doi.org/10.22498/pages.26.2.64

Constituent databases and data stewards in the Neotoma Paleoecology Database: History, growth, and new directions

Eric C. Grimm[1], J.L. Blois[2], T. Giesecke[3], R.W. Graham[4], A.J. Smith[5] and J.W. Williams[6]

The Neotoma Paleoecological Database provides critical cyberinfrastructure for paleoenvironmental research. The database can accommodate virtually any type of fossil data or paleoenvironmental proxy, and is extensible to new data types.

Scientists have long harnessed paleodata to study ecosystem dynamics across time and space. For example, to reconstruct the postglacial expansion of tree species, von Post (1924) assembled fossil-pollen data from across Sweden; Szafer (1935) assembled data from Poland and neighboring areas and invented isopolls to summarize the data; while Firbas (1949) collected data from central Europe north of the Alps, which he summarized in various ways including with isopolls, which he called "Pollenniederschlagskarten" (pollen rain maps). These early investigators assembled, organized, and processed data. In other words, they created "databases", although that term was not yet invented. Their work demonstrated the power of data collections to address emergent questions. With the advent of computers, this power was greatly amplified, for both data management and data analysis.

An early effort to harness computing power was the Cooperative Holocene Mapping Project (COHMAP Members 1988; Wright et al. 1993) in the 1970s, which developed an archive of pollen data as flat files. Many scientists contributed data to this project, which produced numerous publications and spinoff projects. Nevertheless, the data were not publicly available, accompanied by rich metadata, or stored in a relational database. That changed with the advent of the North American Pollen Database (NAPD) in the early 1990s, which was made available for public access by the National Geophysical Data Center of the U.S. National Oceanic and Atmospheric Administration. NAPD was first populated with data from COHMAP, then continued to acquire additional legacy and new data over about 15 years. The European Pollen Database (EPD) was developed simultaneously and in collaboration with NAPD, but the two databases remained separate. The FAUNMAP database, which included Quaternary data from the conterminous United States, was also launched in the early 1990s and made available on floppy disk included with its publication (FAUNMAP Working Group 1994). Following the success of these three databases, other databases were developed for other regions and data types, including the Latin America Pollen Database (LAPD), African Pollen Database, North American Plant Macrofossil Database, North American Non-Marine Ostracode Database (NANODe), Diatom Paleolimnology Data Cooperative, Northern Eurasian Palaeoecological Database, and others.

These database projects assembled large numbers of datasets, involved disciplinary experts, and supported and engendered scientific research. Nevertheless, they suffered from funding lapses and inability to cross-communicate. These issues and others led to the creation of the Neotoma Paleoecology Database (neotomadb.org) following a 2007 workshop at Pennsylvania State University (Williams et al. 2018). This database is named after the rodent genus *Neotoma*, prodigious collectors of diverse materials within their territories and which under the right conditions preserve a multiproxy record of environmental change.

Neotoma provides the underlying cyberinfrastructure for a variety of disciplinary database projects and can accommodate virtually any type of fossil data or paleoenvironmental proxy. All data in Neotoma are stored in a single centralized database but are conceptually organized into virtual constituent databases. These constituent databases, which may be organized according to data type or region, involve disciplinary specialists for data types and regions, thus providing domain scientists with quality control over their portions of the data. Neotoma is a curated resource with governance and control by disciplinary experts. "Curation" implies a high level of quality control. All data added to Neotoma are reviewed and uploaded by

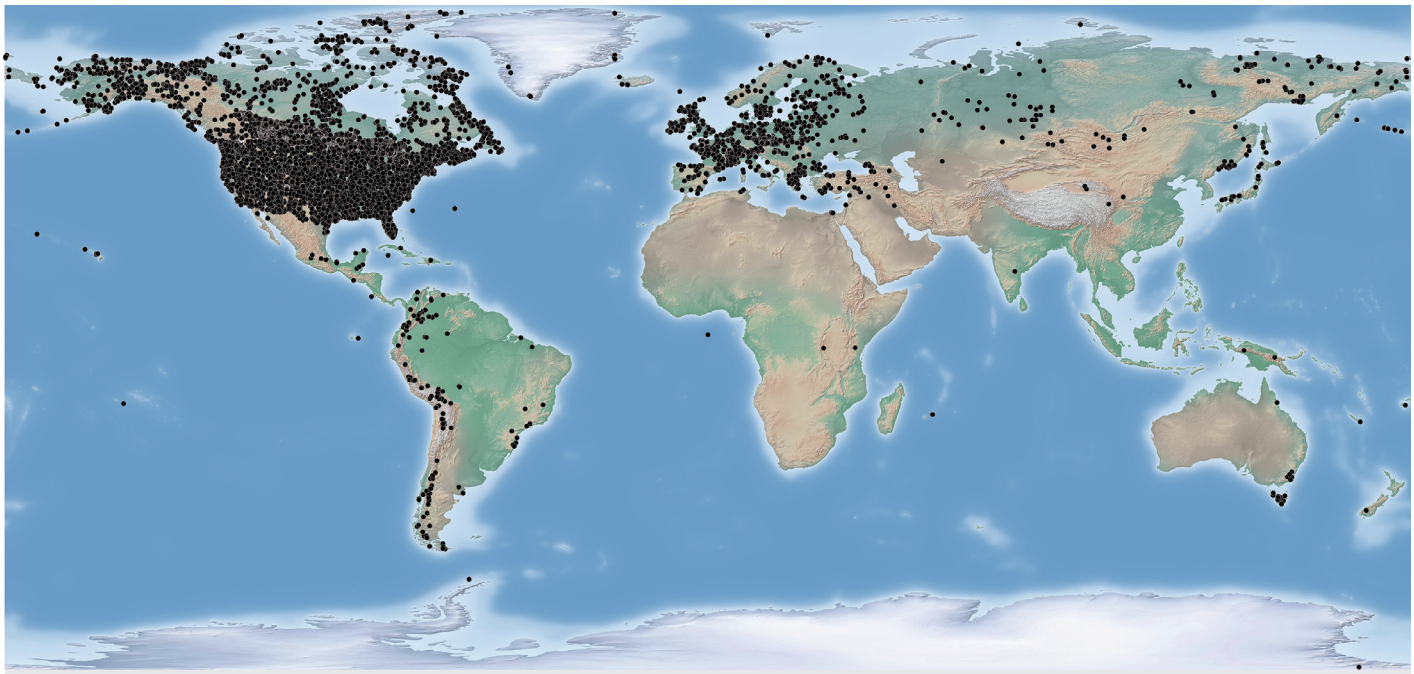

Figure 1: All data holdings in Neotoma: 22,012 datasets from 12,656 sites.

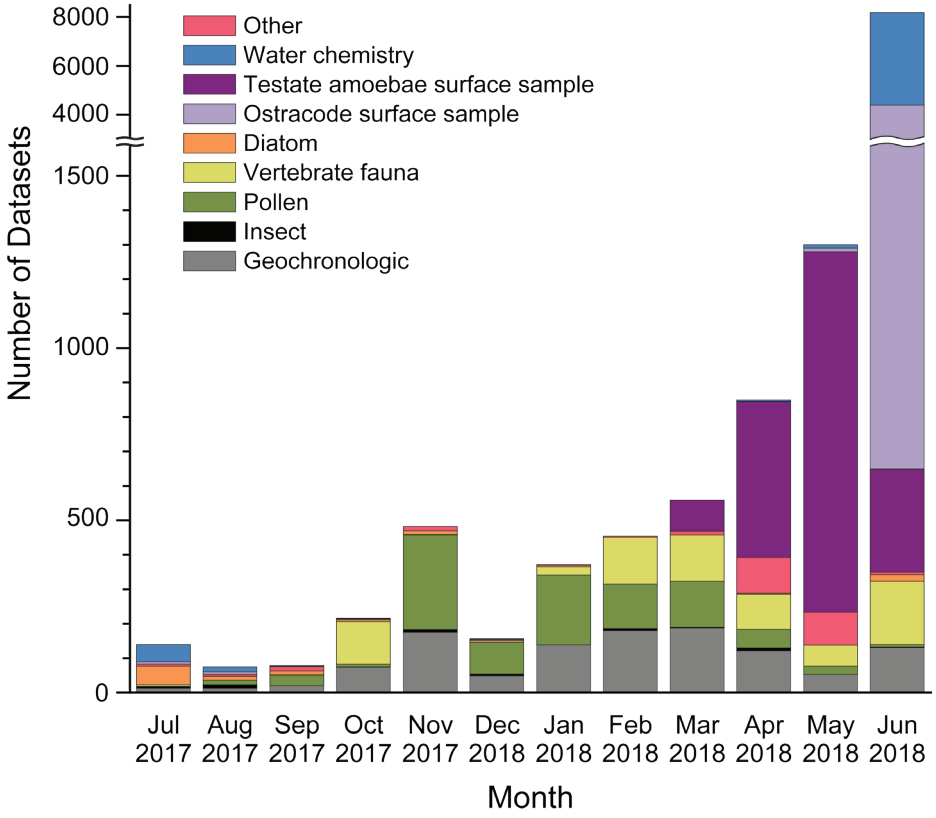

Figure 2: Recent dataset uploads to Neotoma by data type and month.

data stewards, who are appointed by the leaders of the various constituent databases.

Neotoma began by incorporating data from existing databases such as NAPD, EPD, and FAUNMAP. However, Neotoma's design model is flexible and expandable, with many open pathways for participation by new members, data contributors, stewards, and research communities. The Neotoma data model supports, or can be extended to support, any kind of paleoecological or paleoenvironmental data from sedimentary archives (Williams et al. 2018). As of 30 June 2018, Neotoma held 22,012 datasets from 12,656 sites (Fig. 1). New datasets are added almost daily at an increasing rate over the past year (Fig. 2).

Over the past year, there has been considerable push to upload surface samples for aquatic proxies, specifically testate amoebae, diatoms, and ostracodes. The latter two proxies include paired water samples, which comprise "calibration" datasets for quantitative calibration of water chemistry from diatom or ostracode assemblages. As of 30 June 2018, uploads include 1886 testate amoebae surface samples, 640 diatom surface samples, 4515 ostracode surface samples, and 5297 water chemistry samples. The ostracode surface samples have been ported from NANODe and from the Canadian Museum of Nature-Delorme Ostracoda-Surface Samples database. Most of the samples have been uploaded from the Delorme database: 3769 ostracode samples and 3776 water chemistry samples. Although these samples are from other databases, they are not ported en masse, but are subjected to the validation procedures to ensure data quality and compliance with Neotoma metadata standards.

Major efforts have been undertaken to upload data from the EPD (Giesecke et al. 2016) and FAUNMAP 2 (Uhen et al. 2013) databases and to inventory and upload pollen data from Latin America. EPD data contributed before 2007 were included in the Global Pollen Database (GPD), which was available from the World Data Center for Paleoclimatology at NOAA. Following a workshop in November 2017, EPD stewards have uploaded to Neotoma 881 new pollen datasets from 685 sites. After the new datasets are uploaded, the older EPD data ported to Neotoma from GPD will be replaced by the current EPD data, which include many updates and new age models. The original FAUNMAP database (FAUNMAP 1), was an initial compilation into Neotoma. The FAUNMAP 2 database, which includes Canada, Alaska, and the Pliocene (Blancan land mammal age), was compiled but never released nor fully vetted. Since November 2017, Allison Stegner and Mona Colburn have uploaded about a third (1009) of the FAUNMAP 2 datasets to Neotoma. For Latin America, Suzette Flantua and colleagues (Flantua et al. 2013, 2015, 2016) have inventoried pollen and associated geochronological data, and in 2017 over 50 new LAPD pollen datasets from Colombia were uploaded to Neotoma, including important, classic datasets from Thomas van der Hammen and Henry Hooghiemstra.

Another recent improvement particularly relevant for vertebrate fauna, but also other data types, is the ability to store data about individual specimens, including taxonomic and element identification, and museum catalog numbers. Other data can then be associated with these specimens, including radiocarbon dates, GenBank sequence identifiers, and isotopic measurements. In recent years, many high-quality AMS radiocarbon dates on purified collagen have been published (e.g. Widga et al. 2017), and many of these are from sites that are already in Neotoma. These new radiocarbon dates can now be added to existing or new geochronological datasets, and new age models can be built. AMS dates on identified plant macrofossils also comprise another valuable temporal record of taxon occurrences.

The flexible and expandable Neotoma data model has prompted the formation of cooperatives for data types that previously had no appropriate database. Two, in particular, are working groups for stable isotopes and organic biomarkers. The data model of Neotoma has been expanded to accommodate these proxies, and the input software has been modified to upload and validate them. Test datasets have been uploaded, and these holdings should increase during future data mobilization campaigns. We welcome inquiries from researchers interested in contributing data or launching new constituent databases. The continued growth of Neotoma in terms of data holdings and data types will increasingly enable and support paleoenvironmental reconstructions, building upon those first initiated by von Post, Szafer, Firbas, and their contemporaries in the pre-computer era.

AFFILIATIONS

[1]Department of Earth Sciences, University of Minnesota, Minneapolis, USA
[2]School of Natural Sciences, University of California, Merced, USA
[3]Department of Palynology and Climate Dynamics, University of Göttingen, Germany
[4]Department of Geosciences, Pennsylvania State University, State College, USA
[5]Department of Geology, Kent State University, USA
[6]Department of Geography, University of Wisconsin–Madison, USA

CONTACT

Eric C. Grimm: eric.c.grimm@outlook.com

REFERENCES

COHMAP Members (1988) Science 241: 1043-1052

FAUNMAP Working Group (1994) Illinois State Mus Sci Pap 25: 1-690

Firbas F (1949) Spät- und nacheiszeitliche Waldgeschichte Mitteleuropas nördlich der Alpen. G. Fischer, 480 pp

Flantua SGA et al. (2013) PAGES news 21: 88

Flantua SGA et al. (2015) Rev Palaeobot Palynol 223: 104-115

Flantua SGA et al. (2016) Clim Past 12: 387-414

Giesecke T et al. (2016) PAGES Mag 24: 48

Szafer W (1935) Bulletin de l'Académie Polonaise des Sciences et des Lettres, Série B 1: 235-239

Uhen MD et al. (2013) J Vert Paleontol 33: 13-28

von Post L (1924) Geol Fören Förh 46: 83-128

Widga C et al. (2017) Boreas 46: 772-782

Williams JW et al. (2018) Quat Res 89: 156-177

Wright HE Jr et al. (1993) Global climates since the last glacial maximum. University of Minnesota Press, 584 pp