Open Data Web Management Tools for the Eurasian Modern Pollen Database (EMPD) version 2

13 May 2019

Summary

The European Modern Pollen Database (EMPD), version 1, established in 2013 by Davis et al. (2013), is a fully documented and quality-controlled dataset of modern pollen samples. Recent efforts by more than 60 data contributors almost doubled the number of samples in the database and increased it's spatial domain, such that it is now released as the Eurasian Modern Pollen Database, version 2 (Davis et. al, in prep.) with around 8000 samples.

The EMPD is the only public and openly accessible database of modern pollen data in the Eurasian continent and is entirely driven by the community of it's data contributors. This effort of creating an open and accessible database led to the development of new open source data management tools to simplify and reveal the management of multiple contributions from different sources and people. The EMPD2 is now hosted on the version control platform Github at github.com/EMPD2 with a dedicated web viewer at EMPD2.github.io and a automated administration app, the EMPD-admin.

The EMPD web framework

In favor of open science, the hosting on Github allows a fully transparent and comprehensible development of the database, without the need of additional funding for dedicated web services. The implementation on Github additionally provides the advantage of continuous integration services, such as Travis CI (travis-ci.org) and DockerHub (hub.docker.com) to further process and distribute the database. The following sections describe the different tools and integrations in more detail.

The EMPD viewer

The main interface into the EMPD is an interactive web viewer accessible via EMPD2.github.io. This JavaScript-based application provides an interface into the database in an informative way without requiring any particular computer expertise. It enables the user to view the data on a map and select and download subsets of the database. The webpage involves no server-side processing and such it can be hosted for free using on Github allowing stability, independent on the availability of funding. This also allows a local use the viewer with data in a local repository, without the need for further installations of dependencies.

Being inspired by the open source climate proxies finder (Bolliet et al. (2016), Brockmann (2016)), it is mainly based on 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. This allows a quick and efficient filtering of the database. The viewer 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 to review and edit their contributions before they are merged into the database, and to raise issues without the need to have an own Github account.

The EMPD2 data repository

The raw data of the EMPD2 is accessible as plain text files in a Github repository. This allows a transparent traceback of changes made to the EMPD via version control and it allows the EMPD viewer to interface into the database (see previous section). In an automated post-processing, and as an additional check of the database, we also provide a relational postgres dump that is automatically generated based on the plain text files. Meta data tests and a combination with the continuous integration service of Travis CI, standard tools for open source software development, allow a continuous check of the community-based database and facilitates the implementation of new contributions. The management and testing of new contributions through pull requests on Github contributions additionally enables the integration with the automated administration web app EMPD-admin (see next section).

The EMPD-admin

To facilitate the check of new contributions to the EMPD, we developed the EMPD-admin webapp. Inspired by the web management tools of the conda-forge community (Kirkham et al. (2019)), this tool provides an automated handling of data contributions from within Github Pull Requests, including testing, fixing and querying the submission. An additional integration with the EMPD viewer

(see previous section) allows the interactive editing of data contributions and the reporting of issues through the web interface.

The EMPD-admin webapp is hosted for free at Heroku (https://www.heroku.com) at empd-admin.herokuapp.com. This, again, allows stability independent on the availability of funding. It's core functionality can, however, also be installed locally and used from the command-line, independent of Github and Heroku.

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

Accessibility

The EMPD is hosted within the EMPD2 Github organization (https://github.com/EMPD2) at https://github.com/EMPD2/EMPD-data. The source files of the viewer are accessible at https://github.com/EMPD2/EMPD2.github.io and for the EMPD-admin at https://github.com/EMPD2/EMPD-admin.

The EMPD data and the EMPD-admin are additionally both available as Docker images at https://hub.docker.com/u/empd2 and can be accessed via

```
docker pull empd2/empd-data
docker pull empd2/empd-admin
```

The EMPD-admin can also be installed through the python package manager `pip` via

```
pip install EMPD-admin
```

Acknowledgements

We gratefully acknowledge funding by the Swiss National Science Foundation (SNF) through the HORNET project (200021_169598).

References

Agafonkin, Vladimir, and leaflet contributors. 2019. "Leaflet - an Open-Source Javascript Library for Mobile-Friendly Interactive Maps." 2019. http://crossfilter.github.io/crossfilter/.

Bayer, Michael. 2012. "SQLAlchemy." In *The Architecture of Open Source Applications Volume Ii: Structure, Scale, and a Few More Fearless Hacks*, edited

by Amy Brown and Greg Wilson. aosabook.org. http://aosabook.org/en/sqlalchemy.html.

Bolliet, T., P. Brockmann, V. Masson-Delmotte, F. Bassinot, V. Daux, D. Genty, A. Landais, et al. 2016. "Water and Carbon Stable Isotope Records from Natural Archives: A New Database and Interactive Online Platform for Data Browsing, Visualizing and Downloading." *Climate of the Past* 12 (8): 1693–1719. https://doi.org/10.5194/cp-12-1693-2016.

Brockmann, Patrick. 2016. "ClimateProxiesFinder - Dc.js + Leaflet Application to Discover Climate Proxies." 2016. https://github.com/PBrockmann/ClimateProxiesFinder.

Davis, B. A. S., M. Zanon, P. Collins, A. Mauri, J. Bakker, D. Barboni, A. Barthelmes, et al. 2013. "The European Modern Pollen Database (Empd) Project." Journal Article. *Vegetation History and Archaeobotany* 22 (6): 521–30. https://doi.org/10.1007/s00334-012-0388-5.

Kirkham, John, Isuru Fernando, Dougal J. Sutherland, Phil Elson, Marius van Niekerk, Filipe Pires Alvarenga Fernandes, and Eric Dill. 2019. "Conda-Forge-Webservices." 2019. https://github.com/conda-forge/conda-forge-webservices/.

McKinney, Wes, S van der Walt, and J Millman. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, 445:51–56. Austin, TX.

Square, Inc., and crossfilter contributors. 2019. "Crossfilter - Fast Multidimensional Filtering for Coordinated Views." 2019. http://crossfilter.github.io/crossfilter/.

Zhu, Nick, and the dc.js Developers. 2019. "Dc.js - Dimensional Charting Javascript Library." 2019. https://dc-js.github.io/dc.js/.