Technical Note: PaleoDataView - A software toolbox for the collection, homogenization and visualization of marine proxy data

Michael Langner[1], Stefan Mulitza[1]

[1] MARUM – Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

5 *Correspondence to*: Stefan Mulitza (smulitza@marum.de)

Abstract. We present a software toolbox that allows the efficient collection, management and maintenance of larger paleoceanographic data sets. The program combines a graphical user interface (GUI) with a simple document-based database and functionality for visualization, stratigraphy (visual alignment and radiocarbon calibration), age modelling and efficient
10 ensemble time series generation to create larger homogenous data compilations. Data can be imported from Excel or text files, are stored locally in netCDF format and can be easily exchanged between collaborating scientists. Within a data collection, data can either be imported to proxy-specific sub-collections or to a multi-proxy ("miscellaneous") sub-collection that allows to import and manage any downcore data. A single age model is shared among all proxies of a core within a collection. The stand-alone software can be used with Windows and macOS and does not require web access.

1 Motivation

15

Paleoenvironmental data provide valuable information on climate variability and the functioning of the Earth System on timescales not covered by instrumental data. Information on past climates is usually based on proxy data, parameters like stable isotope or elemental ratios of organism remains or sedimentary components that can be translated into climate variables via transfer functions (i.e. Wefer et al. 1999). Proxies are usually measured on samples taken from stratified climate archives (e.g.
20 marine sediments) that allow to derive a time scale through radiometric dating or other stratigraphic methods. Meanwhile proxy data are available in large quantities and with considerable coverage in time and space which allows a detailed spatio-temporal investigation of paleoclimate evolution and variability. When compiled into homogenous and consistent data products with error estimates, paleoclimate data can provide useful information to benchmark climate model experiments. However, the compilation of paleoclimate data is often a difficult and time-consuming process. Paleoclimate data available
25 from public databases (i.e. https://www.pangaea.de/) preserve the state of the data at the time of publication, but usually do not provide a standardized data format that would allow an automatic processing of the data into larger consistent products. In addition, proxy data from public databases are often incomplete or fragmented, and can contain outdated age scales due to the availability of new radiocarbon calibration curves or reference curves that are published after deployment of the data set in the

data base. Furthermore, compilations of paleoceanographic data are often somewhat intransparent and difficult to assess with respect to the quality, resolution and stratigraphic integrity of the included records. Here we provide an overview on the free open-access and cross-platform software tool PaleoDataView (PDV) that allows the efficient collection, homogenization and visualization of marine proxy data. Our primary intention is to support a workflow that allows the creation of continuously

5   updatable and transparent data products, comparable to software (i.e. Ocean Data View, Schlitzer, 2019) or atlas products (i.e. Levitus, 1982) in oceanography.

2 Description

2.1 Software architecture

The program code has been developed in C++ computer language using the Qt Framework (http://www.qt.io/). This integrated

10   development environment allows cross-platform development for Windows and macOS operating systems. A Linux version would in principle be possible, but has not yet been compiled. PDV works with a document-oriented database. After import, data are locally stored in binary form as netCDF files (Rew and Davies, 1990) using the netCDF C++ library (https://www.unidata.ucar.edu/software/netcdf/). This format is machine-independent and self-describing and allows the storage of meta-data. The use of a binary format is necessary to allow the fast reading/writing of data and the efficient storage

15   of large amounts of data, i.e. the age ensembles resulting from Bayesian age-depth modelling. Data can be imported from specifically formatted Excel (*.xlsx) or text files (Figure 1). For each imported record, two types of files are created: A file with the proxy data vs. depth and a file with the stratigraphic/age model data. PDV connects the two file types via the core name. An identical core name for the age model and the proxy file is guaranteed when simultaneously imported from Excel templates, but has to be ensured when files are added manually. Several proxy records can exist for the same core, but only

20   one age/age model file. Hence, when the age model is changed, it changes for all proxy records connected to the age file. To allow fast access to all data in the database (i.e. for plotting purposes), an inventory file containing the meta data of all records in the database is generated/updated at program start or when the proxy type is changed. Since all meta data are stored within the file header, netCDF files generated with PDV can also be added manually to the appropriate folder (i.e. when exchanged between collaborating scientists) and will be added to the inventory when the program is restarted. PDV comes with an installer

25   that extracts the necessary resources to the appropriate folder on your hard disc, including a set of published proxy data for testing purposes (see references and data sources in the data files). The source code of PDV is available under a GNU General Public License, Version 3 and can be unpacked during the installation process. The size of downloadable installers is ~120 MB and the installation of PDV requires about 1GB of free disk space, most of which is required for data resources (i.e. temperature and salinity fields from the World Ocean Atlas).

2.2. Software functionalities

PDV provides specific functionalities needed to streamline the data collection and homogenization process. These include

- the creation of multiple data collections. It is possible to create independent collections and to change between collections/proxy types when using PDV,

- the import and export from/to specifically formatted Excel and text files. Proxy-specific template files are provided with the distribution. Imported data are saved in netCDF files. These files can be exchanged between PDV users and included into other collections (also on other computers with PDV installed) by deploying copies in the respective folder of the PDV data collection,

- an import helper for text files downloaded from the PANGAEA information system (https://www.pangaea.de/). Missing meta data can be added within the software,

10
- a zoomable map interface to quickly explore the availability of data in specific regions. Raw downcore data are plotted when a core position in the map is selected (Figure 2). Alternatively, profile views with a similar functionality are available for major ocean basins,

- an editor for meta data. Meta data can be corrected/added and saved in the netCDF file of the record,

15
- the possibility to comment and flag individual downcore proxy data. Comments (i.e. sediment disturbances) can be added to every sampling depth in a table view,

- the option to make clipboard copies of most of the displayed data via an inbuilt data viewer accessible through the context menu,

- a radiocarbon calibration tool, allowing to calibrate raw radiocarbon dates against IntCal13 (Reimer et al., 2013), taking custom or modelled (Butzin et al., 2017) reservoir ages into account,

20
- a visual alignment tool (comparable to AnalySeries, Paillard et al., 1996) allowing to derive correlation ages from standard reference curves (i.e. Lisiecki and Stern, 2016), also in combination with existing calibrated radiocarbon ages,

- a map interface for the fast visualization/comparison of modern ocean atlas hydrographies for the core sites and the global ocean (Antonov et al. 2009, Locarnini et al. 2009), including proxy relevant parameters like $\delta^{18}O$ of sea water (LeGrande and Schmidt, 2006) and pre-industrial $\delta^{13}C$ of dissolved CO_2 (Eide et al., 2017),

25
- a tool to compare different proxy records in the time or depth domains,

- a direct interface to BACON (Bayesian Age modelling, Blaauw and Christen, 2011), with the option to store/reload age models and BACON parameters in/from netCDF files,

30
- proxy-specific tools to apply transfer function or to correct data. Proxy-specific tools will be activated when a proxy is selected,

- the calculation and visualization of time series ensemble statistics derived from BACON age models.

3 Potential audiences and application scenarios

3.1 Data curation for individual scientist

Individual scientists usually collect their published and unpublished data on their personal computers often in form of Excel or text files, with arbitrary structure. However, the amount of data in Paleoceanography and Marine Geology is increasing

5   exponentially due to more efficient measurement techniques, i.e. through non-destructive downcore scanners. This makes it more and more difficult to keep track of the existing data and their characteristics. Published data should in principle be available in public databases, but these databases are usually static and preserve the state of data and research upon publication. Data have to be downloaded before they can be investigated or modified, for example with respect to stratigraphy. Most scientists prefer to investigate and publish their data before deployment in public databases. This means that it can take years

10   or even decades until newly produced paleoceanographic data are available in the public domain and there is the risk of incomplete documentation or even data loss, i.e. if scientists retire. PDV combines a simple database with visualization and data analysis tools and helps to organize, document, compare and investigate individual data collections. PDV encourages the standardized storage of data at a very early state and thus increases the probability that unpublished data may finally end up in the public domain with the proper documentation.

15   The development of age models for sediment cores is a key technique to produce time series of past climate variability on longer time scales. However, since calibration and reference curves are constantly improved, age models for individual cores need to be revised from time to time. PDV can be employed to efficiently create and update age models based on radiocarbon calibration or graphic alignment to standard reference curves. Ages/age errors from other events (i.e. tephras or biostratigraphic dates) can be added manually in the alignment tool. An added value of PDV is that these dating techniques can be combined

20   and that all details of age model production are preserved. Using the "Miscellaneous" proxy category (see user manual), PDV might be used as a simple dating tool for any downcore parameter.

3.2 Development and maintenance of proxy-specific databases

Millions of proxy measurements are available in public data libraries. For example, the databases PANGAEA

25   (https://www.pangaea.de/) and NOAA (https://www.ncdc.noaa.gov/) host foraminiferal stable isotope and radiocarbon data for thousands of cores, and this database is constantly growing. The combination of a simple document-oriented database with functionality for data processing and documentation allows the efficient collection and transformation of foraminiferal stable isotope and other proxy data into quality controlled and homogenous compilations. Since all modifications are preserved and documented, it is possible to efficiently update the proxy database and the resulting data product. A similar workflow is

30   established in other instrumental data products such as the World Ocean Atlas, where the raw data (World Ocean Database, i.e., Johnson et al., 2009) and the resulting derived products (World Ocean Atlas, i.e., Antonov et al., 2009; Locarnini et al., 2009) are regularly updated as important resources for oceanography and neighbouring disciplines.

Since PDV does not provide a complex relational database, very specific database queries are not possible. However, the advantage of a document-oriented database is the exchangeability of the files among personal computers operated with Windows and macOS. For example, netCDF files can be exchanged among cooperating users of PDV via e-mail or file server. The netCDF format is platform independent and can also be processed with other software. If PDV data collections are
5 published, they can be provided as a zipped collection and later easily be reviewed or extended by other scientists.

3.3 Educational and explorational use

PDV also offers opportunities for educational purposes in Geoscience. Students can interactively explore the regional paleoclimate evolution, compare records from different locations and sedimentation regimes and test stratigraphic methods.
10 Furthermore, PDV also comes with modern ocean atlas hydrographic data and allows to interactively explore the distribution of sea water properties, including proxy-relevant parameters such as $\delta^{13}C_{DIC}$ and $\delta^{18}O$ of sea water.

PDV might also be used as a tool for site survey planning (i.e. at the proposal stage to identity sampling gaps) or to find suitable coring sites before or even during marine expeditions (PDV does not require any web access and does not communicate via the internet). Finally, PDV can be used as a tool to discuss data with fellow scientists on meetings and workshops.

15

4. Concluding remarks and further developments

The combination of a proxy toolbox and a proxy database increases the transparency and reusability of proxy collections and increases the efficiency of the required homogenisation steps. The current version of PDV includes proxy-specific functionalities for foraminiferal isotopes, foraminiferal Mg/Ca and alkenones in addition to a "miscellaneous" proxy category
20 for collections of unspecified downcore records. The "miscellaneous" category is relevant if the user aims at multi-proxy studies on single cores or downcore-parameters for which no proxy-category exists. We plan to extend future versions of PDV to include modules and functionalities more proxies and downcore parameters. The built-in stratigraphic methods can also be extended and refined. The visual alignment with standard records might benefit from the inclusion of an automatic alignment algorithm such as Dynamic Time Warping (Rabiner et al. 1978). Furthermore, a core-to-core alignment of downcore
25 parameters like XRF elemental ratios or susceptibility would help to improve stratigraphies where radiocarbon or stable isotope measurements are not available or associated with large errors. Our goal is to allow a workflow where the proxy values can be saved into organized and quality-controlled atlas products with the possibility to quickly review the raw data foundation. These atlases can then be very useful products for neighbouring disciplines such as paleoclimate modelling.

30 **Code availability.** Installers of the current version of PDV for both Windows 10 and macOS including the C++ code can be downloaded from https://www.marum.de/Stefan-Mulitza/PaleoDataView.html along with a detailed user guide that will be continuously updated. The source code and all resources needed for compilation can be unpacked during the installation process.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Paleoclimate data synthesis and analysis of associated
5 uncertainty (BG/CP/ESSD inter-journal SI)".

Acknowledgements. This work was supported by the German Federal Ministry of Education and Research (BMBF) as a
Research for Sustainability initiative (FONA) through the PalMod project (FKZ: 01LP1509B).

References

10 Antonov, J. I., Seidov, D., Boyer, T. P., Locarnini, R. A., Mishonov, A. V., Garcia, H. E., Baranova, O. K., Zweng, M. M.,
and Johnson, D. R.: World Ocean Atlas 2009, Volume 2: Salinity, NOAA Atlas NESDIS, edited by: Levitus, S., U.S.
Government Printing Office, Washington, D.C., 184 pp., 2010.

Blaauw, M., and Christen, J. A.: Flexible Paleoclimate Age-Depth Models Using an Autoregressive Gamma Process, Bayesian
Anal., 6, 457-474, 10.1214/11-ba618, 2011.

15 Butzin, M., Köhler, P., and Lohmann, G.: Marine radiocarbon reservoir age simulations for the past 50,000years, Geophys.
Res. Lett., 44, 8473-8480, 10.1002/2017gl074688, 2017.Eide, M., Olsen, A., Ninnemann, U. S., and Johannessen, T.: A global
ocean climatology of preindustrial and modern ocean delta C-13, Global Biogeochem. Cy., 31, 515-534,
10.1002/2016gb005472, 2017.

Johnson, D. R., Boyer, T. P., Garcia, H. E., Locarnini, R. A., Baranova, O. K., and Zweng, M. M.: World Ocean Database
20 2009 Documentation, NOAA Printing Office, Silver Spring, MD, 175, 2009.

LeGrande, A. N., and Schmidt, G. A.: Global gridded data set of the oxygen isotopic composition in seawater, Geophys. Res.
Lett., 33, 5, 10.1029/2006gl026011, 2006.

Levitus, S.: Climatological Atlas of the World Ocean, NOAA Professional Paper 13, U.S. Gov. Printing Office, Rockville,
M.D., 190 pp, 1982.

25 Lisiecki, L. E., and Stern, J. V.: Regional and global benthic O-18 stacks for the last glacial cycle, Paleoceanography, 31,
1368-1394, 10.1002/2016pa003002, 2016.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., and Johnson,
D. R.: World Ocean Atlas 2009, Volume 1: Temperature, NOAA Atlas NESDIS, edited by: Levitus, S., U.S. Government
Printing Office, Washington, D.C., 184 pp., 2010.

30 Paillard, D., Labeyrie, L., and Yiou, P.: Macintosh program performs time-series analysis, Eos, Transactions, American
Geophysical Union, 77, 379, 1996.

Rabiner, L. R., Rosenberg, A. E., and Levinson, S. E.: Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition, IEEE Trans. Acoust., Speech, Signal Processing, 26, 575-582, 10.1109/tassp.1978.1163164, 1978.

Reimer, P. J., Bard, E., Bayliss, A., Beck, J. W., Blackwell, P. G., Ramsey, C. B., Buck, C. E., Cheng, H., Edwards, R. L., Friedrich, M., Grootes, P. M., Guilderson, T. P., Haflidason, H., Hajdas, I., Hatte, C., Heaton, T. J., Hoffmann, D. L., Hogg,

5   A. G., Hughen, K. A., Kaiser, K. F., Kromer, B., Manning, S. W., Niu, M., Reimer, R. W., Richards, D. A., Scott, E. M., Southon, J. R., Staff, R. A., Turney, C. S. M., and van der Plicht, J.: IntCal13 and Marine13 Radiocarbon Age Calibration Curves 0–50,000 Years cal BP, Radiocarbon, 55, 1869-1887, 2013.

Rew, R., and Davis, G.: NetCDF: an interface for scientific data access, IEEE Comp. Graphics App., 10, 76-82, 10.1109/38.56302, 1990.

10   Schlitzer, R., Ocean Data View, https://odv.awi.de, 2019.

Wefer, G., Berger, W., Bijma, J., and Fischer, G.: Clues to ocean history: a brief overview of proxies, in: Use of proxies in paleoceanography, Springer, 1-68, 1999.
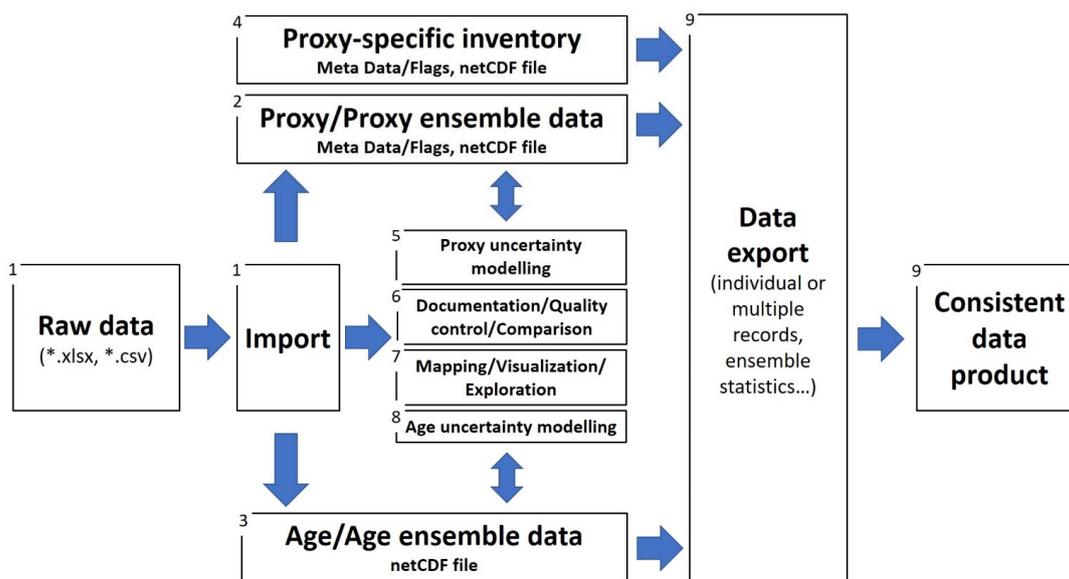
Figure 1: Workflow of the PDV-Toolbox. Data can be imported (1) from Excel or text files and are separated into a proxy
netCDF file containing meta data and downcore proxy data (2) and a netCDF file containing all age-related data (3). An
inventory file is created (4) and constantly updated containing only the meta data of all records in the database. This file allows
fast access to individual files in the database. Functionality within the toolbox includes proxy uncertainty modelling (5),
commenting, flagging and documentation of data/meta data (6) and tools for visualization (7) and age modelling (8). Individual
time series or proxy ensemble time series statistics for individual or a collection of cores can be exported in various file formats
(9).
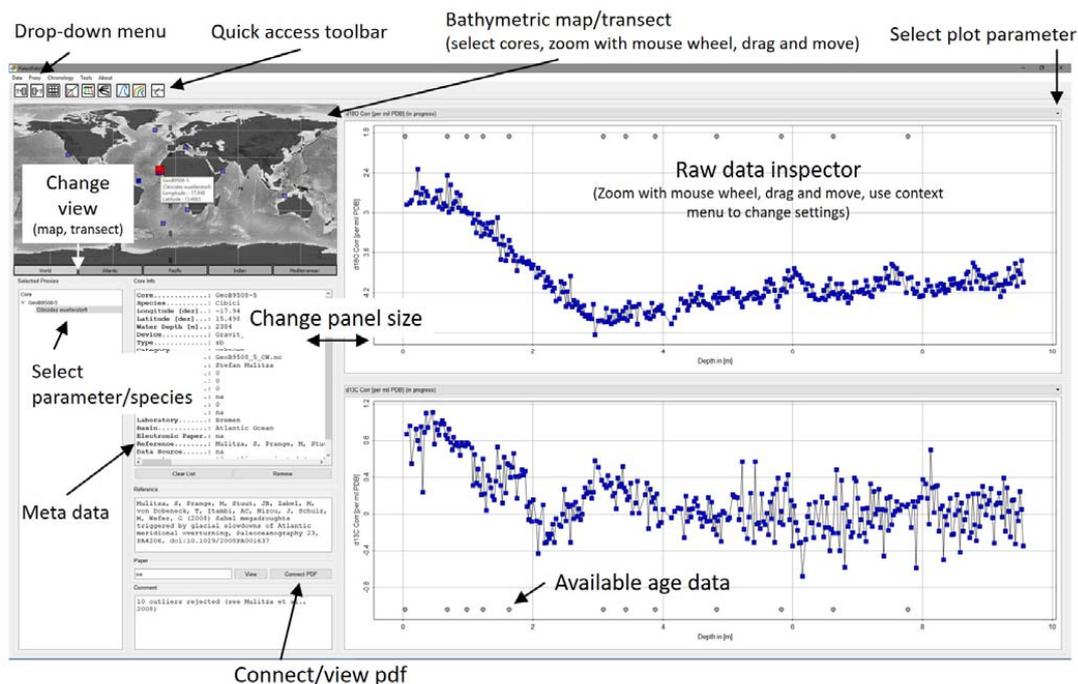
Figure 2: Main window of PDV. Core locations can be selected by pointing and double-clicking with the mouse into the map on the left panel. Both the graphics window and the map are interactively movable and scalable. Available datings/tie points are indicated by cycles at the bottom of the graph. A click on the individual data points reveals depth and isotope value. An available PDF document (i.e. from the corresponding publication) can be connected to the data set and viewed. Further functionality can be accessed through the menu or the symbol bar at the top of the window.