information on environmental factors affecting vegetation distribution and traits, namely climate and soil variables (**Table 3**).

We developed a data structure that provides readily available environmental context for each vegetation (and trait) record by creating spatio-temporal links between these records and abiotic information. We first unified the resolution of the original rasters by resampling all abiotic data into ~ 25km resolution at 500-year time slices, reducing the amount of data. Next, as the rasters are not suitable to be stored in an SQLite database, we created artificial points, called '*gridpoints'***,** located in the middle of each raster cell. This resulted in a uniform spatio-temporal matrix of gridpoints holding the abiotic information (see **Figure 2B**or an example area). Next, we linked *gridpoints* and other non-*gridpoint Samples*, namely *vegetation_plot*, *fossil_pollen_archive*, and *traits* (see **Internal Database Structure** section), and calculated the spatial and temporal distances between them. We discarded any *gridpoint Sample* beyond 50 km and/or 5000 years to any other non-*gridpoint Sample*. See **Figure 2C** for an example of linking abiotic information at *gridpoints* and vegetation data. In VegVault, users can select the information for each non-*gridpoint Sample* from the closest spatio-temporal abiotic data or get the average from all surrounding *gridpoints*. By providing such comprehensive and well-linked structure between vegetation and abiotic data, VegVault enhances the ability to study the interactions between vegetation and their environment, facilitating the workflow to advanced ecological research and modelling efforts.

Data Records

The VegVault database v1.0.0 is an SQLite database of a size of ~ 110 GB stored in the *National Repository*[34] (‘*REPO*’). It contains > 480 000 *Datasets*, > 13 000 000 *Samples*, > 110 00 *Taxa*, 6 vegetation *Traits*, > 11 000 000 *Trait* values, and 8 abiotic variables. The spatio-temporal overview of all vegetation data can be found in **Figure 3**.

Currently, the SQLite database consists of 29 interconnected tables with 87 fields (variables). However, the internal structure of VegVault may not be of relevance to a user, as the {*vaultkeepr*}R package processes all the data to output only the most relevant information in a “ready-to-analyse” format (see **Usage Notes** section).

Internal Database Structure

VegVault is organized into several sections to systematically manage the varying datasets it integrates. The main structure is the *Dataset*, which serves as the cornerstone of the database structure. *Datasets* are composed of *Samples*, representing individual data points within each dataset. Each *Dataset* will only contain several *Samples* if these differ in age. There are four types of *Datasets*, namely (i) contemporary vegetation plots, (ii) past vegetation (fossil pollen records), (iii) functional traits, and (iv) gridpoint (a dataset type created to hold abiotic data, see details in **Methods** section).

For the contemporary (i) and past (ii) vegetation *Datasets*, the *Samples* hold information about *Taxa* as derived directly from the primary data sources. Trait information is organised in separate *Datasets* (as they are associated with unique information about their spatio-temporal location, references, etc) but linked to the same *Taxa* codes as those taxa in the vegetation *Datasets* (if present). Moreover, each *Taxa*, disregarding of the source, is associated with classification information (i.e. species, genus and family name; see **Methods** section). While the vegetation data is the main focus of the database, the structure of separate *Datasets* ensures a comprehensive organization of the interlinked data. Note that there is more in-depth and updated documentation of database structure on a project websites (*https://bit.ly/VegVault*).

Dataset

Here we will explain some, but not all, of the features of the *Dataset*.

Dataset Type

The *Dataset Type* defines the most basic classification of each *Dataset*, ensuring that the vast amount of data is categorized systematically. Currently, VegVault contains the following types of *Datasets*:

- ***`vegetation_plot`***: This type includes contemporary vegetation plot data, capturing contemporary vegetation characteristics and distributions.
- ***`fossil_pollen_archive`***: This type encompasses past vegetation plot data derived from fossil pollen records, providing insights into historical vegetation patterns.
- ***`traits`***: This type contains functional trait data, detailing specific characteristics of plant species that influence their ecological roles.
- ***`gridpoints`***: This type holds artificially created *Datasets* to manage abiotic data, here climate and soil information

See Error! Reference source not found. for overview of number of *Datasets* in each *Dataset Type*.

Dataset Source-Type

Each *Dataset* is derived from a specific *Source-Type* including:

- ***BIEN***[6] (Botanical Information and Ecology Network).
- ***sPlotOpen***[15] (The open-access version of sPlot).
- ***TRY***[8] (TRY Plant Trait Database).
- ***Neotoma-FOSSILPOL*** (FOSSILPOL[16]: Workflow for processing and standardizing global paleoecological pollen data from *Neotoma*[7]) Note that we specifically state *Neotoma-FOSSILPOL* and not just Neotoma, as *FOSSILPOL* not only provide the data acquisition but also alters it (e.g., creating new age-depth models). It also addresses major challenges in paleoecological data integration, such as age uncertainty, by incorporating probabilistic age-depth models and their associated uncertainty matrices. This enables the propagation of temporal uncertainty in subsequent analyses, a critical advancement for robust macroecological studies, previously flagged as major issue with paleo- data[35].
- ***Gridpoints*** (artificially created *Datasets* for abiotic data).

See **Figure 4** for overview of number of *Datasets* for each *Dataset Source-Type*. VegVault maintains the detailed information on the source used to retrieve the original data, thereby enhancing the findability and referencing of primary data sources.

Sampling Method

Information on *Sampling Method* is only present for both *vegetation_plot* and *fossil_pollen_archive Datasets*, and vary substantially reflecting the diverse nature of the data collected. Such information is crucial for understanding the context and limitations of each *Dataset Type*. For contemporary vegetation plots, sampling involves standardized plot inventories and surveys that capture detailed vegetation characteristics across various regions. Fossil pollen data are collected from sediment records from numerous different depositional environments[16] representing past vegetation and climate dynamics.

Dataset References

To support robust and transparent scientific research, each *Dataset* in VegVault can have multiple *References* at different levels as the raw data was pooled from other databases that provide data from various authors. The *Dataset Source-Type*, *Dataset Source*, and *Sampling Method* can all have their own *References*, providing detailed provenance and citation information. This multi-level referencing system enhances the traceability and validation of the data. The VegVault database v1.0.0 currently includes 706 sources of *Datasets* where each dataset can also have one or more direct references to specific data, further ensuring that users can accurately cite and verify the sources of their data.

Samples

Samples represent the main unit of data in the VegVault database, serving as the fundamental building blocks for all analyses. There are currently over 13 million *Samples* in the VegVault database v1.0.0 (of which ~ 1.6 million are *gridpoints* of abiotic data, see **Methods** section). Each *Sample* is linked to a specific *Dataset* via the *Dataset-Sample* table, which ensures that every sample is correctly associated with its corresponding *Dataset Type* (whether it is *vegetation_plots*, *fossil_pollen_archive*, *traits*, or *gridpoint*) and other *Dataset* properties (e.g., geographic location).

Vegetation plot size

The size of vegetation plots can vary substantially. This detail is crucial for ecological studies where plot size can influence species diversity and abundance metrics, thus impacting follow-up analyses and interpretations. To account for this variability, information about the plot size is recorded separately for each contemporary *Sample*. See **Figure 5** for an overview of number of *Samples* in each size category.

Sample age

VegVault encompasses both contemporary and paleo-data, necessitating accurate age information for each *Sample*. Contemporary *Samples* are assigned an age of 0, while *Samples* from fossil pollen records are in calibrated years before present (cal yr BP). The "present" is here specified as 1950 AD. In addition, each *Sample* from *fossil_pollen_archive Dataset* is also associated with an uncertainty matrix. This matrix provides a range of potential ages derived from age-depth modelling, reflecting the inherent uncertainty in dating paleoecological records. See an example of detailed age uncertainty information for one fossil pollen record in **Figure 6**.

Sample reference

Each *Sample* in VegVault can have specific *References* in addition to the those at *Dataset*-level. These individual *Sample References* provide detailed provenance and citation information, ensuring that users can trace the origin and validation of each data point. This level of referencing enhances the transparency and reliability of the data, especially when the database continues to be updated in the future.

Taxa

The VegVault database records taxa names derived directly from the primary data sources. Each individual *Taxon* is linked to corresponding *Samples* through the *SampleTaxa* table, ensuring accurate and systematic association between species and their ecological data. Currently, the VegVault database v1.0.0 holds over 110 thousand taxonomic names. Each taxonomic name undergoes an automated classification (see **Methods** section). Through this procedure, only 1312 (1.2%) of all *Taxa* remained unclassified. See **Figure 7** for overview of number of species, genus and family classification.

Traits

Functional traits of vegetation taxa follow the same structure of *Dataset* and *Samples* obtained directly from *Dataset Source-Types*. Storing trait values in VegVault involves linking information across *Datasets*, *Samples*, *Taxa*, and *Traits*. This comprehensive linkage ensures that each *Trait* value is accurately associated with its relevant ecological, environmental and taxonomic context. Yet, it is up to the user to decide on how to further aggregate trait values if multiple trait samples of one *Trait Domain* are available for the same environmental or taxonomic entity. *To* ensure clarity and reproducibility, each *Trait* in the VegVault database can have additional *References* beyond the

general *Dataset* and *Sample References*. These *Trait*-specific *References* provide detailed provenance and citation information, supporting rigorous scientific research and enabling users to trace the origins and validation of each trait value.

Traits are further grouped into *Trait Domains* for easier summarisation of *Traits* across data sources (see **Methods** section). In total, six Trait domains are present: *Stem specific density*, *Leaf nitrogen content per unit mass*, *Diaspore mass*, *Plant height*, *Leaf area*, *Leaf mass per area*, following Diaz et al. (2016)[33]. See **Table 2** and for overview of *Traits* in various *Trait Domain* and number of *Trait* values.

Abiotic data

Currently, VegVault includes abiotic data from *CHELSA*[9], *CHELSA-TRACE21*[18,19], and *WoSIS*[17]. The first two data sources provide high-resolution downscaled climatic data, while the latter offers detailed soil information (only available for contemporary data). The abiotic variables included in the database are summarized in **Table 3**. Values of abiotic data are stored in *gridpoints*, artificially created *Datasets* and *Samples* (see **Methods**), with one *Dataset* holding more *Samples* only if they differ in age.

Technical Validation

The data validity of VegVault is driven by the completeness, accuracy, and reliability of the source data. Moreover, licence of the primary data sources prohibits the changing of subset of the data (e.g., *BIEN* data has *CC BY-NC-ND* licence). Nevertheless, several steps have been undertaken to ensure that VegVault provides high-quality data for research applications.

For the fossil pollen records, strict data-quality rules have been applied by using the *FOSSILPOL* workflow[16]. This includes selection of depositional environments, ecological groups, chronology control point types, and minimum number of chronology control points to construct age-depth models. Individual samples and records have been filtered by age limits, number of pollen grains, maximum age interpolation, and number of valid levels. In addition, the accuracy of fossil pollen data is increased by also re-estimation of all age-depth models using the Bayesian probabilistic approach[36] and include the information about individual age uncertainty. For more information about *FOSSILPOL* workflow, see https://bit.ly/FOSSILPOL. For exact values used for accessing, processing, and filtration of fossil pollen data for VegVault v1.0.0, see https://github.com/OndrejMottl/VegVault-FOSSILPOL/tree/v1.0.0.

Usage Notes

Guidelines for data reuse and update

Due to the version system of VegVault, the database is expected to be updated using new data and/or changing the process of data migration and processing. Therefore, we recommend citing the latest VegVault version with corresponding DOI (v1.0.0 can be found here: doi.org/10.48700/datst.m7ky4-4de51).

As the primary data sources are held under various licences, the VegVault has an *CC BY-NC-ND 4.0* Licence (https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en), allowing the usage of more restricted data (i.e. *BIEN* data is under *CC BY-NC-ND*). We recommend users to carefully read the constricts associated with data usage. In addition, when citing the usage of VegVault, users should always cite the primary sources of the data, as well as any individual datasets. For those purposes, we created a function `*vaulkeeper::get_references()*` within the {*vaultkeepr*} R package (see below) that will output a list of recommended references for the specific data compilation created by the user.

Limitations and considerations for future research

Potential limitation lies in the completeness, accuracy, and reliability of the primary data sources. For example, it has been suggested that improper data practices erode the quality of global ecological databases such as *TRY*[37], which have been used in VegVault. We do not attempt to solve these problems as VegVault's main purpose is the re-usability of the data. Note that for some cases we are not permitted to make modifications based on the original data (e.g. *BIEN* data has *CC BY-NC-ND* licence).

{vaultkeeper}: R-package for accessing the VegVault database

In order to make usage of the VegVault database as easy as possible, we have developed an R-package called {*vaultkeepr*}, providing a suite of functions to effectively interact with the database directly from the R programming environment. Functions include opening the database, extracting datasets, filtering samples, and accessing specific taxa and traits. This package is a well-tested interface (>95% code coverage). For more details, see {*vaultkeepr*} website (https://bit.ly/vaultkeepr).

Here we present 3 examples of potential projects and how to obtain data for such projects from VegVault using {*vaultkeepr*} R-package. Note that we are specifically not trying to do any analysis, only presenting a way to obtain data, which can be used for such projects.

1) Spatio-temporal patterns of the *Picea* genus across North America since the Last Glacial Maximum

The first example demonstrates how to retrieve data for the genus Picea across North America by selecting both modern and fossil pollen plot datasets, filtering samples by geographic boundaries and temporal range (0 to 15,000 yr BP), and harmonizing taxa to the genus level. The resulting dataset allows users to study spatiotemporal patterns of *Picea* distribution over millennia (**Figure 8A**). This can be accomplished by running the following code:

```{r}
data_na_plots_picea <-
  # Access VegVault
  vaultkeepr::open_vault(path = "<path_to_VegVault>") %>%
  # Start by adding dataset information
  vaultkeepr::get_datasets() %>%
  # Select both contemporary and paleo plot data
  vaultkeepr::select_dataset_by_type(
    sel_dataset_type = c(
      "vegetation_plot",
      "fossil_pollen_archive"
    )
  ) %>%
  # Limit data to North America
  vaultkeepr::select_dataset_by_geo(
    lat_lim = c(22, 60),
    long_lim = c(-135, -60)
  ) %>%
  # Add samples
  vaultkeepr::get_samples() %>%
  # Limit the samples by age
  vaultkeepr::select_samples_by_age(
    age_lim = c(0, 15e3)
  ) %>%
  # Add taxa & classify all data to a genus level
  vaultkeepr::get_taxa( classify_to = "genus") %>%
  # Extract only Picea data
  vaultkeepr::select_taxa_by_name(sel_taxa = "Picea") %>%
  vaultkeepr::extract_data()
```

2) Joined-species Distribution Model for vegetation within Czech Republic

In the second example, the project aims to do species distribution modelling for plant taxa in the Czech Republic based on contemporary vegetation plot data and mean annual temperature. The code includes selecting datasets and extracting relevant abiotic data (**Figure 8B**) as followed:

```{r}
data_cz_contemporary <-
  # Access VegVault file
  vaultkeepr::open_vault(path = "<path_to_VegVault>") %>%
  # Add the dataset information
  vaultkeepr::get_datasets() %>%
  # Select contemporary plot data and climate
  vaultkeepr::select_dataset_by_type(
    sel_dataset_type = c(
      "vegetation_plot",
      "gridpoints"
    )
  ) %>%
  # Limit data to Czech Republic
  vaultkeepr::select_dataset_by_geo(
    lat_lim = c(48.5, 51.1),
    long_lim = c(12, 18.9)
  ) %>%
  # Add samples
  vaultkeepr::get_samples() %>%
  # select only contemporary data
  vaultkeepr::select_samples_by_age(
    age_lim = c(0, 0)
  ) %>%
  # Add abiotic data
  vaultkeepr::get_abiotic_data() %>%
  # Select only Mean Annual Temperature (bio1)
  vaultkeepr::select_abiotic_var_by_name(sel_var_name = "bio1") %>%
  # add taxa
  vaultkeepr::get_taxa() %>%
  vaultkeepr::extract_data()
```

3) Patterns of community weighted mean of plant height for South and Central America between 6-12 ka cal yr BP

The third example focuses on obtaining data to be able to reconstruct plant height for South and Central America between 6-12 ka cal yr BP (thousand years before present). This example project showcases the integration of trait data with paleo-vegetation records to subsequently study historical vegetation dynamics and functional composition of plant communities (**Figure 8C**):

```{r}
data_la_traits <-
  # Access VegVault file
```

```
  vaultkeepr::open_vault(path = "<path_to_VegVault>") %>%
  # Add the dataset information
  vaultkeepr::get_datasets() %>%
  # Select contemporary plot data and climate
  vaultkeepr::select_dataset_by_type(
    sel_dataset_type = c(
      "fossil_pollen_archive",
      "traits"
    )
  ) %>%
  # Limit data to South and Central America
  vaultkeepr::select_dataset_by_geo(
    lat_lim = c(-53, 28),
    long_lim = c(-110, -38),
    sel_dataset_type = c(
      "fossil_pollen_archive",
      "traits"
    )
  ) %>%
  # Add samples
  vaultkeepr::get_samples() %>%
  # Limit to 6-12 ka yr BP
  vaultkeepr::select_samples_by_age(
    age_lim = c(6e3, 12e3)
  ) %>%
  # add taxa & classify all data to a genus level
  vaultkeepr::get_taxa(classify_to = "genus") %>%
  # add trait information & classify all data to a genus level
  vaultkeepr::get_traits(classify_to = "genus") %>%
  # Only select the plant height
  vaultkeepr::select_traits_by_domain_name(sel_domain = "Plant height") %>%
  vaultkeepr::extract_data()
```

Code Availability

The VegVault database is built on SQLite, chosen for its robustness and compatibility with various data types. The database itself is under *CC BY-NC-ND 4.0* licence, due to the licences of undelaying data (see **Usage Notes** section).

VegVault utilizes several tools and software to facilitate data management and access. All code used to generate the VegVault database is publicly available under MIT licence (see **Methods** section and **Figure 1** for details about individual GitHub repositories).

To enable easy access and manipulation of data, {*vaultkeepr*} R package has been developed as an open access software under MIT licence.

The code to create the figure in this manuscript is available under DOI:10.5281/zenodo.14777448.

Acknowledgements

OM is funded by the *Czech Science Foundation PIF* grant (GN23-063861), by the *Charles University Research Centre program* (UNCE/24/SCI/006), and by the *Institutional Support for Science and Research* of the Ministry of Education, Youth and Sports of the Czech Republic. SGAF is funded by the *Trond Mohn Research Foundation* (TMF) and University of Bergen by starting grant no. TMS2022STG03).

Author contributions

OM conceived the study and programmed all related code; SGAF, FG, and IS performed critical comments to the technical infrastructure of database; OM & SGAF wrote the first draft of the manuscript, SGAF, FG, and IS provided several rounds of revisions. All authors approved of the final version of the manuscript.

Competing interests

The author(s) declare no competing interests.

Figures

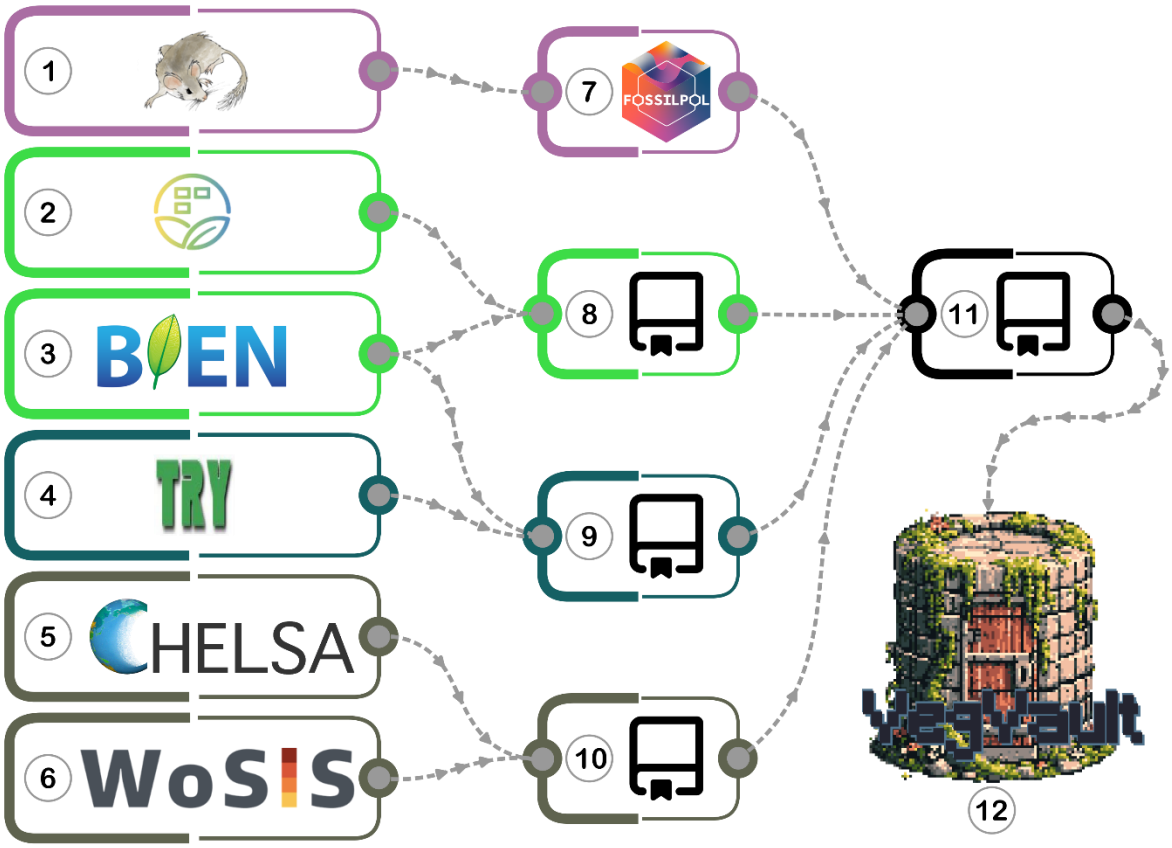

Figure 1. Schematic workflow process of accessing and processing all primary data sources into VegVault v1.0.0 database. 1) *Neotoma Paleoecological Database*[7]; **2)** *sPlotOpen*[15]; **3)** *Botanical Information and Ecology Network*[6]; **4)** *TRY Plant Trait Database*[8]; 5) *CHELSA*[9] and *CHELSA-TRACE21*[18,19]; **6)** *World Soil Information Service*[17]; **7)** *FOSSILPOL: Workflow for processing and standardizing global paleoecological pollen data*. A project-specific version with all the settings used for VegVault is stored as GitHub repository *VegVault-FOSSILPOL* (can be access as DOI: 10.5281/zenodo.14745060); **8)** A GitHub repository *VegVault-Vegetation_data* used to processes data for contemporary vegetation plots (can be accessed as DOI: 10.5281/zenodo.14745273); **9)** A GitHub repository *VegVault-Trait_data* used to processes data for plant vegetation traits (can be accessed as DOI: 10.5281/zenodo.14745086); **10)** A GitHub repository *VegVault-abiotic_data* used to process climate and soil abiotic data (can be access as DOI: 10.5281/zenodo.14745293); **11)** A GitHub repository *VegVault* used to integrate all data together and migrate them int SQLite database (can be accessed as DOI: 10.5281/zenodo.14777448). Additional steps such as the creation of *gridpoints*, grouping of *Traits* into *Trait Domains,* and classification of taxa are also present (for more details see **Methods**); **12)** VegVault - SQLite database (v1.0.0; DOI: 10.48700/datst.m7ky4-4de51); For details about primary data sources, see **Table 1**.

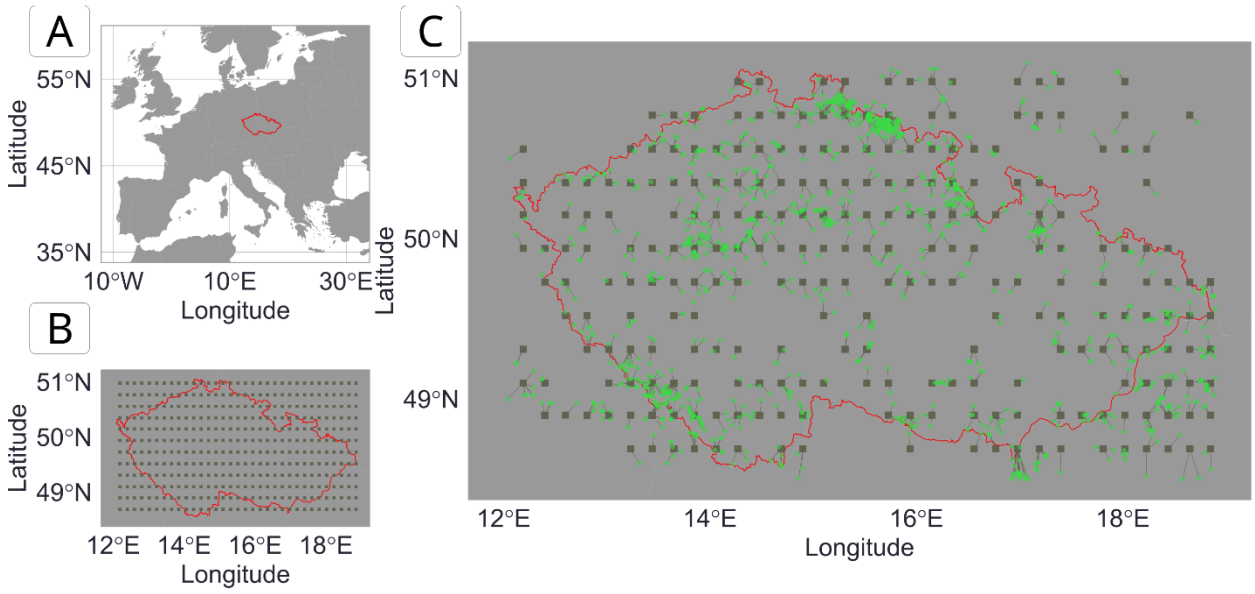

Figure 2. **Example of the resolution of the *gridpoint Datasets* extracting abiotic information.** The red line indicates country borders with Czech Republic as the focus of the map. **A**: map of Europe. **B**: Distribution of the individual *gridpoints* (grey squares) within the Czech Republic. **C**: Link between *gridpoints* and vegetation plots. The green circles represent the contemporary vegetation plots. The grey lines connect the vegetation points to their closest *gridpoint*. Note that the abiotic data is provided to each vegetation plot from their closest *gridpoint* (i.e., the vegetation plots connected to the same gridpoint will have the same abiotic data). *Gridpoint* without vegetation plot in their spatio-temporal vicinity (here contemporary data) were excluded from the VegVault database to reduce the database size. See **Methods** section for more details.

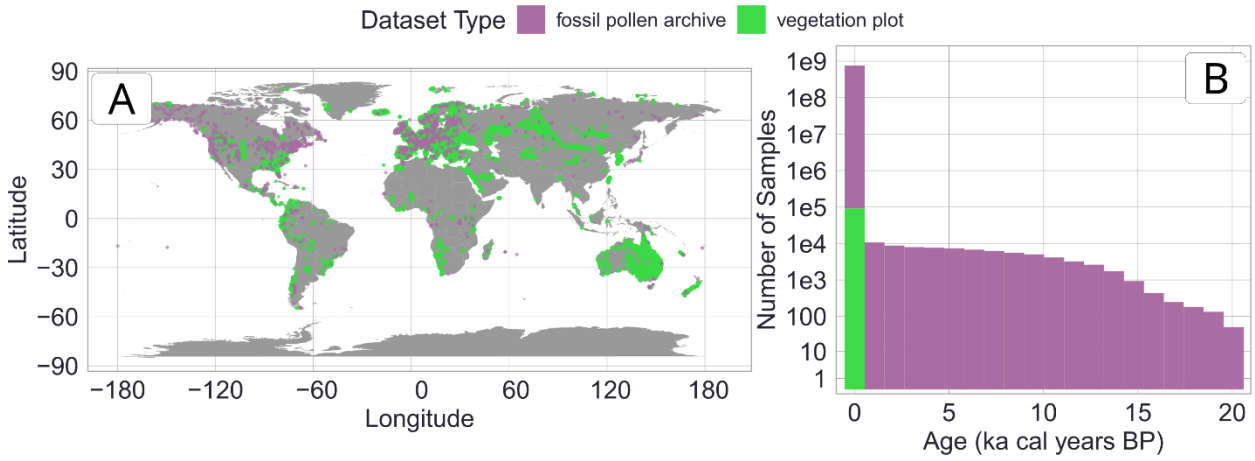

Figure 3. **Spatio-temporal distribution of vegetation Samples** (for both vegetation plots and fossil pollen archives). Note that this is overview of current state of VegVault (v1.0.0.). **A**: Spatial distribution with each point representing one plot. **B**: Temporal distribution of individual Samples.

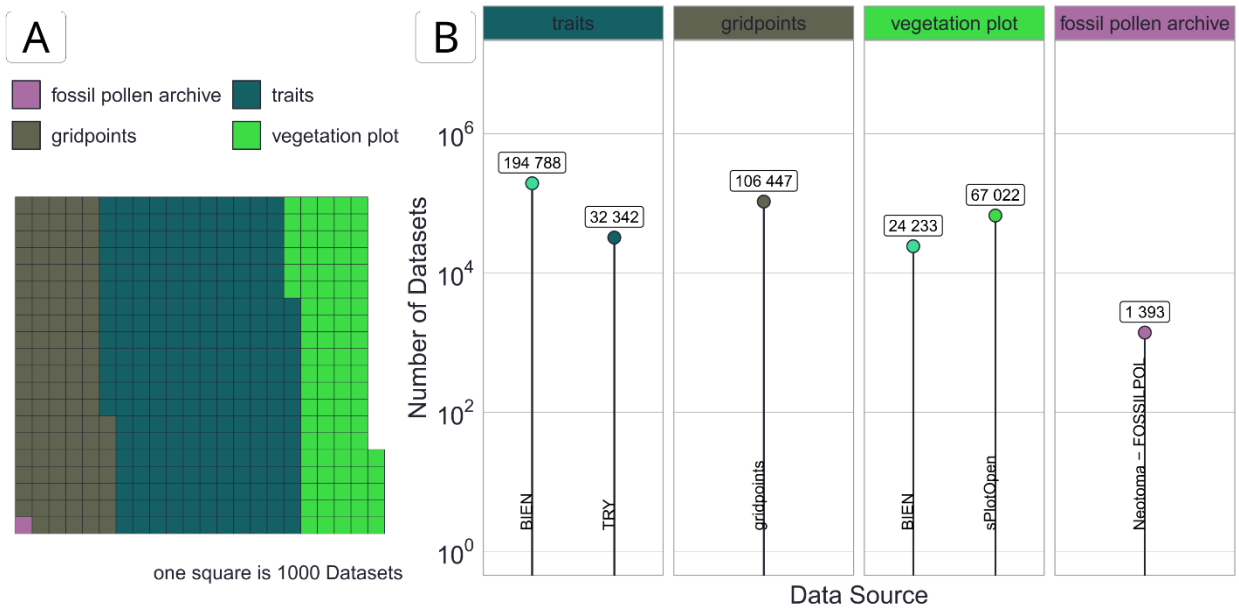

Figure 4. **Visualisation of the number of *Datasets*. A**: waffle plot coloured by *Dataset Type* with each square representing ~1000 *Datasets*. **B**: Number of *Datasets* by *Dataset Source-Type* grouped by *Dataset Type*. See **Internal Database Structure** section for more details.

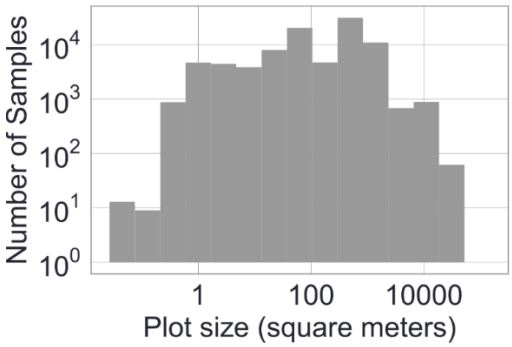

Figure 5. **Distribution of *Samples* by vegetation plot sizes.** In VegVault, only *vegetation_plot Samples* and *Datasets* has such information. See **Internal Database Structure** section for more details.

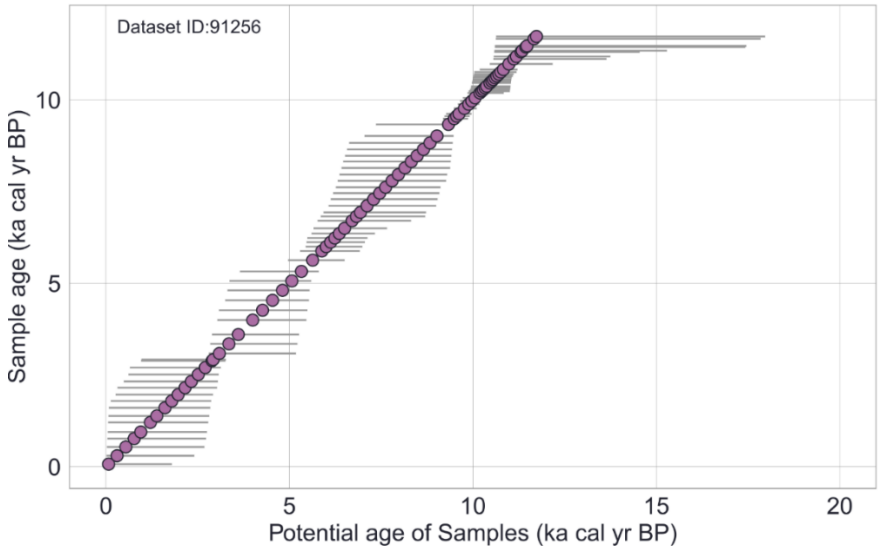

Figure 6: **Example of age uncertainty of *Samples* for selected fossil pollen *Dataset* (ID: 91256).** Each point represents one *Sample*. Each horizontal line represents a range of all potential ages associated with that *Sample*. Potential ages are gathered from age-depth models and stored in *SampleUncertainty* table. See **Internal Database Structure** section for more details about the data structure and *FOSSILPOL* workflow[16] for more details about age uncertainty from age-depth models.

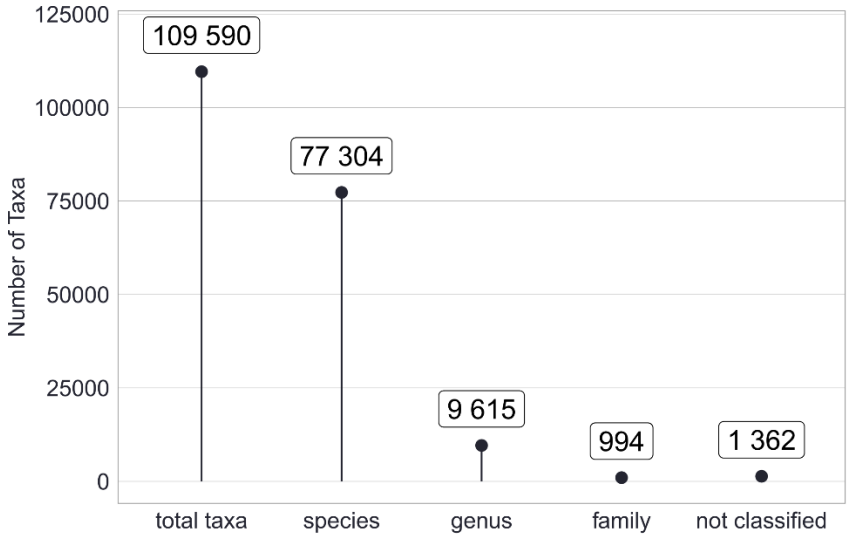

Figure 7. Total number of *Taxa* and their classification present in VegVault. See Methods section for more information about the classification process.

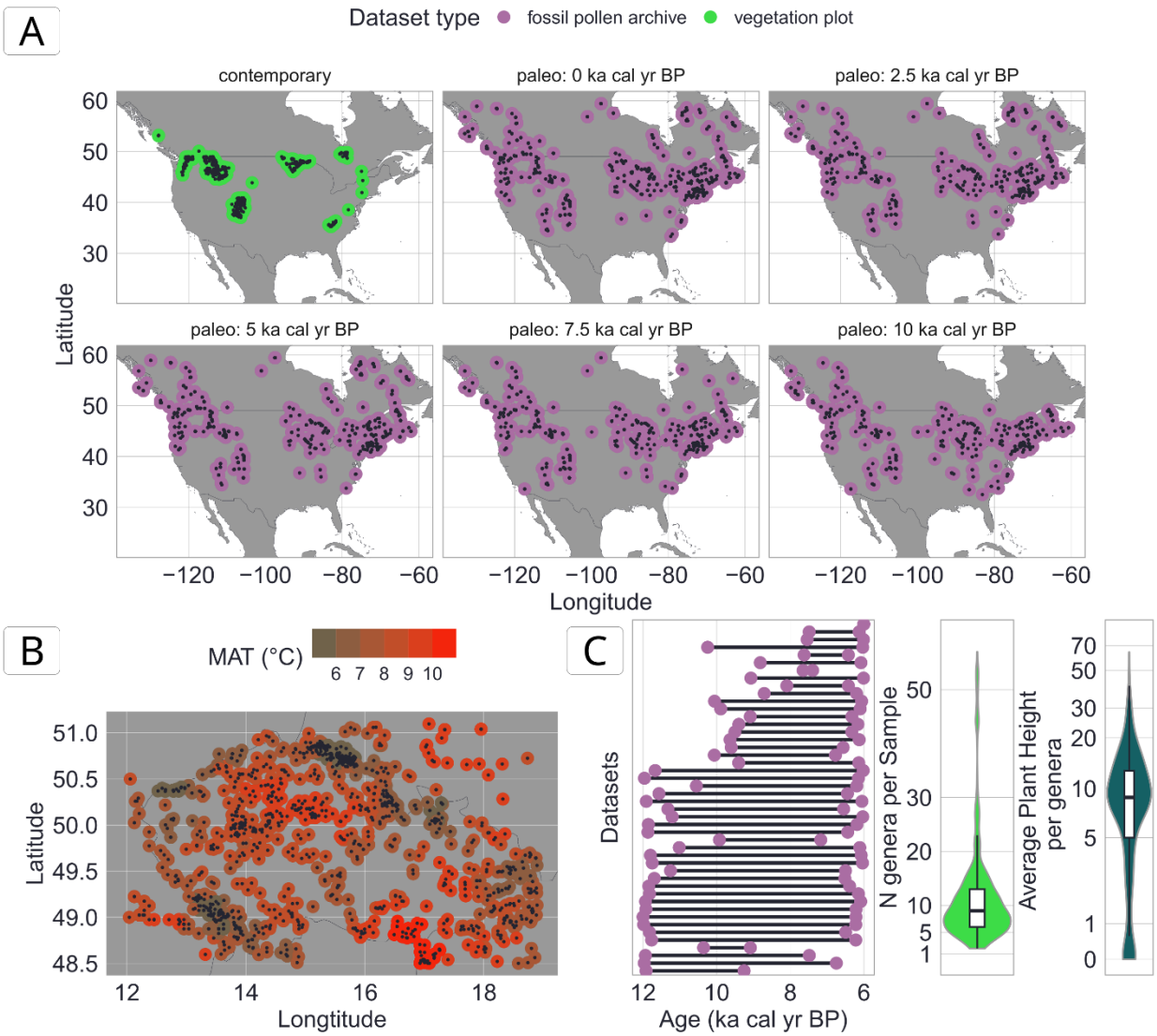

Figure 8. Visualisation of example data usage of VegVault using vaultkeepr package. See **Usage Notes** section for code used to extract such data from VegVault and for details about the aim of each data compilation. **A**: Data for "*Spatio-temporal patterns of the Picea genus across North America since the Last Glacial Maximum*" example**.** The individual points represent vegetation data (*Samples* for fossil pollen or contemporary vegetation plot) within each age time slice. **B**: Data for "*Joined-species Distribution Model for vegetation within Czech Republic*" example. Each point represents one contemporary vegetation plot. The colour represents the Mean Annual Temperature (MAT) associated with that plot. Note that several vegetation plots can have the same value due to sharing the same closest *gridpoint* (see **Figure 2**). **C**: Data for "*Patterns of community weighted mean of plant height for South and Central America between 6-12 ka cal* yr BP". The left figure represents temporal length of fossil pollen *Datasets*, which are available for Central and South America between 6 and 12 ka BP (thousand years before present). The middle figure represent the distribution various genera present in this data assembly. The right figure represents distribution of average *Plant height Trait* per each genus present in the data assembly

Tables

Table 1: Main sources of data presented in VegVault v1.0.0.

Source	Type	Comment	References
Botanical Information and Ecology Network (*BIEN*)	Vegetation – contemporary data	Plot-based vegetation data	Botanical Information and Ecology Network (http://bien.nceas.ucsb.edu/bien/)[6]
sPlotOpen	Vegetation - contemporary data	The open-access version of *sPlot*[5]	Sabatini et al. (2021)[15]
FOSSILPOL: Workflow for processing and standardizing global paleoecological pollen data	Vegetation - Fossil pollen records	We specifically state FOSSILPOL and not Neotoma Paleoceologcal database[7], which is the primary source of the data, as FOSSILPOL not only accesses the data but also applies several selection criteria and adds new information (e.g., new age-depth models)	Flantua et al. (2023)[16]
TRY Plant Trait Database (*TRY*)	Functional traits of vegetation		Kattge et al. (2020)[8]
CHELSA	Abiotic data – climate	global downscaled climate data at 1km resolution	Karger et al. (2017)[9]
CHELSA-Trace21k	Abiotic data – past climate	global downscaled climate data at 1km resolution since the LGM	Karger et al. (2020)[18] Karger et al. (2023)[19]
World Soil Information Service (*WoSIS*)	Abiotic data – soil types	global downscaled climate data at 1km resolution	Batjes et al. (2024)[17]

Table 2: Selected *Traits* presented in the VegVault v1.0.0. The *Traits* were selected based on similarity to Diaz et al. (2016)[33] and grouped into six main *Trait Domains* to simplify the navigation between different data sources.

Trait Domain	***Trait* name**	***Trait* units**	**Source of *Datasets***	**Number of *Datasets***
Stem specific density	*Stem specific density (SSD, stem dry mass per stem fresh volume) or wood density*	g/cm^3	*TRY*	8058

Stem specific density	*stem wood density*	g/cm^3	*BIEN*	348
Leaf nitrogen content per unit mass	*Leaf nitrogen (N) content per leaf dry mass*	mg/g	*TRY*	11464
Leaf nitrogen content per unit mass	*leaf nitrogen content per leaf dry mass*	mg/g	*BIEN*	707
Diaspore mass	*Seed dry mass*	mg	*TRY*	1635
Diaspore mass	*seed mass*	mg	*BIEN*	48
Plant height	*Plant height vegetative*	m	*TRY*	11476
Plant height	*whole plant height*	m	*BIEN*	193492
Leaf area	*Leaf area (in case of compound leaves: leaf, petiole excluded)*	mm^2	*TRY*	1415
Leaf area	*Leaf area (in case of compound leaves: leaf, petiole included)*	mm^2	*TRY*	1374
Leaf area	*Leaf area (in case of compound leaves: leaf, undefined if petiole in- or excluded)*	mm^2	*TRY*	897
Leaf area	*Leaf area (in case of compound leaves undefined if leaf or leaflet, undefined if petiole is in- or exluded)*	mm^2	*TRY*	2202
Leaf area	*Leaf area*	mm^2	*BIEN*	902
Leaf mass per area	*Leaf area per leaf dry mass (specific leaf area, SLA or 1/LMA): petiole included*	mg/mm^2	*TRY*	1935
Leaf mass per area	*Leaf area per leaf dry mass (specific leaf area, SLA or 1/LMA): undefined if petiole is in- or excluded)*	mg/mm^2	*TRY*	8777
Leaf mass per area	*Leaf mass per area*	mg/mm^2	*BIEN*	835

Table 3: Overview of abiotic variables present in VegVault v1.0.0.

Abiotic variable name	**Abiotic variable description**	**Abiotic variable unit**	**Source of data**
bio1	Annual Mean Temperature	°C	*CHELSA & CHELSA-TRACE21*
bio4	Temperature Seasonality	°C	*CHELSA & CHELSA-TRACE21*
bio6	Min Temperature of Coldest Month	°C	*CHELSA & CHELSA-TRACE21*
bio12	Annual Precipitation	kg m-2 year-1	*CHELSA & CHELSA-TRACE21*
bio15	Precipitation Seasonality	Unitless	*CHELSA & CHELSA-TRACE21*
bio18	Precipitation of Warmest Quarter	kg m-2 quarter-1	*CHELSA & CHELSA-TRACE21*
bio19	Precipitation of Coldest Quarter	kg m-2 quarter-1	*CHELSA & CHELSA-TRACE21*
HWSD2	Soil type	Unitless	*WoSIS*

References

1. Fordham, D. A. *et al.* Using paleo-archives to safeguard biodiversity under climate change. *Science* **369**, eabc5654 (2020).
2. Nieto-Lugilde, D. *et al.* Time to better integrate paleoecological research infrastructures with neoecology to improve understanding of biodiversity long-term dynamics and to inform future conservation. *Environ. Res. Lett.* **16**, 095005 (2021).
3. Reitalu, T., Kuneš, P. & Giesecke, T. Closing the gap between plant ecology and Quaternary palaeoecology. *Journal of Vegetation Science* **25**, 1188–1194 (2014).
4. GBIF.org. GBIF Home Page. https://www.gbif.org (2024).
5. Bruelheide, H. *et al.* sPlot – A new tool for global vegetation analyses. *Journal of Vegetation Science* **30**, 161–186 (2019).
6. Botanical Information and Ecology Network (BIEN). BIEN: Botanical Information and Ecology Network. http://bien.nceas.ucsb.edu/bien/ (n.d.).
7. Williams, J. W. *et al.* The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quat. res.* **89**, 156–177 (2018).
8. Kattge, J. *et al.* TRY plant trait database – enhanced coverage and open access. *Global Change Biology* **26**, 119–188 (2020).
9. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci Data* **4**, 170122 (2017).
10. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302–4315 (2017).
11. Schipper, E. L. F., Dubash, N. K. & Mulugetta, Y. Climate change research and the search for solutions: rethinking interdisciplinarity. *Clim Change* **168**, 18 (2021).

12. Intergovernmental Panel On Climate Change. *Climate Change and Land: IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems*. (Cambridge University Press, 2022). doi:10.1017/9781009157988.
13. Smith, J. *et al*. BioDeepTime: A database of biodiversity time series for modern and fossil assemblages. *Global Ecology and Biogeography* **32**, 1680–1689 (2023).
14. Weigelt, P., König, C. & Kreft, H. GIFT – A Global Inventory of Floras and Traits for macroecology and biogeography. *Journal of Biogeography* **47**, 16–43 (2020).
15. Sabatini, F. M. *et al*. sPlotOpen – An environmentally balanced, open-access, global dataset of vegetation plots. *Global Ecology and Biogeography* **30**, 1740–1764 (2021).
16. Flantua, S. G. A. *et al*. A guide to the processing and standardization of global palaeoecological data for large-scale syntheses using fossil pollen. *Global Ecology and Biogeography* **32**, 1377–1394 (2023).
17. Batjes, N. H., Calisto, L. & de Sousa, L. M. Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023). *Earth System Science Data* **16**, 4735–4765 (2024).
18. Karger, D. N., Nobis, M. P., Normand, S., Graham, C. H. & Zimmermann, N. E. CHELSA-TraCE21k: Downscaled transient temperature and precipitation data since the last glacial maximum. (2020) doi:http://dx.doi.org/10.16904/envidat.211.
19. Karger, D. N., Nobis, M. P., Normand, S., Graham, C. H. & Zimmermann, N. E. CHELSA-TraCE21k – high-resolution (1km) downscaled transient temperature and precipitation data since the Last Glacial Maximum. *Climate of the Past* **19**, 439–456 (2023).
20. GBIF Secretariat. GBIF Backbone Taxonomy. (2023).
21. Mottl, O. taxospace - v0.0.0.9003. https://doi.org/10.5281/zenodo.14699491 (2024).

22. Mottl, O. vaultkeepr - v0.0.5. https://doi.org/10.5281/zenodo.14749734 (2024).
23. Mottl, O. *et al.* Global acceleration in rates of vegetation change over the past 18,000 years. *Science* **372**, 860–864 (2021).
24. Šímová, I., Ordonez, A. & Storch, D. The dynamics of the diversity–energy relationship during the last 21,000 years. *Global Ecology and Biogeography* **32**, 707–718 (2023).
25. Brown, K. A. *et al.* Trait-based approaches as ecological time machines: Developing tools for reconstructing long-term variation in ecosystems. *Functional Ecology* **37**, 2552–2569 (2023).
26. Shipley, B., Dilkina, B. & McGuire, J. MegaSDM: Modelling Species Ranges in The Past And Future. *Bulletin of the Florida Museum of Natural History* **60**, 115–115 (2023).
27. Pollock, L. J. *et al.* Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution* **5**, 397–406 (2014).
28. Ordonez, A. & Gill, J. L. Unravelling the functional and phylogenetic dimensions of novel ecosystem assemblages. *Philosophical Transactions of the Royal Society B: Biological Sciences* **379**, 20230324 (2024).
29. Wieczynski, D. J. *et al.* Climate shapes and shifts functional biodiversity in forests worldwide. *Proceedings of the National Academy of Sciences* **116**, 587–592 (2019).
30. Zurell, D., Fritz, S. A., Rönnfeldt, A. & Steinbauer, M. J. Predicting extinctions with species distribution models. *Cambridge Prisms: Extinction* **1**, e8 (2023).
31. Chacon, S. & Straub, B. *Pro Git*. (Springer Nature, 2014).
32. Global Names Architecture. Global Names Resolver. https://resolver.globalnames.org/.
33. Díaz, S. *et al.* The global spectrum of plant form and function. *Nature* **529**, 167–171 (2016).
34. CESNET. National Repository. https://data.narodni-repozitar.cz/ (2024).
35. Dillon, E. M. *et al.* Challenges and directions in analytical paleobiology. *Paleobiology* **49**, 377–393 (2023).

36. Haslett, J. & Parnell, A. A Simple Monotone Process with Application to Radiocarbon-Dated Depth Chronologies. *Journal of the Royal Statistical Society Series C: Applied Statistics* **57**, 399–418 (2008).

37. Augustine, S. P., Bailey-Marren, I., Charton, K. T., Kiel, N. G. & Peyton, M. S. Improper data practices erode the quality of global ecological databases and impede the progress of ecological research. *Global Change Biology* **30**, e17116 (2024).