The backbone of PAGES 2k: data management and archiving

Lucien von Gunten, D.M. Anderson[2], B. Chase[1], M. Curran[1], J. Gergis[1], E.P. Gille[2], W. Gross[2], S. Hanhijärvi[1], D.S. Kaufman, T. Kiefer, N.P. McKay[1], I. Mundo[1], R. Neukom[1], M. Sano[1], A. Shah[2], J. Tyler[1], A. Viau[1], S. Wagner[1], E.R. Wahl[2] and D. Willard

[1]PAGES 2k data managers
[2]NOAA Paleoclimatology branch members
Full affiliations listed here: www.pages-igbp.org/products/newsletters/ref2013_2pdf

The PAGES 2k Network and NOAA collaborate closely to optimize data compilations, and to build structures to facilitate ongoing supply and dynamic use of data. It is thus an successful example for a large trans-disciplinary effort leading to added value for the scientific community.

The PAGES 2k Network has formed to study climate change over the last two millennia at a regional scale, based on the most comprehensive dataset of paleoclimate proxy-records possible. In 2011, at its second network meeting in Bern, Switzerland (von Gunten et al. 2012) the network formally acknowledged that the envisioned data-intensive multi-proxy and multi-region study must be built on the foundations of efficient and coordinated data management. In addition, the group committed to PAGES general objective to promote open access to scientific data and called for all records used for, or emerging from the 2k project to be publicly archived upon publication of the related 2k studies.

Architects of the home for 2k data

The National Climatic Data Center at the National Oceanic and Atmospheric Administration (NOAA) offered to host the primary 2k data archive. They set up a dedicated NOAA task force to tailor the 2k data archive to the specific needs of the 2k project and to coordinate archiving with NOAA's data architecture and search capabilities. The 2k groups nominated regional data managers to provide input from the users' end.

Over the last two years, the regional 2k data managers have worked closely with NOAA to tailor the database infrastructure and prepare the upload of the 2k data. In addition, they provided expertise to help promote improvements in NOAA's archival of paleoscientific data in general. Since the data managers of the regional 2k groups are spread across the globe, the collaboration was organized around bi-monthly teleconference meetings under the lead of NOAA. In spite of the occasional unearthly meeting hours for some, the interaction between the 2k data and NOAA database groups has worked fruitfully, as the following achievements show.

The 2k database

The paleoclimatology program at NOAA has set up a dedicated 2k project site with subpages for all regional groups (www.ncdc.noaa.gov/paleo/pages2k/pages-2k-network.html). This page was created early in the project to provide the regional groups with a central place to continuously compile datasets considered relevant to their studies.

Populating the database

A two-step approach was applied for entering the 2k data into the database in order to serve demands for both speediness and thoroughness.

First, all records used for the first synthesis article on regional temperature reconstructions (PAGES 2k Consortium 2013) were made available on a "data synthesis products" page dedicated to the paper (hurricane.ncdc.noaa.gov/pls/paleox/f?p=519:1:::::P1_STUDY_ID:14188). This ensured that the records were made publicly available exactly at the time of publication and in a format that will remain identical with the data files supplementing the article. In a second step, all these records are currently being (re)submitted to NOAA with more detailed metadata information than before using a new submission protocol. Additionally, many new and already stored records that were not used for the PAGES 2k temperature synthesis are (re)formatted to the new submission protocol. This will allow improved search and export capabilities for a wealth of records that can currently only be accessed individually.

Improved data submission protocol

The data submission process is a crucial step for the long-term success of a database. On the one hand, it should contain as much relevant information as possible in order to maximize the value of the data. On the other hand, it should remain simple enough to keep the threshold for data providers as low as possible. The NOAA task force and 2k data managers therefore created a substantially revised submission template file. This new protocol allows including more comprehensive information relating to the proxy records, and, crucially, is organized in a structured format that allows machine reading and automated searching for defined metadata information. This is critical in order to maximize the usefulness of the data to other scientists, as it additionally allows them to reprocess underlying features of the records such as the chronology or proxy calibrations.

The new data submission template is also optimized for taking advantage of NOAA's archival structure, which follows international conventions for data description and archiving, and the Open Archive Initiative Protocol for Metadata Harvesting. This allows PAGES 2k data to be visible beyond the NOAA web site.

A new feature of the NOAA-Paleo archive is the search capabilities that allow for project-specific searches by a logical operator (e.g. "PAGES 2K AND Monsoon"). Additional functionalities are planned that will, for example, allow the user to select a subset of proxy data for a region, and generate a single downloadable file of the requested data in NetCDF, ASCII, or Excel™ formats.

2k data management - next steps

In the next phase of the project starting now, the 2k network will work on completing the database of paleoclimate records of the last 2000 years to eventually produce new synoptic climate reconstructions. The proxy records will be collected according to the new NOAA data submission protocol. The prior use of this template during the collection and analysis phases of the project has the following advantages: 1) all records are collected in the same, uniform format allowing for the inclusion of all relevant information, 2) the files are easily computer readable for data analyses, and 3) no additional formatting is required for the subsequent submission to the NOAA-Paleo archive.

For large, data-intensive studies a good data management strategy is crucial. The experience from the PAGES 2k project suggests that setting up a data manager team and involving archivist partners such as NOAA at an early stage of the project is key to handling data efficiently.

This new collaborative data storing effort is only possible thanks to the members of the regional 2k groups who provide their data and metadata inventories with the aim to make the global network of paleoclimate datasets publicly available.

References

PAGES 2k Consortium (2013) *Nature Geoscience* 6: 339-346
von Gunten L, Wanner H, Kiefer T (2012) *PAGES news* 20(1): 46