UNIVERSITÉ DE LAUSANNE

DOCTORAL THESIS

Software and Numerical Tools for Paleoclimate Analysis

Author:
Philipp S. SOMMER

Supervisor:
Dr. Basil A. S. Davis

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Science

in the

The Davis Group
Institute of Earth Surface Dynamics (IDYST)

October 4, 2019

Contents

Chapter 4

The EMPD and POLNET web-interfaces

4.1 Summary

The Eurasian (née European) Modern Pollen Database (EMPD) was established in 2013 as a public database of quality controlled and standardized modern pollen surface sample data to compliment the European Pollen Database (EPD) for fossil pollen (B. A. S. Davis et al., 2013). The first version of the EMPD (referenced herein as the EMPD1) contained almost 5000 samples, submitted by over 40 individuals and research groups from all over Europe. Over the last 6 years more data has continued to be submitted, and more efforts have been made to incorporate more data held in open data repositories such as PANGAEA, and as supplementary information in published studies. This data is now released as the Eurasian Modern Pollen Database, version 2 (Basil A. S. Davis et al., in prep) with an increase of 80 percent to 8663 samples (see figure 4.1).

The EMPD remains the only public and open access database of modern pollen samples covering the Eurasian continent and is entirely driven by the community of its data contributors. This effort of creating an open and accessible database led to the development of new open source data management tools that we present in this article. The EMPD2 is now hosted on the version control platform Github, with a dedicated web viewer at EMPD2.github.io and a automated administration app, the EMPD-admin (see table 4.1 for a list of the web resources). The new framework provides a simplified and transparent administration of multiple contributions from different sources and people to the database. All web components are hosted without any additional costs and as such, they have the potential to be applied for other community-based (regional) pollen databases, such as the Latin American Pollen

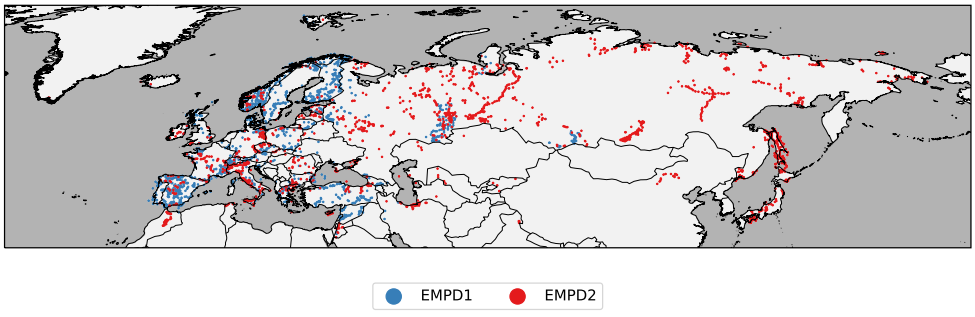


FIGURE 4.1: Modern calibration samples in the Eurasian Modern Pollen Database (EMPD).

TABLE 4.1: EMPD Web resources

	Description	Online Access
EMPD2	Github Organization	github.com/EMPD2
EMPD-Viewer	Map-based web interface to the EMPD database	github.com/EMPD2/EMPD-Viewer empd2.github.io
EMPD-Data	Version controlled data repository of the EMPD	github.com/EMPD2/EMPD-data
EMPD-Admin	Automated administration web app for the EMPD	github.com/EMPD2/EMPD-admin empd-admin.herokuapp.com EMPD2.github.io/EMPD-admin

Database (LAPD) or African Pollen Database (APD), for instance. Especially the light-weight EMPD-viewer web interface can be ported to other database (as shown in section 4.3) to make diverse data accessible to the broad public.

4.2 The EMPD web framework

The EMPD web framework is built on very common open source software development tools that have been adopted for a transparent data management, in favor of open science. The EMPD is now hosted on the web platform Github (github.com/EMPD2). This web platform, free of charge, hosts the source code for many popular open source software packages but can also be used to host a diverse, but small database (in terms of megabytes), such as the EMPD. Github builts upon the version control system *git* that transparently manages changes to documents by providing a full history of their revisions. The web platform is intrinsically designed for community-based projects that focus on collaboration and contains many features for a transparent communication between users, maintainers and contributors of a project. Besides others, the platform provides repository (i.e. project) specific discussion pages, so-called issues, where users can provide feedback, report bugs, or discuss any other aspect of the project. These issues are often linked to so-called pull requests, where each pull request is a proposal for a change in the source files of the project. This is then discussed between project maintainer and contributor in a dedicated discussion/review page.

Another common feature for Github repositories are integrations with so-called Continuous Integration (CI) services, e.g. for automated testing and/or packaging the software. These services run predefined scripts (for example test scripts) every time someone contributes to the repository, or creates a pull request.

The following sections describe how these software development tools are implemented in the three components of the EMPD web framework, the EMPD-viewer (section 4.2.1), the EMPD data repository (section 4.2.2) and the EMPD-admin (section 4.2.3).

4.2.1 The EMPD viewer

The main public interface into the EMPD is an interactive web viewer accessible from EMPD2.github.io. This JavaScript-based application (see figure 4.2 for a screenshot) provides an intuitive interface into the database without requiring any particular computer expertise. It enables the user to view the data on a map and select and

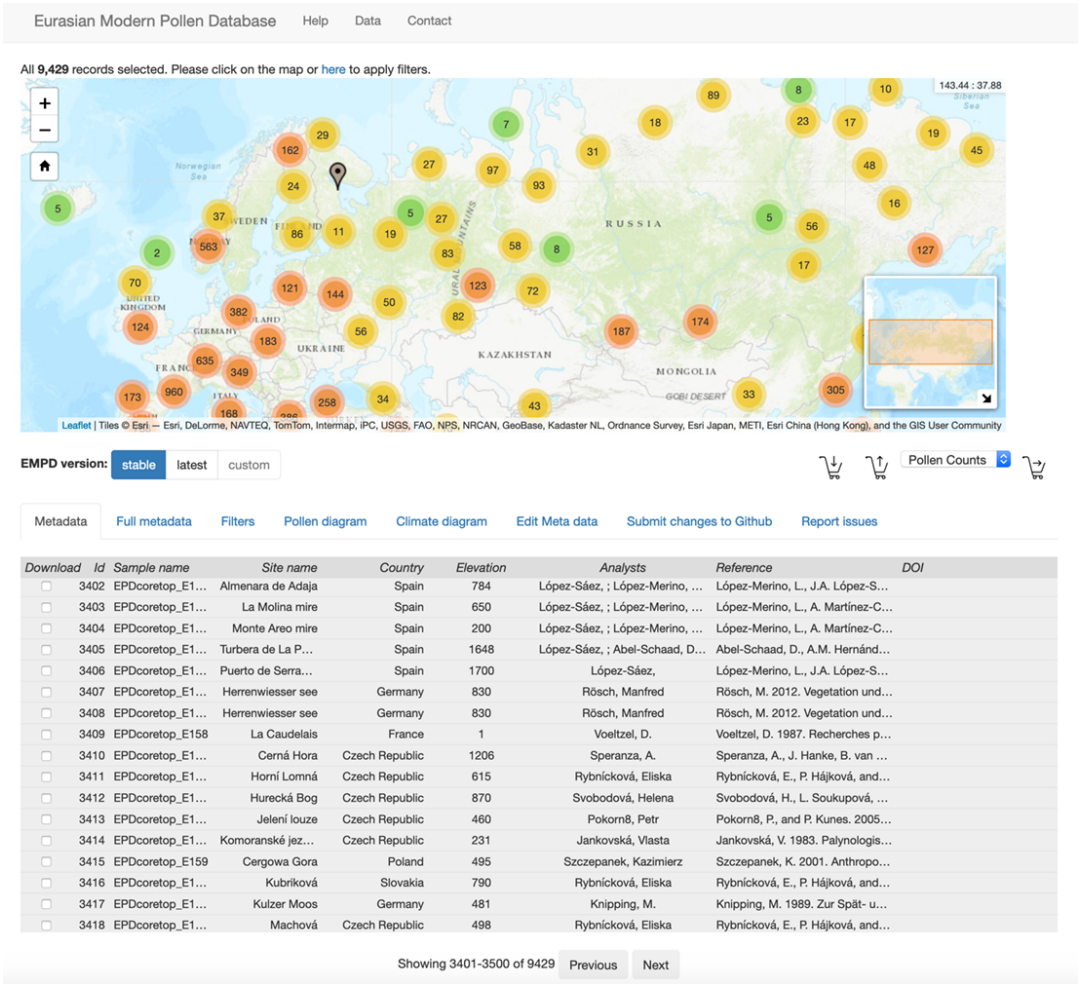


FIGURE 4.2: Screenshot of the EMPD viewer

download subsets of the database. The webpage involves no server-side processing and such it can be hosted for free using the service provided by Github Pages (pages.github.com). This provides a stable access to the database, independent of funding availabilities.

The Web Interface

The EMPD-Viewer has been initially based on the climate proxies finder (Bolliet et al., 2016; Brockmann, 2016) which can still be seen in it the layout and design of its graphical interface (i.e. its front-end). The code base, however, has been changed entirely, updated to the latest available versions of the underlying JavaScript dependencies and extended with multiple additional tools, shown in table 4.2. The central element of the viewer is a map to show the sample locations. It also allows to intuitive access to the essential meta data of every sample through the popup of the corresponding marker on the map. The detailed meta data can also be seen in the meta data table, together with all the other samples. Another key element of the viewer are the meta data filters, that subset the data using efficient and intuitive filtering tools. This allows to search the database, or to select specific countries, climatic regimes, sample types, samples of a specific data contributor/analyst, and more.

needs implementation

Additional information on the sample is revealed through a bar diagram of the associated pollen data, as well as a diagram showing the monthly, seasonal and annual precipitation and temperature at the side, based on the WorldClim dataset, version 2 (Fick and Hijmans, 2017).

Finally, the viewer contains elements that allow scientists to contribute to the database, even without dedicated knowledge about the Github framework. The meta data editor allows to edit a sample and then submit it via the data submission form. The request is handled by the EMPD-admin webapp (see section 4.2.3) that pushes the data to the corresponding pull request on Github that is then reviewed by the core database maintainers. Another implemented element is an issue report form that allows the user to highlight erroneous sample information which is then, again through the EMPD-admin, submitted as a Github issue to the data repository.

The web app is fully integrated into the Github framework of the EMPD and loads the displayed data from the online repository. As such, it also provides a further quality control check and allows the data contributors/maintainers to review and edit new contributions before they are merged into the database.

Implementation details

The viewer itself is very light-weight and can be flexibly adapted to other database systems (see for example section 4.3). As the climate proxies finder (Bolliet et al., 2016; Brockmann, 2016), the EMPD-viewers main viewing/filtering functionality it is built upon the *dc* (Zhu and the dc.js Developers, 2019), *crossfilter* (Square, Inc. and crossfilter contributors, 2019) and *leaflet* (Agafonkin and leaflet contributors, 2019) open source JavaScript libraries. We ported the app to the *npm* package manager (npmjs.com) which enables a better and more secure monitoring of the app dependencies. This package manager is also used for an automated testing of the viewer on a Continuous Integration (CI) service, prior to deployment on the official web page. Due to time constrains, the viewer is not yet fully adapted to mobile devices.

4.2.2 The EMPD2 data repository

The raw data of the EMPD2 is accessible as plain text files in the *EMPD-data* Github repository (see table 4.1). The software development framework of Github (see introductory part of section 4.2) is adopted such that issues in the data repository can highlight errors in the database, or provide room for the discussion of potential new efforts that should be considered within the community-database. Pull requests into the repository are new data contributions that can be reviewed by the maintainers before being merged into the official database.

This methodology allows a fully transparent traceback of changes made to the EMPD through version control. The online access to the raw data files through Github also allows the EMPD viewer to interface with different versions of the database (see previous section).

The EMPD-data repository additionally uses the CI services from Travis CI (travis-ci.org) for automated tests of the meta data in each sample.

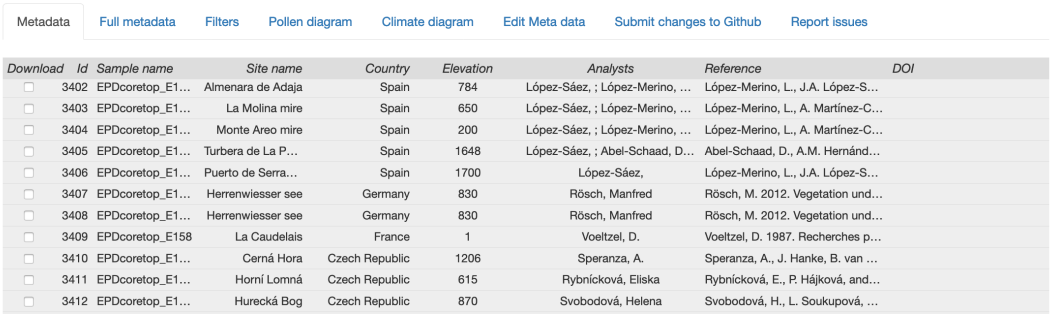
4.2.3 The EMPD-admin

In addition to the standard CI services, we developed the EMPD-admin webapp. Inspired by the web management tools of the conda-forge community (conda-forge.org), this tool provides an automated handling of data contributions from within Github

TABLE 4.2: Tools in the EMPD viewer

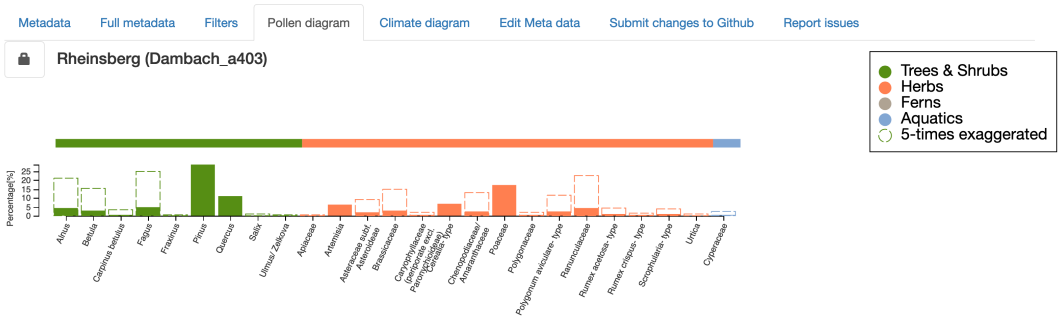
Map interface

Meta data table



Download	Id	Sample name	Site name	Country	Elevation	Analysts	Reference	DOI
	3402	EPDcoretop_E1…	Almenara de Adaja	Spain	784	López-Sáez, ; López-Merino, …	López-Merino, L., J.A. López-S…	
	3403	EPDcoretop_E1…	La Molina mire	Spain	650	López-Sáez, ; López-Merino, …	López-Merino, L., A. Martínez-C…	
	3404	EPDcoretop_E1…	Monte Areo mire	Spain	200	López-Sáez, ; López-Merino, …	López-Merino, L., A. Martínez-C…	
	3405	EPDcoretop_E1…	Turbera de La P…	Spain	1648	López-Sáez, ; Abel-Schaad, D…	Abel-Schaad, D., A.M. Hernánd…	
	3406	EPDcoretop_E1…	Puerto de Serra…	Spain	1700	López-Sáez,	López-Merino, L., J.A. López-S…	
	3407	EPDcoretop_E1…	Herrenwiesser see	Germany	830	Rösch, Manfred	Rösch, M. 2012. Vegetation und…	
	3408	EPDcoretop_E1…	Herrenwiesser see	Germany	830	Rösch, Manfred	Rösch, M. 2012. Vegetation und…	
	3409	EPDcoretop_E158	La Caudelais	France	1	Voeltzel, D.	Voeltzel, D. 1987. Recherches p…	
	3410	EPDcoretop_E1…	Cerná Hora	Czech Republic	1206	Speranza, A.	Speranza, A., J. Hanke, B. van …	
	3411	EPDcoretop_E1…	Horní Lomná	Czech Republic	615	Rybnícková, Eliska	Rybnícková, E., P. Hájková, and…	
	3412	EPDcoretop_E1…	Hurecká Bog	Czech Republic	870	Svobodová, Helena	Svobodová, H., L. Soukupová, …	

Pollen Data



Climate Data

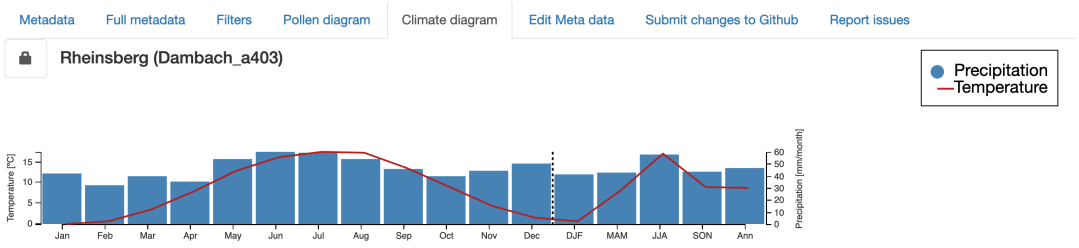
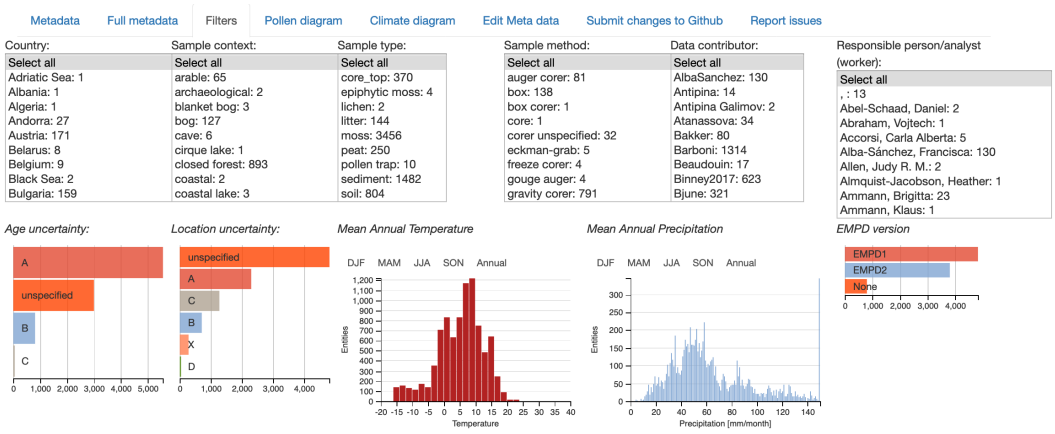
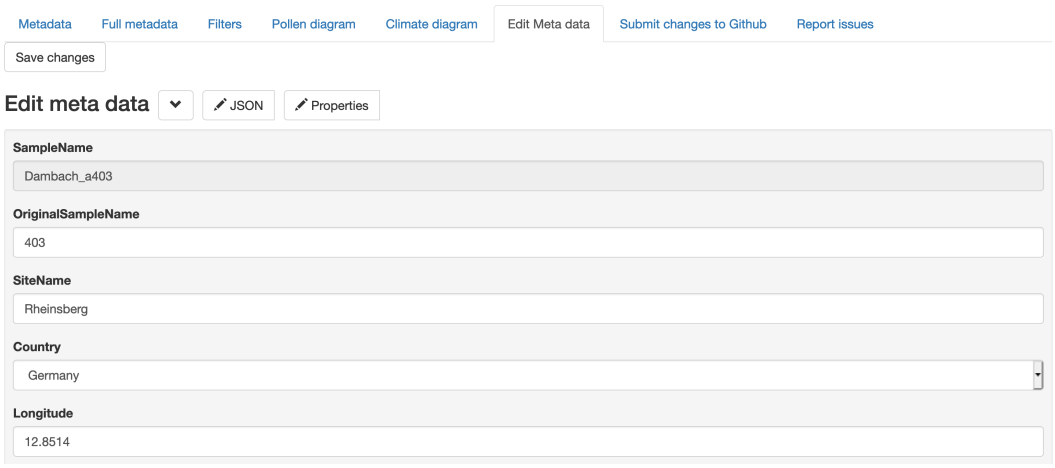


TABLE 4.2: Tools in the EMPD viewer (continued)

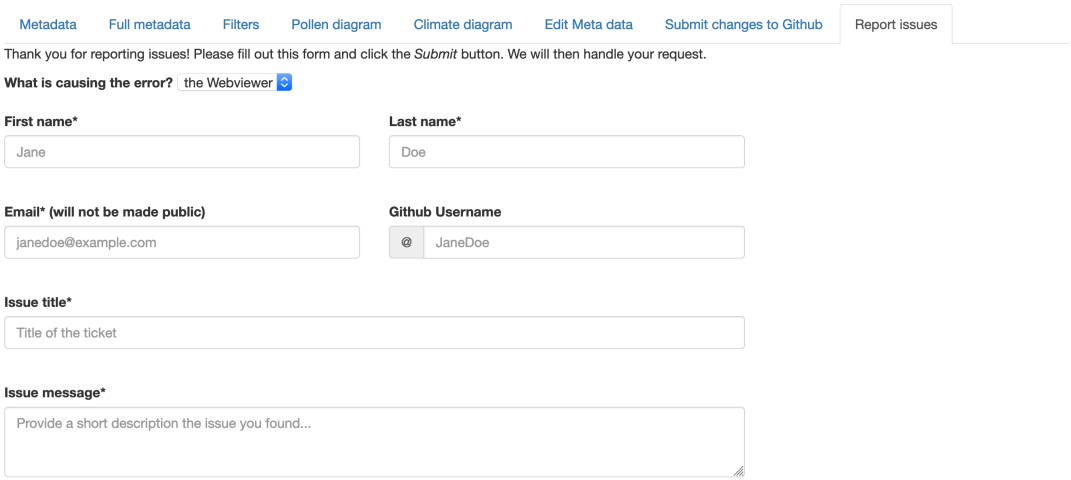
Meta data filter



Meta Data Editor



Issue submission form



Pull Requests. It acts like standard CI service and runs tests on the data contribution, every time changes have been made to the pull request.

But the main purpose of the EMPD-admin is to provide a web tool for an automated administration of the database, which is helpful for a community-project with changing maintainers. Hence, the EMPD-admin web app acts like a bot that reacts on comments within a pull request (i.e. data contribution). Maintainers and contributors can use this functionality and directly contact the bot, for instance, to subset the data, run specific tests on subsets of the data, or automatically fix certain meta data issues, such as wrong countries or missing elevation.

The bot is also integrated in the EMPD-viewer (see previous section 4.2.1). Bug reports or edited data are processed by the EMPD-admin and put online as an issue in the github repository, or it updates the corresponding data contribution.

As such, the administration of the database can be done entirely remotely, without having to install dedicated software on a local computer.

Implementation details

The EMPD-admin webapp is hosted for free at Heroku (https://www.heroku.com) at empd-admin.herokuapp.com with a software package documentation hosted at EMPD2.github.io/EMPD-admin. This, again, allows stability independent on the availability of funding. The package can, also be installed locally and used from the command-line, independent of Github and Heroku, which is sometimes helpful for very large data contributions..

The Python library is based on the tornado web framework www.tornadoweb.org, as well as pandas (McKinney, 2010), a tabular data analysis library for Python, and sqlalchemy (Bayer, 2012), a Python SQL toolkit.

4.2.4 Distribution of the tools

The EMPD is hosted within the EMPD2 Github organization (github.com/EMPD2) at github.com/EMPD2/EMPD-data. The source files of the viewer are accessible at github.com/EMPD2/EMPD-viewer, and for the EMPD-admin in the EMPD2/EMPD-admin repository (see also table 4.1).

The EMPD-data and the EMPD-admin are additionally both available as so-called Docker container image at `https://hub.docker.com/u/empd2`. These containers are lightweight, standalone, executable packages of software that include everything needed to run an application: code, runtime, system tools, system libraries and settings. As such, they extend standard software packaging systems by providing an entire operating system that contains the target application. This makes it especially useful for web applications (such as the EMPD-admin) that can, as such, operate in a well-defined and portable environment.

The EMPD-admin can, however, also be installed through the standard python package manager `pip`.

4.3 The POLNET viewer

The adaptability of the EMPD-viewer gave the motivation for an application with the POLNET database. This database, currently in development status, is a northern hemispheric, sub-tropical collection of modern and fossil pollen assemblages (Basil A. S. Davis and Kaplan, 2017; Sommer et al., 2019). The purpose of this database is to generate the source for large-scale climate reconstruction during the Holocene (past 12'000 years) that can be used for model-data comparisons. It contains about 3'300 fossil pollen sites and about 13'200 modern surface samples (see figure 4.3) and

Check this number

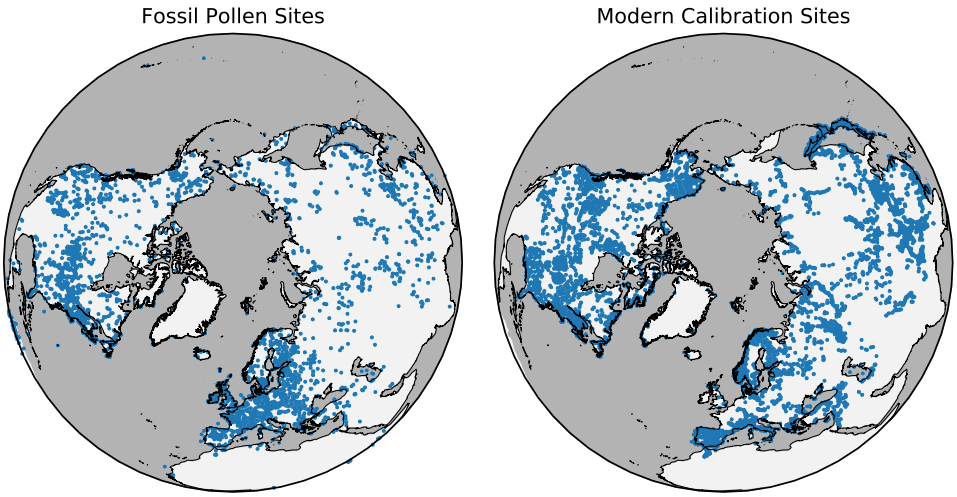


FIGURE 4.3: Maps of (left) fossil and (right) modern pollen sites in the POLNET database.

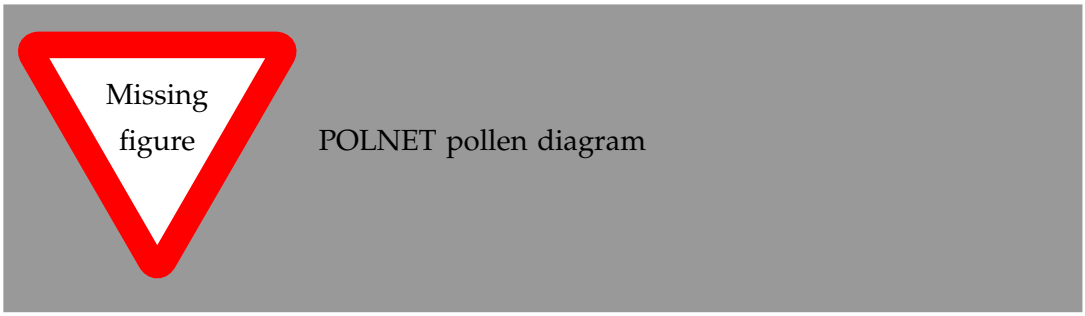


FIGURE 4.4: An exemplary pollen diagram from the POLNET viewer

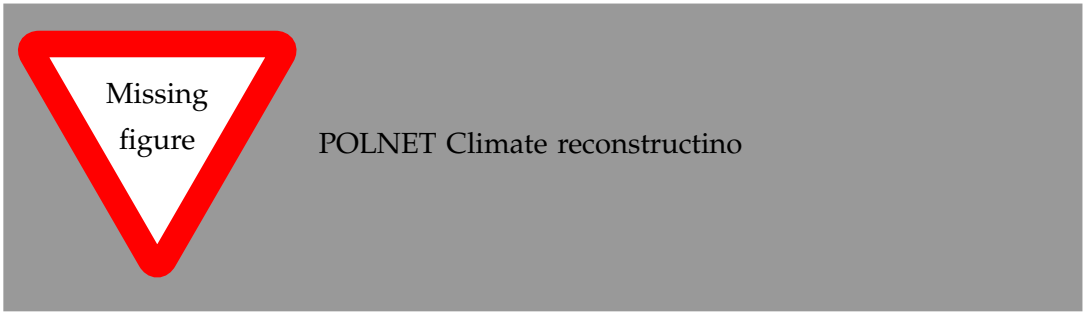


FIGURE 4.5: The climate reconstruction visualized in the POLNET viewer.

is at present the largest existing collection of fossil and modern pollen samples. The database will soon be made publicly available through a dedicated web interface, the POLNET viewer. We present it here as a sample application of the EMPD-viewer to demonstrate how this web interface can be extended and applied to other datasets, in order to make them more accessible.

Like its core application, the EMPD-viewer, the POLNET-viewer is a map-based interface with implemented meta data filters. As it is a data exploration and distribution tool only, we did not include the functionalities to edit the meta data or to submit issues. Instead we implemented new features to visualize the essential aspects of this database: fossil pollen data and climate reconstructions.

The fossil pollen data is loaded upon request from the dedicated Github repository. It is afterwards visualized in form of a stratigraphic pollen diagram, with the age of the samples on the vertical y-axis, and the pollen taxa organized as vertically aligned diagram columns (see figure 4.4).

Climate reconstructions are displayed in two different manners: Climate reconstructions are displayed in two different manners: The site-based reconstructions are visualized as line plots in a separate diagram, together with their associated uncertainties. The gridded temperature reconstruction, i.e. the final product of the database (see also chapter 6) is visualized as an overlay on the map of the web application. This results in a combined visualizations of site-based and gridded reconstructions (see figure 4.5) which enables an intuitive regional analysis of the reconstruction method.

References

Agafonkin, Vladimir and leaflet contributors (2019). *Leaflet - an open-source JavaScript library for mobile-friendly interactive maps*. URL: http://crossfilter.github.io/crossfilter/ (visited on 05/14/2019).

Bayer, Michael (2012). "SQLAlchemy". In: *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. Ed. by Amy Brown and Greg Wilson. aosabook.org. URL: http://aosabook.org/en/sqlalchemy.html.

Bolliet, Timothé, Patrick Brockmann, Valérie Masson-Delmotte, Franck Bassinot, Valérie Daux, Dominique Genty, Amaelle Landais, Marlène Lavrieux, Elisabeth Michel, Pablo Ortega, Camille Risi, Didier M. Roche, Françoise Vimeux, and Claire Waelbroeck (2016). "Water and carbon stable isotope records from natural archives: a new database and interactive online platform for data browsing, visualizing and downloading". In: *Climate of the Past* 12.8, pp. 1693–1719. DOI: 10.5194/cp-12-1693-2016.

Brockmann, Patrick (2016). *ClimateProxiesFinder: dc.js + leaflet application to discover climate proxies*. Last accessed: 2019-08-30. URL: https://github.com/PBrockmann/ClimateProxiesFinder (visited on 10/06/2016).

Davis, B. A. S., M. Zanon, P. Collins, A. Mauri, J. Bakker, D. Barboni, A. Barthelmes, C. Beaudouin, A. E. Bjune, E. Bozilova, R. H. W. Bradshaw, B. A. Brayshay, S. Brewer, E. Brugiapaglia, J. Bunting, S. E. Connor, J. L. de Beaulieu, K. Edwards, A. Ejarque, P. Fall, A. Florenzano, R. Fyfe, D. Galop, M. Giardini, T. Giesecke, M. J. Grant, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuhl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. Guiot, S. Jahns, V. Jankovska, S. Juggins, M. Kahrmann, M. Karpinska-Kolaczek, P. Kolaczek, N. Kuehl, P. Kunes, E. G. Lapteva, S. A. G. Leroy, M. Leydet, J. A. L. Saez, A. Masi, I. Matthias, F. Mazier, V. Meltsov, A. M. Mercuri, Y.

Miras, F. J. G. Mitchell, J. L. Morris, F. Naughton, A. B. Nielsen, E. Novenko, B. Odgaard, E. Ortu, M. V. Overballe-Petersen, H. S. Pardoe, S. M. Peglar, I. A. Pidek, L. Sadori, H. Seppa, E. Severova, H. Shaw, J. Swieta-Musznicka, M. Theuerkauf, S. Tonkov, S. Veski, W. O. van der Knaap, J. F. N. van Leeuwen, J. Woodbridge, M. Zimny, and J. O. Kaplan (2013). "The European Modern Pollen Database (EMPD) project". In: *Vegetation History and Archaeobotany* 22.6, pp. 521–530. ISSN: 0939-6314. DOI: 10.1007/s00334-012-0388-5. URL: http://link.springer.com/article/10.1007/s00334-012-0388-5.

Davis, Basil A. S. and Jed O. Kaplan (2017). *HORNET Holocene Climate Reconstruction for the Northern Hemisphere Extra-tropics*. SNF-Research-Plan. last accessed Jan, 30th, 2018. URL: http://p3.snf.ch/project-169598#.

Davis, Basil A. S., Manuel Chevalier, Philipp S. Sommer, et al. (in prep). "The Eurasian Modern Pollen Database (EMPD), Version 2". In: *Earth System Science Data ESSD*.

Fick, Stephen E. and Robert J. Hijmans (2017). "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". In: *International Journal of Climatology* 37.12, pp. 4302–4315. DOI: 10.1002/joc.5086. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5086. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086.

McKinney, Wes (2010). "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman, pp. 51–56.

Sommer, Philipp S., Basil A. S. Davis, Manuel Chevalier, Jian Ni, and John Tipton (2019). "The HORNET project: applying 'big data' to reconstruct the climate of the Northern Hemisphere during the Holocene". In: *20th Congress of the International Union for Quaternary Research (INQUA)*. International Union for Quaternary Research. URL: https://app.oxfordabstracts.com/events/574/program-app/submission/94623.

Square, Inc. and crossfilter contributors (2019). *Crossfilter - Fast Multidimensional Filtering for Coordinated Views*. URL: http://crossfilter.github.io/crossfilter/ (visited on 05/14/2019).

Zhu, Nick and the dc.js Developers (2019). *dc.js - Dimensional Charting Javascript Library*. URL: https://dc-js.github.io/dc.js/ (visited on 05/14/2019).

Appendices

Todo list