The Neotoma Paleoecology Database: A multi-proxy, international community-curated data resource

Author Names & Affiliations

Williams, John W.[1,2,*]
Grimm, Eric C.[3,*]
Blois, Jessica[4]
Charles, Donald F.[5]
Davis, Edward[6]
Goring, Simon J.[1]
Graham, Russell W.[7]
Smith, Alison J.[8]
Anderson, Michael[9]
Arroyo-Cabrales, Joaquin[10]
Ashworth, Allan C.[11]
Betancourt, Julio L.[12]
Bills, Brian W.[13]
Booth, Robert K.[14]
Buckland, Philip[15]
Curry, B. Brandon[16]
Giesecke, Thomas[17]
Jackson, Stephen T.[18]
Latorre, Claudio[19]
Nichols, Jonathan[20]
Purdum, Timshel[21]
Roth, Robert E.[1, 22]
Stryker, Michael[12]
Takahara, Hikaru[23]

[1]Department of Geography, University of Wisconsin-Madison, Madison, WI 53706
[2]Center for Climatic Research, University of Wisconsin-Madison, Madison, WI 53706
[3]Department of Earth Sciences, University of Minnesota, Minneapolis, MN 55455
[4]School of Natural Sciences, University of California, Merced, CA 95343

[5]Earth and Environmental Science, Drexel University and Patrick Center, Academy of Natural Sciences of Drexel University, Philadelphia, PA 19103
[6]Department of Earth Sciences and Museum of Natural and Cultural History, University of Oregon, Eugene, OR 97403
[7]Department of Geosciences, College of Earth and Mineral Sciences, The Pennsylvania State University, State College, PA 16802
[8]Department of Geology, Kent State University, Kent, OH 44242
[9]SpatialIT, State College, PA 16802
[10]Laboratorio de Arqueozoología, Instituto Nacional de Antropología e Historia, 06060 Ciudad de Mexico, CdMx
[11]Department of Geosciences, North Dakota State University, Fargo, ND 58108
[12]National Research Program, Water Mission Area, U.S. Geological Survey, Reston VA 20192
[13]Center for Environmental Informatics, The Pennsylvania State University, State College, PA 16802
[14]Earth and Environmental Sciences Department, Lehigh University, Bethlehem, PA 18015
[15]Environmental Archaeology Lab, Dept. of Historical, Philosophical & Religious Studies, Umeå University, Umeå, SE-90187
[16]Illinois State Geological Survey, Champaign, IL 61820
[17]Department of Palynology and Climate Dynamics, Albrecht-von-Haller-Institute for Plant Sciences, University of Göttingen, Göttingen
[18]Southwest Climate Science Center, U.S. Geological Survey, Tucson, AZ 85721 and Department of Geosciences, University of Arizona, Tucson, AZ 85721
[19]Departamento de Ecologia, Facultad de Ciencias Biológicas, Pontificia Univeridad Católica de Chile, Casilla 114-D, Santiago and Institute of Ecology & Biodiversity (IEB), Santiago, Chile
[20]Lamont-Doherty Earth Observatory, Palisades, NY 10964
[21]Academy of Natural Sciences of Drexel University, Philadelphia, PA 19103
[22]Cartography Lab, University of Wisconsin-Madison, Madison, WI 53706
[23]Laboratory of Forest Vegetation Dynamics, Kyoto Prefectural University, Hangi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522
*These authors contributed equally to this manuscript.

Corresponding Author
John W. Williams
jww@geography.wisc.edu

Department of Geography
550 North Park St
University of Wisconsin-Madison
Madison, WI 53706
608-265-5537 office
608-265-3994 fax

ABSTRACT

The Neotoma Paleoecology Database is a community-curated data resource that supports interdisciplinary global change research by enabling broad-scale studies of taxon and community diversity, distributions, and dynamics during the large environmental changes of the past. By consolidating many kinds of data into a common repository, Neotoma lowers costs of paleodata management, makes paleoecological data openly available, and offers a high-quality, curated resource. Neotoma's distributed scientific governance model is flexible and scalable, 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. Data additions to Neotoma are growing and now include >3.8 million observations, >17,000 datasets, and >9,200 sites. Dataset types currently include fossil pollen, vertebrates, diatoms, ostracodes, macroinvertebrates, plant macrofossils, insects, testate amoebae, geochronological data, and the recently added organic biomarkers, stable isotopes, and specimen-level data. Multiple avenues exist to obtain Neotoma data, including the Explorer map-based interface, an Application Programming Interface, the *neotoma* R package, and digital object identifiers. As the volume and variety of scientific data grow, community-curated data resources such as Neotoma have become foundational infrastructure for big data science.

KEYWORDS

Biogeography; Geoinformatics; Global; Micropaleontology; Paleoclimatology; Paleodatabases; Paleoecology; Paleoecoinformatics; Paleolimnology; Paleontology

INTRODUCTION

The Neotoma Paleoecology Database (hereafter called Neotoma, www.neotomadb.org) was launched in 2009 with a mission to provide an open, community-curated, sustainable, and high-quality repository for multiple kinds of paleoecological data. Neotoma's name refers to the behavior of woodrats or packrats (genus *Neotoma*), which (inadvertently) serve paleoecology by gathering diverse biological materials into their nests, there to be preserved for future generations. Although Neotoma itself is relatively young, it builds upon decades of effort by paleoecologists, paleoclimatologists, and paleontologists to gather individual records into larger spatial networks for the purpose of studying ecological, evolutionary, biogeographic, climatic, and cultural processes at spatial scales beyond the scope of any single site-level paleoecological record. The gathering of these records is expensive, with substantial investments in money and time, and their scientific value is multiplied when aggregated into larger networks. By bringing these resources into a single open data resource with an accompanying distributed governance framework, Neotoma seeks to accelerate our capacity to do global-scale paleoscience, serve as an open-source platform for new kinds of analytics and visualizations, enhance reproducibility, and increase the longevity and sustainability of our communities' hard-won data.

Neotoma's creation and design is motivated by the goal of enabling global-scale science from long-term, site-level data. Ecological processes operate across a wide range of interacting spatial and temporal scales (Heffernan et al., 2014). Dynamics at one location are often interpretable only in the context of larger-scale biogeographic and climatic processes (Webb, 1997), and ecosystems can be affected by slow processes that were triggered by events centuries or even millennia ago (Goring and Williams, 2017; Svenning and Sandel, 2013). Networks of paleoecological records, therefore, provide fundamental scientific infrastructure for understanding the responses of species to large and abrupt environmental changes, the mechanisms that promote resilience, and the interplay between climatic and biotic interactions (Blois et al., 2013; Dawson et al., 2011; Jackson and Blois, 2015; Moritz and Agudo, 2013). Examples include the processes controlling contemporary and past patterns of community, species, and genetic diversity (Blarquez et al., 2014; Cinget et al., 2015; De La Torre et al., 2014; Fritz et al., 2013; Gutiérrez-García et al., 2014; Jezkova et al., 2015; Sandom et al., 2014); identification of species refugia (Bennett and Provan, 2008; Gavin et al., 2014; Vickers

and Buckland, 2015); rates of species expansion (Giesecke et al., 2017; Ordonez and Williams, 2013); the reshuffling of species into no-analog communities during climate change (Finsinger et al., 2017; Graham et al., 1996; Radeloff et al., 2015); the timing and patterns of abrupt ecological and climate change (Seddon et al., 2015; Shuman et al., 2009), quantification of the time lags between abrupt climate change and local ecological response (Ammann et al., 2013; Birks, 2015); and the timing, causes, and consequences of late-Quaternary megafaunal extinctions (Doughty et al., 2013; Emery-Wetherell et al., in press; Lorenzen et al., 2011).

The scientific communities interested in paleoecological data extend well beyond paleoecology and biogeography. Paleoecological data such as fossil pollen, diatoms, and marine foraminifera are the backbone of continental- to global-scale paleoclimatic reconstructions developed to benchmark climate models and assess feedbacks within the earth system (Bartlein et al., 2011; CLIMAP Project Members, 1976; Marcott et al., 2013; MARGO Project Members, 2009; Shakun et al., 2012; Trouet et al., 2013; Viau et al., 2012; Wright et al., 1993) and constrain estimates of climate sensitivity (Schmittner et al., 2011). Paleoecological data help establish ecosystem baselines and trajectories for managers seeking to conserve species and ecosystems of concern (Barnosky et al., 2017; Clarke and Lynch, 2016; Dietl et al., 2015; Panagiotakopulu and Buchan, 2015; Whitehouse et al., 2008). Similarly, paleoecological data are necessary for understanding the interactions between past environmental change and early human evolution, land use, cultural and technological innovation, and dispersal at local to global scales (deMenocal, 2001; Ellis et al., 2013; Gaillard et al., 2010; Grant et al., 2014; Kaplan et al., 2011; Kaplan et al., 2009; Muñoz et al., 2010).

In response to these scientific drivers, multiple databases have been developed by multiple teams over the past 30 years for different kinds of Pliocene-Quaternary fossil data. These prior efforts, beginning in North America and Europe in the 1970s (Bernabo and Webb, 1977; Davis, 1976; Grimm et al., 2013; Huntley and Birks, 1983; Sadler et al., 1992), resulted in multiple paleoecological databases, each usually restricted to a particular dataset type or region, e.g. the North American, European, African, and Latin American Pollen Databases (Flantua et al., 2015; Fyfe et al., 2009; Grimm et al., 2013; Vincens et al., 2007); the FAUNMAP and MIOMAP terrestrial vertebrate databases (Carrasco et al., 2007; Graham et al., 1996), the Mexican Quaternary Mammal Database (MQMD, Arroyo-Cabrales et al., 2007, 2009), NANODe (Forester et al., 2005), the Diatom Paleolimnology Data Cooperative (https://diatom.ansp.org/dpdc/) (Sullivan and Charles, 1994), the North American packrat midden (Betancourt et al., 1990) and plant macrofossil databases (Jackson et al., 2000; Jackson et al., 1997), the Base de Données Polliniques et Macrofossiles du Quebec (Richard,

1995), the BUGS insect database (Sadler et al., 1992), and others. Other, newer proxies, such as testate amoebae and organic biomarkers, are just beginning to be gathered for use in broad-scale studies and need a common platform for data archival, sharing, and re-use.

The importance of these databases cannot be overstated. Hundreds of scientific papers have utilized them (Fig. 1); and entire dissertations and subsequent papers have been based on them (e.g. Buckland, 2007; Li, 2004; Lyons, 2001). Nevertheless, the various paleoecology databases were typically established with either one-time or sporadically funded projects. Long-term maintenance, sustainability, and development have plagued virtually all paleoecological and paleoclimatic database efforts mainly because continuous funding is needed for both information technology (IT) and data preparation and cleaning. Some databases have had curatorial support from museums and government entities (e.g. Canada Museum of Nature's curation of Delorme ostracode database), others have not. Funding hiatuses rarely cause these databases to disappear entirely, but have caused data backlogs and delays, with long delays between data contributions by individual scientists and their release to the public. Additionally, early databases were maintained in flat files or standalone, desktop-database management systems such as Paradox or Microsoft Access (Grimm et al., 2013), versus the new standard of client/server systems that serve data over the internet. Copies of databases sometimes proliferated and individual copies rapidly became obsolete (e.g. the initial release of FAUNMAP was distributed by hard copy and floppy disks, as well as on-line, FAUNMAP Working Group, 1994). Because paleoecological data were dispersed across different resources, with differing data architectures and degrees of accessibility, it has been difficult to synthesize data across resources.

Neotoma builds on these prior efforts by providing 1) a consolidated and hence more cost-efficient and sustainable IT structure, 2) an open and flexible data model based on decades of experience with paleoecological data, and 3) a distributed and extensible governance structure that promotes high-quality, curated data and establishes pathways for new scientists and research groups to join and contribute. The Neotoma data model supports, or can be extended to support, any kind of paleoecological or paleoenvironmental data from sedimentary archives. Neotoma has focused on the Quaternary to Miocene section of the geological record, and primarily supports research about ecological processes operating at timescales of 10^2 to 10^6 years, but there is no hard limit to Neotoma's temporal extent. Neotoma employs a distributed governance structure based upon Constituent Databases (see *Governance and Data Use*), because the scientific expertise necessary for curating paleoecological data is widely distributed across the scientific community. Each Constituent

Database corresponds to a particular dataset type (e.g. ostracodes, insects, vertebrates, diatoms, pollen, organic biomarkers) or region and each is led by expert Data Stewards (list of Stewards available at http://bit.ly/2tzjEsZ), with opportunities available for scientists who would like to contribute their data and become Stewards, either for existing Constituent Databases or to launch new ones. Neotoma also acts as a boundary organization and translator (Fig. F2, Guston, 2001) among multiple interacting groups, bridging across the field- and lab-oriented communities that contribute data to Neotoma, the diverse multiple research and educational communities that use paleoecological data, and the informatics communities that build systems for translating big data to knowledge.

Neotoma is part of the emerging field of paleoecoinformatics (Brewer et al., 2012), which itself contributes to larger efforts in bioinformatics and geoinformatics to overcome bottlenecks associated with data access, mobilize dark data, and maximize the power of scientific data collected by networks of researchers (Ferguson et al., 2014; Hampton et al., 2013; Heidorn, 2008; Howe et al., 2008; Lynch, 2008). Related efforts include the Paleobiology Database (https://paleobiodb.org/#/), the International Tree Ring Databank (https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/tree-ring), the Paleoclimatology data holdings at the NOAA National Centers for Environmental Information (https://www.ncdc.noaa.gov/data-access/paleoclimatology-data), Pangaea (https://www.pangaea.de/), iDigBio and iDigPaleo (https://www.idigpaleo.org/), MorphoBank (https://morphobank.org/), the Limnological Research Center and Continental Scientific Drilling Office (https://csdco.umn.edu/), the Interdisciplinary Earth Data Alliance (http://www.iedadata.org/), and The Strategic Environmental Archaeology Database (http://www.sead.se). Interlinking efforts are also underway through e.g. NSF's EarthCube program.

Neotoma also seeks to support and leverage on-going advances in paleosciences and data sciences. In paleoecology and paleobiology, the rate of publications is increasing exponentially (Uhen et al., 2013), which requires scalable informatic solutions; new paleoecological proxies continue to emerge (e.g. organic biomarkers and compound-specific stable isotopes, Bush and McInerney, 2013; Sachse et al., 2012; Zhang et al., 2006); radiometric dating techniques and age modeling software continue to improve, enabling more precise ecological inferences (e.g. Zazula et al., 2014); and the growth of data assimilation and ecological forecasting approaches is requiring a closer and interactive coupling between ecological data and mechanistic models (Dietze, 2017; Dietze et al., 2012). Relevant advances in the data sciences include the development of efficient protocols (e.g. JSON, SPARQL) that

enable the establishment of networks of reliable and linked distributed resources across the internet (https://www.w3.org/blog/SW/2008/01/15/sparql_is_a_recommendation/) (ECMA International, 2013), the development of international standards for unique data identifiers (e.g. DOI, www.datacite.org), and the rapid advances in open-source and collaborative programming environments such as R, GitHub, or Jupyter that place state-of-the-art analytical tools at the fingertips of scientists (Goring et al., 2015; Ohri, 2014). Neotoma also supports the now-common requirement by funding agencies and journals for scientific data to be made publicly available (Nature, 2017).

Here we provide an overview of the Neotoma Paleoecology Database, its design principles, key concepts, software ecosystem, governance structure, and a snapshot of current data holdings and uploads. We summarize the structure of the Neotoma data model and the available capabilities for uploading, finding, exploring, analyzing, and validating data. A supplementary section points readers to additional information about the technical details of implementation, available in open-source code repositories (www.github.com/NeotomaDB), several on-line help manuals, and other publications (Goring et al., 2015; Grimm et al., 2014). We begin by describing key design principles and core semantic concepts that underlie the Neotoma data model, software design, and governance structure. Finally, we highlight ways that interested scientists can contribute to building this community resource, and we look ahead to current opportunities, challenges, and their solutions.

NEOTOMA DATABASE: DESIGN PRINCIPLES

Seven core principles and philosophies govern Neotoma's design.

1) Neotoma is a Spatiotemporal Paleoecological Database: The core mission of Neotoma is to store and openly share well-organized and curated information about the past occurrences and abundances of organisms, their geobiological signatures, and associated paleoenvironmental variables, in space and time (see *Data Holdings and Data Types* for a fuller listing). Informed interpretation of these data also requires Neotoma to store and curate information about the geographic and sedimentary characteristics of the field site and stratigraphic horizons from which fossils were collected or variables measured, age controls and age-depth models used to estimate time, and the identity and contact information of investigators. Other kinds of information are also relevant to informed paleoecological interpretation, but fall outside Neotoma's core mission and may be best curated by other communities and data resources, e.g. ecological traits, digital images of fossils, contemporary

genetic data, archaeological excavation data, reconstructions of past sea level and paleogeography, and paleoclimatic simulations from earth system models. Neotoma's data and governance models are designed to be flexible and extensible to other paleoecological or paleoenvironmental proxies.

2) Neotoma Consolidates IT, Distributes Scientific Governance: Neotoma combines a centralized database structure with a system of distributed scientific governance. All Neotoma data are housed in a single relational database (see *Technical Specifications and Software Ecosystem*) and organized according to a common set of core semantic concepts (see *Neotoma Data Model: Fundamental Concepts*). The use of one data structure for multiple paleoecological proxies reduces developer support costs, facilitates data discoverability and reuse, and increases data interoperability. Neotoma comprises virtual Constituent Databases, each encompassing a particular dataset type or region (e.g. North American Pollen Database, European Pollen Database, North American Non-marine Ostracode Database, FAUNMAP, Bugs) and each with its own Data Stewards (Fig. 3). Constituent Databases and allied cyberinfrastructure resources can also create their own front-end portals into Neotoma through the use of the Neotoma Application Programming Interface (API; see *Technical Specifications and Software Ecosystem*). New Constituent Databases can be created to bring in new data types and regions (see *Governance and Data Use Policy*) and new Data Stewards can be readily trained (see *Governance* and *Next Steps* sections). This distributed governance structure addresses the challenge of distributed scientific expertise and allows Neotoma to scale as new records and proxy types are added to it.

3) Neotoma is a Community-Curated Data Resource. Neotoma's Data Stewards are functionally similar to a journal's Board of Editors, charged with ensuring that data stored in Neotoma conform to community-established data standards. Data input into Neotoma is led by trained Data Stewards and Data Processors (see *Governance and Data Use Policy*) appointed by their communities, with taxonomic names approved by Taxonomic Experts. Distributed scientific governance is essential for Neotoma because no single individual or institution can be expert in all the dataset types, regions, and time periods represented within Neotoma. Software systems provide Stewards with automated tools to check for data inconsistencies, metadata completeness, and taxonomic conformity with Neotoma standards prior to uploading (see *Technical Specifications and Software Ecosystem*). This process of data validation, standardization, and cleaning adds significant value (Lehnert and Hsu, 2015) and distinguishes Neotoma from general-purpose and comprehensive data depositories such as DataDryad or FigShare.

4) Neotoma Data are Open: Neotoma data are available to anyone with an internet connection and are accessible through several interfaces (Fig. 4), each serving distinct user communities. Neotoma uses a CC-BY license, allowing free reuse of data with proper attribution to Neotoma and the original data contributors (see *Governance and Data Use Policy*, http://www.neotomadb.org/data/category/use). By supporting open data, Neotoma prevents the underuse and eventual loss of valuable paleodata that languish on individual computers (Hampton et al., 2013; Heidorn, 2008) while promoting scientific transparency and reproducibility across the community. Neotoma has an embargo policy for data contributed to Neotoma prior to publication (*Governance and Data Use*) and is working on an embargo management system, to promote good data management practice of organizing data early in a project cycle, rather than at the moment of publication, when details may be difficult to recollect. This embargo will also allow contributors to analyze their data in the context of the whole database without releasing it for public access until publication. Neotoma supports data archival and management plans required by many funding agencies and journals.

5) Neotoma is a Living Database. The life cycle of a paleoecological dataset does not end with its first publication. Data are reused, as the original investigators and new teams synthesize existing data to answer new questions. Errors may be caught and corrected during subsequent syntheses (e.g. inaccurate transcriptions of geographic coordinates, inaccurate or incomplete capture of all fossil data from a site, missing metadata). Derived inferences may be updated with newer analytical methods (e.g. newer Bayesian age models, Blaauw and Christen, 2011; Parnell et al., 2008). Neotoma stores relatively stable raw data (e.g. number of fossil specimens of a species in an assemblage, radiocarbon dates) as well as derived data (e.g. age models), subject to change as scientific understanding advances. Of these, the most changeable data tend to be age estimates and taxonomic names. Age estimates will change as dating techniques and age-depth models improve. Taxonomic names may change if taxonomic identifications of individual specimens are revised or if taxonomic nomenclature is revised or updated.

6) Community Engagement and Empowerment are Essential. Neotoma serves diverse communities (Figs. 2, 3). The constituent groups are essential to Neotoma's ability to grow and scale upwards. Hence, Neotoma continually seeks to enlarge and support its community of Data Stewards, contributors, and third-party developers (see *Governance* and *Next Steps* sections). The sustainability of Neotoma is ultimately determined by the degree to which it supports key research priorities of these users.

7) Neotoma is Part of a Larger Ecosystem. Many communities are gathering and assembling their data, while others establish standards, vocabularies, and systems for sharing data across systems. Data types are many and Neotoma's resources are few. Hence, whenever possible, Neotoma will partner with other allied resources. For example, other organizations have set standards for storing and representing information about individual investigators (ORCID, https://orcid.org/ , physical samples (IGSNs, http://www.geosamples.org/igsnabout), geospatial data (Open Geospatial Consortium, http://www.opengeospatial.org/), or funding agencies (FundRef registry built by CrossRef, https://www.crossref.org/services/funder-registry/); Neotoma is adopting or moving towards adopting these common standards. Similarly, Neotoma seeks to develop partnerships to intersect paleoecological and paleoenvironmental data with other kinds of climatic, archaeological, and ecological data (see *Next Steps*).

NEOTOMA DATA MODEL: FUNDAMENTAL CONCEPTS

This section describes high-level semantic concepts embedded within Neotoma's data model. We first describe Neotoma's system for representing and storing the many kinds of sedimentary sampling designs used by paleoecologists, then describe the kinds of information linked to variables, and finally Neotoma's system for representing time. We do not attempt to describe in detail Neotoma's relational database structure because most scientific users do not come into direct contact with the actual relational database. However, many of the concepts described here correspond to one or more data tables in Neotoma's relational database. For interested scientists and developers, further information is available in the Neotoma Database Manual (http://www.neotomadb.org/uploads/NeotomaManual.pdf).

Sites, Collection Units, Analysis Units, Samples, and Datasets

Paleoecological data from sedimentary archives have many commonalities: They typically involve measurements of fossil organisms or *proxies* found in various geological *archives* along some spatial direction, usually vertical *depth*, for which we estimate *time* with uncertainty (see also Evans et al., 2013). These commonalities enable a common data model. Within this general framework, many sampling systems exist, that vary within and among subdisciplines and depositional environments. For example, paleolimnologists may collect one or more sediment cores from a lake, with multiple kinds of measurements made on the cores and subsamples from it; archaeologists may collect botanical or faunal specimens from surface

scatter or excavations; or vertebrate paleontologists may measure stable isotopes on bones collected from a sediment section or cave deposit. In order to flexibly store data from these different sampling methods, Neotoma uses a hierarchical arrangement of ***sites***, ***collection units***, ***analysis units***, ***samples***, and ***specimens,*** with samples further grouped by type into ***datasets*** (Fig. F5, Grimm et al., 2013).

A ***site*** is a geographic place from which paleoecological data have been collected. Examples of sites include lakes, caves, archaeological excavations, and stratigraphic sections. The spatial extent of sites is flexible, and tends to be defined based on field practice, ranging from a single point to a lake or archaeological site with an extent measured in hectares. Key properties include name, geographic coordinates, altitude, and areal extent. In Neotoma, the spatial extent of sites is represented by bounding boxes with north and south latitudes and east and west longitudes. The bounding box can circumscribe the site, (e.g. a lake) or may circumscribe a larger area containing the site, either because site location is imprecisely known (e.g. described as 'on a gravel bar 5 miles east of town') or because location is purposely kept vague (e.g. to prevent looting and vandalism). Many legacy sites in Neotoma have point coordinates. A site will have one to many collection units.

A ***collection unit*** is a place within a site from which a set of fossil specimens or samples has been collected. Typical collection units include individual or composite cores from lakes and peatlands, profiles from stratigraphic sections (e.g. river cutbanks, quarry walls), archaeological or paleontological excavation features or contexts, isolated specimens (e.g. a bone collected from a gravel bar), and surface samples. Collection units typically have spatial Cartesian or geographic coordinates within a site. Collection unit properties include name, latitude-longitude coordinates (represented as point coordinates with error), altitude, date of collection, and metadata about collection methods and the depositional environment. The definition of collection units and their spatial extent is flexible. For example, in a pit cave with three fossil-rich sediment cones, each with several excavation squares, the collection units could be defined as the individual squares, or as three composite collection units, one from each sediment cone. In another example, consider a lake with two closely adjacent cores from the lake center, plus a single core near the lake margin. In this case, the two central cores might be merged into one composite collection unit, with a second collection unit representing the lake-margin core. A collection unit will have one to many analysis units.

An ***analysis unit*** is a unique sampling location within a collection unit. Analysis units are typically arrayed along a depth transect, usually oriented vertically. Analysis units may be identified by depth and optionally thickness or by name and optionally ordinal position. Analysis

units may be arbitrary intervals or natural strata. Examples of analysis units include individual depth intervals along a sediment core or within an excavation, individual strata or features within an excavation, individual measurements taken along a transect from outer surface to inner core of a speleothem, etc. Natural strata may be vertically superimposed, but may vary in depth and thickness within a section or excavation, particularly in colluvial sections and sediment cones below pit-cave openings. In this case, the "depths" may be ordinal positions or pseudo-depths. However, in some cases, however, analysis units may be single-context features (e.g. archaeological hearths and storage pits), which may be identified by name only, with no explicit depth. An analysis unit will have one to many samples.

A ***sample*** is a single set of measurements of a single dataset type from an analysis unit. Dataset types usually correspond to taxonomic groups that are the loci of scientific expertise and Neotoma's Constituent Databases, e.g. diatoms, ostracodes, pollen, plant macrofossils, and terrestrial vertebrates. Samples and dataset types can also comprise geochemical measurements (e.g. stable isotopes, organic biomarkers), physical measurements (e.g. loss on ignition), or geochronological measurements (e.g. ^{14}C or ^{210}Pb dates). For example, the same analysis unit from a stratigraphic section may have a vertebrate sample and a macrobotanical sample; or an analysis unit from a sediment core may have pollen, diatom, ostracode, and stable isotope samples. The analysis unit links together samples located in the same stratigraphic interval, while datasets links samples from the same collection unit.

Datasets comprise all samples of the same data type from a collection unit. Datasets are typically the subjects of publication and are a primary mode in which data within Neotoma are packaged for delivery to users. For example, clicking on a site in Neotoma Explorer (http://apps.neotomadb.org/explorer/) will return a list of all datasets at that site. Similarly, in the *neotoma* package in R, Neotoma data are primarily passed to R in the form of datasets, using the *get_datasets* and *get_download* functions (which respectively return dataset metadata and data, Goring et al., 2015). Similarly, digital object identifiers (DOIs) are assigned to datasets, but not to samples or analysis units (http://data.neotomadb.org/datasets/5000/index.html).

A ***specimen*** is the physical form of a biological object – e.g. a vertebrate bone or other fossil – retrieved from a sample. Specimens are often curated and housed at museums, geological surveys, or other repositories. Specimens often have catalogue numbers, accession numbers, or other unique identifier assigned by their institution. In 2016-2017, the Neotoma data model was extended to store specimen-level measurements, e.g. radiocarbon dates and stable isotopic measurements from individual teeth or bones. Neotoma does not currently store information about specimen morphometric traits, but the data model could be extended in this

direction, or Neotoma could link out to other databases that store specimen- or species-level trait data (e.g. iDigBio, MorphoBank).

Variables

In Neotoma, variables store information about measured organisms or proxies of any type. Variables have the property ***taxon name*** (equivalent to variable name) and the optional properties ***element***, ***units***, and ***context***.

In Neotoma, **taxon name** is used in the broad sense to include both organismal taxa and physical "taxa" such as stable isotopes, organic biomarker compounds, and inorganic minerals. Neotoma uses defined vocabularies of taxon names and, during the validation process of uploading data to Neotoma, taxon names in uploaded files are automatically checked and flagged if there is no match. New names may be proposed by Stewards and approved by Taxonomic Experts (see *Governance and Data Use Policy*). Taxon names for organisms can include non-Latin modifiers to indicate the level of uncertainty in the taxon identification. For example, *Ambrosia*, *Ambrosia*-type, and cf. *Ambrosia* are three different taxa. The uncertainty modifiers are included in the taxon name, rather than in a separate field, so as to indicate the exact level of uncertainty and to faithfully record the original identification, in which the uncertainty is usually included as part of the name. Thus, *Odocoileus* cf. *O. virginianus* indicates that the genus identification is secure, but the species is uncertain; whereas, cf. *Odocoileus virginianus* indicates that the genus identification is uncertain. This example might be the case for regions in which *Odocoileus virginianus* (white-tailed deer) is the only *Odocoileus* species biogeographically reasonable, thus if it is an *Odocoileus*, it must be *Odocoileus virginianus*. The "cf." designation may occur at any number of taxonomic ranks or cladistic nodes. With the exception of the uncertainty modifiers (cf., aff., sp., spp., undiff., -type, ?), non-Latin modifiers are included in parentheses, e.g. Poaceae (<50 µm), Leporidae (large), Eudicotyledoneae (tricolpate, Hooghiemstra 1984 type 152). The forward slash symbol is used to indicate identifications limited to a small number of taxa (usually two) with elements that cannot be differentiated, e.g. *Ostrya/Carpinus*. This packaging of taxon name and uncertainty into a single field carries the advantage of staying true to paleontological tradition and nomenclature, but hinders integration with contemporary biodiversity databases.

If taxonomic names used by data contributors or publications are changed, the original name can be stored. Database-wide changes to taxon are always stored, for example a name change due to taxonomic revision. Different constituent databases have somewhat differing practices for nomenclatural synonymizations made at the time of initial data validation and

upload: some retain the original name, others replace with the currently accepted synonym. For homotypic synonymizations that do not involve a change in circumscription (e.g. Gramineae ≡ Poaceae), original names do not have to be stored; but for heterotypic synonymizations or any name change that involves a potential change in circumscription, original names should be stored. In addition, name changes due to re-identification should be stored, for example a specimen initially identified as *Mammuthus* but later re-identified as *Mammut*.

For organismal taxa, the ***element*** is the organ or part of the organism that was identified. Thus, for pollen datasets, element names include 'pollen,' 'spore,' and 'stomate.' For plant macrofossil datasets elements include 'leaf,' 'seed,' 'bud scale,' 'microstrobilus,' and 'wood.' For fossil insects, elements include 'heads,' 'pronota' (thoraces), 'elytra' (left and right), and 'aedeagii' (reproductive organs). For vertebrates, the element is generally the bone or tooth identified, such as 'femur' or 'tooth, third molar.' Elements, particularly vertebrate elements, may have components, including symmetry, portion, and maturity, entered in that order and separated by semicolons, e.g. 'femur;left;distal;fused,' or 'tooth, third molar;lower left.' Elements are also used for modifying physical variables. For example, for the variable loss-on-ignition, the element is the temperature, e.g. 500°C. Neotoma also uses defined vocabularies for elements.

Units are the measurement units in which the variable is measured. For organismal taxa, the most common units are NISP (Number of Identified Specimens, often called a “count” for microfossils), MNI (Minimum Number of Individuals), and presence. For example, five left femurs would indicate at least five organisms (MNI=5), but five pollen grains could have one to five source plants (NISP=5). For presence data, a value of 1 is entered for presence or the cell is left blank. Zeroes or absences are not stored in Neotoma for organismal taxon variables, because the true absence of a taxon is difficult to definitively establish, given that probability of detection is a function of sampling effort, ecological rarity, and taphonomic processes (Birks and Line, 1992; Olszewski and Kidwell, 2007; Weng et al., 2006). However, partial evidence for taxonomic absence exists for some samples if that taxon was identified in other samples in the same dataset, under the assumption that the analyst looked for that taxon in all samples in a dataset. Neotoma also allows semi-quantitative systems for measuring abundance, e.g. the 1-5 relative abundance scale often used in the rodent midden literature (Spaulding et al., 1990) or other relative scales used in archaeobotany. Geochemical and physical variables can have many kinds of measurement units, e.g. 'percent,' 'per mille,' 'meq/L,' 'mg/L,' and so on. For physical and geochemical measurements, such as $\delta^{13}N$, for which zero can be a measured value, zeroes are stored. Neotoma uses defined vocabularies for units.

Context refers to a depositional context that may influence the interpretation of the taxon. Examples include anachronic, redeposited, or intrusive, which imply that the taxon was deposited at different time than its sediment matrix. A Cretaceous pollen grain may be reworked and redeposited into more recent sediments; a modern *Sus* (pig) bone may be intrusive in Pleistocene sediment. Anachronic simply implies the taxon is of a different age. Contexts sometimes used with pollen are clump and anther, where clumps of pollen or anther fragments may indicate an over-representation of the taxon, e.g. a bee carrying usually infrequent entomophilous pollen may have fallen into the sediment. It is possible for fossils in a dataset or sample to be from the same taxon but have different contexts, e.g. both clumped and non-clumped pollen of one taxon, or Holocene and older reworked Betulaceae pollen (distinguishable by preservational differences) in the same assemblage.

Time, Age Controls, Relative Ages, Age-Depth Models, Chronologies

In Neotoma, the age estimates attached to samples and specimens are treated and stored as a derived variable, that must be estimated through a combination of absolute age controls, relative age controls, event stratigraphic ages, and age-depth models fitted to those controls (e.g. Parnell et al., 2008). Some types of geochronological information are more stable than others. For example, individual geochronological measurements will remain stable, but the set of age controls available at a site or collection unit can change as new dates are obtained. The fitted age-depth models and derived age inferences are also changeable, as estimates of radiometric decay constants are updated, radiocarbon calibration curves are adjusted, the quality of individual age controls is reassessed, new statistical age-depth models are developed, etc. Hence, the Neotoma data model separately represents and stores these layers of information about time as *age controls*, *relative ages*, *age-depth models*, and *chronologies*. Most of the definitions described here are originally from Grimm et al. (2014).

An ***age control*** is an estimate of absolute age, often with a specified uncertainty, for a sample within a core or stratigraphic profile that is used to constrain an age model for that core or profile. Examples of age controls include radiocarbon and other radiometric dates, optically stimulated luminescence, biostratigraphic events, tephras, core top, coins or other dated cultural artifacts, etc. (Blois et al., 2011; Giesecke et al., 2014). Age controls are primary data in Neotoma and generally assumed to be fixed and unchanging; barring data entry error: new age controls can be added, but existing age controls are not modified. Age control data are stored in tables that are separate from but linked to tables that store information about age-depth

models and chronologies (see below). In Neotoma, age control data are stored in the Geochronology table.

Radiometric ages are the most common kind of age control stored in Neotoma, and radiocarbon dates are the most common kind of radiometric date (17,054 of 18,483 age controls in Neotoma are radiocarbon dates, as of November 7, 2017). Radiocarbon dates are stored in original radiocarbon years, with counting uncertainties stored as one standard deviation. Calibrated ages are stored in Neotoma as components of chronologies (below), or simply regenerated by users as needed. Other radiometric metadata include depth and thickness of sample, material dated, lab identifier, $\delta^{13}C$ (for radiocarbon dates), instrumental measurement system, and publication information.

Relative ages store information about the association of analysis units with formally recognized relative age scales based on the stratigraphic record or a series of geological events. These formations and events have their own age estimates, which can change over time. A relative age encompasses a range of time (i.e. it has an upper and lower age bound), and samples assigned to a relative age event are assigned those ages. Examples include Marine Isotope Stages, Heinrich stadials, geomagnetic chrons, archaeological periods, and North American land mammal ages. Thus, a sample assigned to Marine Isotope Stage 5e would be assigned a sample age of 130-116 ka based on the current authoritative estimate (currently, Lisiecki and Raymo, 2005). Optionally, this sample age could be further constrained by other criteria. A full list of relative age scales is available via the Neotoma API (http://api.neotomadb.org/v1/dbtables/RelativeAgeScales, in JSON format). Currently these relative age scales and associated age estimates are stored in Neotoma in the RelativeAges table and must be updated by Stewards; ideally these ages would be dynamically updated by linking to other authoritative data resources on stratigraphic age.

Event stratigraphic ages are globally synchronous, single-event stratigraphic markers, ideally with an age and error, which can be used in age models similar to geochronological ages. Volcanic deposits tephras often serve this purpose. For example, the Mazama tephra, which occurs over a large area of western North America, has been dated to 7,627±150 cal yr BP (Zdanowicz et al., 1999). The Hekla 1104 tephra originating from Iceland, but distributed as far east as Ireland, is historically documented to AD 1104 or 846 cal yr BP (Boygle, 1999). Some event stratigraphic ages are the boundaries of relative age units, e.g. geomagnetic polarity reversals are event stratigraphic ages and are also the boundary ages for geomagnetic chrons, which are relative ages. Hence, a fossil vertebrate assemblage might occur within a unit stratigraphically assigned to the reversed polarity Matuyama chron, with a relative age of 2.581-

0.781 Ma, while the Brunhes/Matuyama geomagnetic polarity reversal, which might be an age control for an age model in a core or section, is dated to 0.781 Ma (Ogg and Smith, 2005). Event stratigraphic ages are treated in the Neotoma data model as instantaneous; although, of course, the actual events occurred over periods of time, from weeks to a few years for a volcanic eruption to perhaps several hundred or even a few thousand years for a geomagnetic reversal (e.g. Clement, 2004). However, for most practical applications, this error should be relatively small and ideally incorporated within the age error estimate.

An ***age-depth model*** is an algorithm used to estimate the age-depth relationship for a given stratigraphic profile based on the age controls available for that profile and prior knowledge. Age-depth models are used to estimate ages for depths not directly associated with an age control or to resolve discrepancies among age controls. Examples of age-depth modeling programs include classical age modeling approaches (linear interpolation, linear regression, polynomials, and splines, Blaauw, 2010) and Bayesian approaches such as Bacon (Blaauw and Christen, 2011) or BChron (Parnell et al., 2008). In Neotoma, information about age-depth models is stored in the Chronologies table.

A ***chronology*** is a series of estimated ages for a set of samples in a collection unit, ideally with associated uncertainty estimates. Chronologies usually derive from an age-depth model and a set of age controls (radiometric, relative, events, or other) used to constrain that model. In Neotoma, information associated with chronologies is parsed into three tables: 1) The ChronControls table, which stores information about the age controls used to constrain an age-depth model. These age controls can, but do not have, to correspond to the age controls stored in the Geochronology and other primary data tables; this flexibility allows scientists to remove age controls deemed to be inaccurate or add other age constraints to the age-depth model (e.g. assigning a modern depositional age to the top of the stratigraphic profile). 2) The Chronologies table stores information about the age-depth model, its parameters, the analyst, and other metadata. 3) The SampleAges table stores the resultant inferred ages for individual samples. All Neotoma chronologies have a unique identifier and are also linked to a specific collection unit.

Chronologies can be stored as calendar years before present, radiocarbon years before present, or calibrated radiocarbon years before present. The chronology and associated age-depth model originally published for a collection unit are stored provided that the age controls, sample ages, and metadata sufficient to replicate the age model are published or provided by the contributor. A collection unit may have multiple chronologies or none, if no age information is available for the collection unit. Hence, each sample in the collection unit may have multiple

age estimates, each linked to a unique chronology. New chronologies may be added to Neotoma, and we envision this component of the Neotoma age model to be dynamic over time, as users download data and build new chronologies.

To handle the multiplicity of chronologies, Neotoma allows one chronology per collection unit to be designated as a default chronology. Choice of default chronologies is made by Data Stewards, and this choice can be revised as new chronologies are added. Default chronologies may be stored in calendar, radiocarbon, or calibrated radiocarbon years before present. Chronologies in radiocarbon years are not recommended, but exist as legacies in Neotoma. We anticipate that over time, these default radiocarbon chronologies will be replaced by updated chronologies in calibrated radiocarbon years. Researchers using Neotoma data are cautioned to critically examine its chronologies and encouraged to contribute new and revised chronologies.

DATA HOLDINGS AND DATA TYPES: CURRENT STATUS AND TRENDS

As of November 8, 2017, Neotoma holds over 3.8 million data records from 17,275 datasets and 9,269 sites (Fig. 6). Each data record is the measured value of a single taxon or other variable from a single sample. The Neotoma taxa table is a dictionary with over 29,000 taxa, fossil morphotypes, geochemical variables, and other variable names.

Data volumes in Neotoma have been rapidly growing as data are uploaded, with a 30% increase since 2014, when a new wave of data uploads began, following the extension of the Tilia software package (see *Technical Specifications and Software Ecosystem* below) was extended to enable data validation and upload to Neotoma (Fig. 6). As of July 13, 2017, Neotoma holds 2,954 pollen datasets (3,274,501 data records), 2,600 pollen surface sample datasets (56,205 data records) 3,669 vertebrate fauna datasets (59,278 data records), 388 diatom datasets (238,344 data records), 637 diatom surface sample datasets (29,968 data records), 554 ostracode surface sample datasets (2,410 data records), 384 macroinvertebrate datasets (805 data records), 283 plant macrofossil datasets (10,654 data records), 177 insect datasets (19,766 data records), and a number of other dataset types (Table 1). Of these, 5,226 datasets are from sediment cores, 562 from rodent middens, 285 from excavations, and 1,051 from stratigraphic sections. Neotoma holds 3,842 geochronologic datasets, with 18,543 individual geochronologic measurements. These data volumes have made Neotoma one of the largest structured repositories of geochronological data.

627 Although Neotoma has focused primarily on paleoecological and geochronological data,
628 it also stores associated physical and geochemical proxies. It includes modern water chemistry
629 data (1,317 datasets, 18,487 data records), loss-on-ignition data (190 datasets, 24,903 data
630 records), charcoal (100 datasets, 14,428 data records), and physical sediment measurements
631 (55 datasets, 1292 data records). Trial datasets have been uploaded for X-Ray Fluorescence
632 (XRF) and X-Ray Diffraction (XRD), and sedimentary geochemistry data. Surface samples can
633 be flagged in Neotoma during upload, to facilitate their use in the building of modern calibration
634 datasets for transfer functions (Birks, 1995). Some associated measurements of environmental
635 variables (relating to water chemistry) are enabled for the ostracode and diatom surface
636 samples stored in Neotoma, and we plan to extend the Neotoma data model to store other
637 environmental variables (e.g. climate variables) associated with surface samples. The Neotoma
638 data model was recently extended to include stable isotopic data and metadata for $\delta^{18}O$, $\delta^{13}C$,
639 $\delta^{15}N$, δD, $\delta^{34}S$, and $^{87}Sr/^{86}Sr$. Isotopic measurements can be stored for samples from
640 sedimentary profiles or from individual fossil specimens. We are expanding the Neotoma data
641 model to store organic biomarker data and taphonomic measurements on vertebrate fossils.
642 Because the formation of Constituent Databases in Neotoma is a voluntary, bottom-up process
643 (*Governance and Data Use Policy*), whether to extend Neotoma to other proxies largely
644 depends on 1) interest by paleodata communities in using Neotoma to house their data and 2)
645 developer and Steward time to extend data tables and metadata variables as needed.
646 Note that Neotoma currently focuses on storing primary paleoecological measurements
647 and generally does not emphasize the storing of derived inferences. For example, Neotoma
648 currently does not store indices of community diversity (richness, evenness, etc.), paleoclimatic
649 reconstructions, biomes or other paleovegetation reconstructions, etc. Our general philosophy
650 is that, given finite resources and the rapid pace of generating these inferences by the scientific
651 community, these reconstructions are best generated and managed outside of Neotoma, ideally
652 using workflow methods that clearly link all derived inferences back to the primary data
653 resources inside Neotoma. However, this boundary between primary data and secondary
654 inferences is not set in stone, and could be revised depending on research and development
655 priorities within and among Constituent Databases and their research communities.

TECHNICAL SPECIFICATIONS AND SOFTWARE ECOSYSTEM

The Neotoma software ecosystem (Fig. 4) has Neotoma's relational database at its heart and includes multiple systems for finding, exploring, visualizing, downloading, processing, and uploading data. We describe each in turn.

Database

The Neotoma Database is currently hosted on servers at Pennsylvania State University, maintained by the Center for Environmental Informatics (CEI). Neotoma uses a relational database structure that was originally deployed in Microsoft SQL Server and now is being migrated to PostgreSQL, so that it will rely primarily on open source technology. Neotoma's relational database structure continues to evolve over time, as new data types and metadata fields are added. A description of the relational database structure and tables is available in an on-line manual (http://neotoma-manual.readthedocs.org/en/latest/).

Data uploaded to the Neotoma relational database are protected by multiple backup measures at CEI, including redundant disk storage, off-site mirroring, file system snapshotting, regular tape backup, and duplication of the backup set. Complete snapshots of the Neotoma database are posted to the Neotoma website (https://www.neotomadb.org/snapshots) and to figShare (e.g. https://dx.doi.org/10.6084/m9.figshare.3376393.v1). figShare in turn ensures long term data sustainability through its partnership with the Digital Preservation Network who obtain periodic snapshots of the entire figShare collection (including the Neotoma snapshots), and replicate it across at least two Replicating Nodes, including the Academic Preservation, DuraCloud Vault, Stanford Digital Repository, Texas Preservation Node, and HathiTrust. Neotoma also will send these snapshots to the Paleoclimatology branch of the NOAA National Center for Environmental Informatics. Database snapshots are intermittent at present, but we plan to establish a regular automated schedule on a quarterly frequency. Neotoma is a certified member of the International Council for Science's World Data Service (ICSU-WDS), which sets standards for open, quality-assured, and sustained stewardship of scientific data.

Data Retrieval: Finding, Exploring, and Downloading

Multiple avenues exist to find, explore, and obtain data from Neotoma, each serving different needs and users. Neotoma data can be accessed via the Neotoma Explorer interactive map-based interface (https://apps.neotomadb.org/Explorer/), through a RESTful API (https://api.neotomadb.org/), and through digital object identifiers (DOIs) that provide persistent

and unique identifiers for every dataset in Neotoma (e.g. https://data.neotomadb.org/datasets/1001/). Further, Neotoma data can be searched, viewed, and analyzed through Neotoma Explorer and the Neotoma API, and also the *neotoma* R package and stratigraphic and map-based visualizations in Tilia. Through partnership with the Earth Life Consortium (http://earthlifeconsortium.org/), we are developing wrapper APIs that can simultaneously search for paleobiological data in Neotoma, the Paleobiology Database, and other partner databases (http://www.earthlifeconsortium.org/api_v1/ui/). A full set of links, manuals, code repositories, and other resources are provided in the Supplementary Material.

Neotoma Explorer (https://apps.neotomadb.org/explorer/), a map-based web application for searching, visualizing, and downloading data, is Neotoma's primary data discovery portal. Users can generate flexible queries for properties such as taxon name, variable type, time window, location, constituent database, site name, and researcher name. Results are displayed on an interactive map, and users can quickly retrieve site and dataset metadata by clicking on sites of interest. Users interested in deeper exploration of datasets can then view them in Explorer's Stratigraphic Diagrammer to inspect age models, stratigraphic plots, and associated publications. Datasets can be downloaded as delimited text files or shared via links that make use of Neotoma's API (e.g., http://apps.neotomadb.org/Explorer/?datasetid=1768). Searches can be saved as JSON files, which can be archived or shared with other users, who can reopen them in Neotoma Explorer (by dragging the JSON file to the Explorer window) to redisplay the results and map configuration produced from the query. Following our design principle of promoting openness, Explorer is built on the Open Web Platform using HTML5, CSS3, and JavaScript, enabling cross-browser and cross-platform support, and makes use of the open-source Dojo and OpenLayers libraries (Roth et al., 2014).

APIs enable programmatic access to the database by third-party developers and software applications. Likely API users include: scientists who need to incorporate the most current data into analytical workflows such as scripts written in R; organizations that want to distribute Neotoma data (with attribution) via their own data portal or web interface (e.g. customized and branded web portals for individual Constituent Databases); and developers creating standalone applications for data analysis and display. Known third-party users of Neotoma APIs include NOAA Paleoclimatology (https://www.ncdc.noaa.gov/data-access/paleoclimatology-data), which has a data search portal that will search and retrieve data housed in either NOAA or Neotoma; Flyover Country (http://fc.umn.edu/), a mobile app-based program for travelers to discover geological data and knowledge during their journeys (Loeffler

et al., 2015); and the Global Pollen Project (https://globalpollenproject.org/) (Martin and Harvey, 2017), a community platform for pollen identification. The APIs are implemented as platform- and language-independent RESTful web services; response formats include JSON and XML.

The R package ***neotoma*** uses the Neotoma API to pass data into R for further analysis (Goring et al., 2015). R is an open-source fourth-generation programming language for statistical analysis and graphics. Many paleoecological statistical and visualization packages have been developed for R, such as *analogue* (Simpson, 2007; Simpson and Oksanen, 2015), *rioja* (Juggins, 2015), *bchron* (Parnell et al., 2008), *bclim* (Parnell et al., 2016), *clam* (Blaauw, 2010), and *bacon* (Blaauw and Christen, 2011). Development of *neotoma* is ongoing (v1.7.0 is available via the Comprehensive R Archive Network, https://cran.r-project.org/web/packages/neotoma/index.html) and the living *neotoma* code, feature requests, bug reports, and development are open and available through GitHub (https://github.com/ropensci/neotoma).

***DOIs*.** We have recently begun assigning persistent and unique digital object identifiers (DOIs) to all Neotoma datasets, which will facilitate citation of Neotoma data and linked-data systems for sharing and connecting earth science data (Duerr et al., 2011). DOIs are created using the EZID system (https://ezid.lib.purdue.edu/), affiliated with DataCite (https://www.datacite.org/), through a license with the University of Wisconsin-Madison Library and University of Illinois. DOIs are assigned at the level of datasets, and each DOI has a landing page (e.g. https://data.neotomadb.org/datasets/1001/) that is designed to provide information to both human and machine users. The DOI pages point to living versions of the Neotoma data; if the Neotoma data are updated, these changes will be automatically detected and incorporated in the DOI landing pages. This approach differs from e.g., Pangaea (www.pangaea.de), in which DOIs point to static versions of datasets.

Data Validation, Upload, and Management

All data added to Neotoma are reviewed by a Data Steward (see *Governance and Data Use*) before upload; this expert curation is central to Neotoma's mission of providing high-quality scientific data. Data entry, curation, and upload into Neotoma are handled through the ***Tilia*** software (https://www.tiliait.com/). Tilia was originally developed as a DOS program to visualize and analyze pollen stratigraphic data (Grimm, 1988) and is still often used for this purpose. Tilia, now a Windows program, still maintains its end-user functions for managing, analyzing, and visualizing stratigraphic data, but its capabilities have been extended to support validation and upload of data to Neotoma, as well as direct download of data from Neotoma. Only Neotoma

Stewards who have password access to Neotoma have access to the upload capability; any user can download data. Tilia now generates .tlx files in extensible markup language (XML) format, which is a simple, extensible text-based markup language format. Tilia contains multiple validation procedures for quality control during upload, including checks against controlled vocabularies, for missing or duplicate data, for data inconsistencies, common errors, and commonly omitted metadata.

Neotoma uses controlled vocabularies for taxa, elements, units, contexts, geopolitical units, depositional environments, geochronological measurements, chronological controls, and other variables that store text-based names. These names are stored in Neotoma tables, and the Tilia validation process checks names in .tlx files against these names. Central to the database is the Taxa table, which contains the names of all taxa in the database. Names not found in the Taxa table during Tilia validation are listed, and Stewards can check for spelling errors or formatting differences for non-Latin modifiers by searching with wildcards. If the name is valid but not yet in Neotoma, the Steward can upload it to the database. Taxonomic names are separately validated by designated taxonomic experts. If the Steward is a taxonomic expert, then the date of entry is recorded as the date of validation. Otherwise the field for validation date is left empty until the taxon is validated. Tilia has a tool that shows the Neotoma taxa in a taxonomic hierarchy, easily allowing the Steward to place the new taxon in its correct taxonomic position (Fig. 7).

Other Tilia validation steps include checking for valid dataset types, for valid combinations of taxon and element (e.g. disallowing the combination of 'Picea' and 'femur'), that latitude-longitude coordinates for collection units fall within site bounding boxes, that elevations fall within possible limits, and that the younger/older reliable age bounds for chronologies are not reversed (a common error). To validate longitudes and latitudes, the Steward is shown a world map with the hemisphere indicated, which must be approved. Positive $\delta^{13}C$ values associated with radiocarbon dates (a common entry error) are flagged. Tilia will ask the Steward whether the top sample of a stratigraphic sequence should be flagged as a modern surface sample. During the validation process, error messages, warnings, or notes may be issued. Notes indicate omissions of optional metadata items that are not required and often non-existent or unavailable. Warnings indicate omissions of optional metadata items that nevertheless are highly desirable. Warnings are also issued for likely or possible inconsistencies, but which nevertheless may be correct. Errors must be addressed before upload is possible. Once data are validated with no errors, authorized Data Stewards can upload data from Tilia directly to Neotoma through password-protected web services.

Tilia has a customized API for data upload and download. Controlled vocabularies that appear in Tilia dropdown pick lists, such as taxa and geopolitical names, are held in local XML lookup files, which should be synchronized periodically from their counterparts in the central Neotoma database. Datasets downloaded from Neotoma to Tilia include all relevant metadata, including various notes and comments that may have been entered. This functionality enables enhanced review, visualization, and analysis of Neotoma data beyond that possible in Neotoma Explorer. The downloaded data also facilitate training for end users and Data Stewards, by providing model datasets in Tilia format. The Data Steward version of Tilia also allows Data Stewards to amend data already in Tilia. For example, metadata items such as latitude-longitude coordinates may be corrected, or missing metadata items for sites and collection units may be added. Publications can be corrected or added. Contact information can be updated. Datasets can be added to sites already in Neotoma. New chronologies and sample ages can be added. The ultimate goal is to enable Data Stewards to correct or add to any data or metadata item within their Constituent Database through the Tilia interface. Changes made through Tilia are logged, Neotoma preserves snapshots of prior database versions, and we are collaborating with others on building an annotation system, with support from NSF's EarthCube program. The ability to upload and amend data is a significant power available to Data Stewards, hence the need for experts to serve as Data Stewards and for Data Stewards to act judiciously when modifying data and metadata.

Tilia software is available from https://www.tiliait.com/. The free version has all the spreadsheet and metadata form options, including the ability to download data from Neotoma, which can be copied to other spreadsheet programs. A licensed version for Stewards is available at no cost and exposes additional options for data visualization, validation, upload, and management, with a password needed for any action that may alter the database. Other licensed versions of Tilia with graphics capabilities are available and are priced to cover software licensing costs associated with Tilia development.

The Tilia workflow for uploading data follows a model in which site-level datasets and associated metadata are uploaded individually. For larger data ports, where many site-level datasets must be exported from one database into Neotoma (e.g. from the Access tables storing data in FAUNMAP and the European Pollen Database), we have taken two approaches. One is to place the database on the Neotoma server, and write customized SQL procedures and web services to download datasets directly to individual Tilia files for validation and upload to Neotoma. The second is to write customized scripts in R or Python that export datasets from the other database to individual Tilia .tlx files, which can then be opened for validation and

upload by Data Stewards. Examples of these batch export scripts, developed for FAUNMAP are available on GitHub (https://github.com/NeotomaDB/FAUNMAP_Import).

GOVERNANCE AND DATA USE

Governance and data use policies are designed to support Neotoma's core goals of data openness and distributed scientific governance. In particular, these policies are intended to: 1) Make Neotoma data open and available to all interested scientific and public communities; 2) Build a governance structure that accommodates both a centralized cyberinfrastructure and a highly distributed scientific community of expertise; and 3) Empower and facilitate individual Data Stewards and Constituent Databases to set data acquisition priorities, curate data, and establish data quality standards and nomenclatures.

All Neotoma data are free to use through a CC BY 4.0 license. Complete attribution of Neotoma data includes a reference to the Neotoma Paleoecology database, constituent databases where relevant, and references to all original investigators and publications. An embargo policy has been developed and included in the Neotoma website and technical implementation is underway. See the Neotoma website (https://www.neotomadb.org/data/category/use) for a full description of the Neotoma data use policy and for the data use statements and citation formats for specific Constituent Databases.

With respect to governance, key needs include 1) an extendible, scalable governance structure that is easily open to new members, 2) effective executive decision-making and responsibility that is bounded by community oversight, and 3) mechanisms to ensure that Neotoma is curated by a community of professionals and scientific experts. Here we briefly summarize the main elements of Neotoma's governance (Fig. 3) and data use policies. A full description of Neotoma governance is described in its bylaws, available at http://www.neotomadb.org/about/category/governance.

Neotoma is governed by a ***Leadership Council*** (NLC) that sets policy and represents the scientific perspectives of Constituent Databases and their Data Stewards (Fig 3). Councilors serve for four-year renewable terms, with one-fourth of the NLC up for election each year. The Council is elected by Neotoma's ***members***, who are professional researchers and educators who contribute to and use Neotoma data and are interested in helping govern the database. Membership can be requested by any individual through a simple webform (https://tinyurl.com/NeotomaMember), and requests are approved by Neotoma's Nominations

and Membership Working Group, chaired by Neotoma's Associate Chair. Data Stewards are automatically granted Neotoma membership.

The NLC delegates responsibility for day-to-day operations to an **Executive Team** (Fig. 3), consisting of an Executive Chair, Associate Chair, and two other members. All positions on the Executive Team serve staggered four-year terms and are selected from and by the NLC. Other teams within the NLC include the Education and Outreach Working Group, the Informatics and Technology Working Group, and the International Partnerships Working Group.

Constituent Databases are a core concept in Neotoma, and a mechanism by which scientific data governance is distributed among the multiple fields of scientific expertise that Neotoma data embody. (A list of Neotoma's Constituent Databases can be obtained through the Tilia API - https://tilia.neotomadb.org/retrieve/?method=GetConstituentDatabases - which returns a JSON object). All data in Neotoma are associated with a Constituent Database, each curated by a community of Data Stewards, Taxonomic Experts, and Data Uploaders. Constituent Databases are responsible for vetting and uploading their community's data to Neotoma, setting priorities and quality standards for data uploads, managing taxonomic names, and appointing and training their Data Stewards and Taxonomic Experts. Constituent Databases may develop variants on the standard Neotoma Data Use policy (see below). Some Constituent Databases may be active for a few years, during e.g. a data mobilization campaign linked to specific large-scale synthesis project and research grant. Other Constituent Databases may be active indefinitely, when individuals and communities use Neotoma as their platform for archiving, managing, and sharing data.

Data Stewards serve a role analogous to that of editors in a peer-reviewed scientific journal. Much of the day-to-day power and responsibility devolves to Stewards for ensuring high-quality data uploads to Neotoma. Data Stewards are authorized to upload data directly to Neotoma within their Constituent Database, and to modify data within their Constituent Database. Data Stewards often work with one or more **Data Processors**, often students or other assistants who assist in the preparation and entry of data into Tilia for eventual upload to Neotoma. **Taxonomic Experts** are a type of Data Steward that can also authorize the addition of names to Neotoma's list of accepted taxa names and variables.

NEXT STEPS

New Users, Contributors, and Communities

Neotoma welcomes new members, data contributors, users, data stewards, and constituent databases. There are many avenues for participation by interested scientists. One simple step is to become a Neotoma member (https://tinyurl.com/NeotomaMember). Scientists interested in contributing their data to Neotoma should contact a Data Steward (https://www.neotomadb.org/data/category/contribution). Research labs with a lot of data may want to consider steward training for someone in their lab. Training webinars can be scheduled by request and are led by current Stewards. For scientists interested in learning how to access and use Neotoma data, we have posted learning materials on-line and periodically hold user-oriented training workshops (see Supplementary Materials).

Research teams interested in building regional- to global-scale data syntheses, both paleoecological and paleoclimatic, may find Neotoma useful as a data synthesis platform. In some cases, these efforts could lead to targeted data mobilization campaigns, uploads of data to Neotoma, and the chartering of new Constituent Databases.

For example, the ANTIGUA project (e.g. Barnosky et al., 2016) is using Neotoma to store fossil occurrences and age constraints for South American megafaunal species, with datasets currently being processed for upload to Neotoma. The SKOPE project, studying human-environment interactions in the southwestern US (e.g. Bocinsky and Kohler, 2014), is using Neotoma as a platform for accessing paleoecological data and, in the process, discovering data corrections and additional records for addition to Neotoma. PalEON, interested in understanding climate-driven vegetation dynamics over the last 2,000 years, has been discovering new records for addition to Neotoma and updating age models as part of its development of the STEPPS pollen-vegetation model (Dawson et al., 2016; Goring et al., 2016; Kujawa et al., 2016).

Building Partnerships with Allied Resources

In the paleosciences, a distributed network of data resources has emerged, each serving a particular suite of data and research communities: evolutionary biologists, paleoclimatologists, archaeologists, sample and core curators, etc. From this perspective, Neotoma is one node among several in what is emerging as a federated ecosystem of complementary and allied data

resources. The key need is to interlink these resources through adoption of common standards and data identifiers, so that scientific users can easily gather data from multiple sources.

Initial efforts to interlink resources have been supported by the NSF EarthCube program and include the Cyberinfrastructure for Paleogeoscience (C4P) research coordination network (RCN) and the EarthRates RCN, which have brought together leaders and users of cyberinfrastructure resources in the paleogeosciences to identify priorities for developing common standards and integrative analytical tools. Neotoma belongs to the EarthCube Council of Data Facilities (https://earthcube.org/group/council-data-facilities). One outcome of these collaborations is the Earth Life Consortium (http://earthlifeconsortium.org/), which is building easy-to-adopt APIs that can simultaneously search for data from multiple paleobiological data repositories. We are also working with VertNet (http://vertnet.org/) to send Neotoma data to the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/).

Adding Data to Neotoma

Entering, preparing, and validating data for entry into Neotoma requires effort. This effort of data cleaning, validation, and preparation is not unique to Neotoma, of course, and it remains the single largest bottleneck to comprehensive global-scale syntheses of paleoenvironmental data. A great advantage of putting data into Neotoma is that this effort is done just once, and then the data are readily available to multiple groups for multiple data synthesis projects. This unified effort contrasts with the common ad hoc and inefficient system in which multiple groups search for, obtain, and clean the same individual datasets.

Multiple solutions exist to this data ingest barrier, and Neotoma is exploring them all. One ready-to-go solution is to crowdsource it, by encouraging and incentivizing individual scientists to contribute data (e.g. through training workshops, recognition of Data Stewards, and by supporting scientists' research objectives). Another solution is to further streamline data input systems that prepare data and metadata in import-ready formats for Neotoma, at the earliest stages of scientific workflows (e.g. building customized import software for data generated from mass spectrometers or for microfossil analysts working at microscopes). A third is to make use of text-mining programs such as GeoDeepDive (Peters et al., 2017; Peters et al., 2014) for help discover and bring on-line dark data (Heidorn, 2008) and automating the discovery of new papers and the particularly routine parts of data entry, such as bibliographic citations. We welcome innovation by others in all of these areas.

Scalability and Sustainability

Scalability refers to the ability of Neotoma to grow, both with respect to the size of its data holdings and community of scientific data contributors and users, while sustainability refers to the ability of Neotoma to persist over time. The two are closely linked, because in both cases the central solution rests in Neotoma as resource that both supports community research priorities, and is supported by its constituent communities.

With respect to scalability, the key barriers are primarily social, rather than technological. In particular, potential limits to scalability include Neotoma's emphasis on 1) expert curation to ensure high data quality and utility and 2) Neotoma's model of community governance and voluntary participation. The former requires both community crowdsourcing and streamlined support tools for data entry, validation, and correction, while the latter means that Neotoma adoption is a voluntary social process, which can be fast or slow. Neotoma is working on the scalability challenge through the mechanisms described in the *Governance Model* and *New Users, Contributors, and Communities* sections above.

Sustainability covers multiple dimensions, including guarantees of the long-term preservation of the data housed by Neotoma and sustainability of the development efforts linked to Neotoma. Data sustainability is the easier challenge: Neotoma ensures long-term data sustainability through multiple and redundant mechanisms, including partnerships with multiple organizations and by ensuring that Neotoma and its constituent databases exist in multiple repositories (*Technical Specifications and Software Ecosystem*). Data sustainability solutions are well developed.

Sustainability of development is the larger challenge. Neotoma development efforts so far have been primarily supported by the US National Science Foundation. This reliance on a single funding source is risky, but no clear alternatives yet exist. Ongoing hosting of Neotoma data and data services is cheap, requiring just a few servers, and could be maintained indefinitely, in case of a lapse in funding. The real risk involves sustaining the community of developers who build and update these services. Neotoma's data are complex, and its infrastructure development relies on individuals with good practical training in both data sciences and paleosciences. This talent pool is very small but growing, and recruitment and retention of this talent remains a persistent challenge.

Community engagement is critical to sustainability because if Neotoma is closely engaged with its data contributors and users, and is seen by the scientific community as vital cyberinfrastructure that facilitates large-scale earth system science, then its prospects improve

for continued support, growth, and development. If Neotoma is not serving this mission, then it will (and should) ultimately lapse. Our personal view is that funding agencies need to commit resources to long-term support of meso-scale community cyberinfrastructure efforts such as Neotoma, as they do to other forms of physical scientific infrastructure, contingent upon satisfactory demonstration that these efforts are advancing community scientific needs.

One charge for the Neotoma Leadership Council is to explore multiple funding sources and business models. Because of Neotoma's commitment to open data, there are no plans to charge users for data access, but other options exist. One is for other national science agencies to support Neotoma's data ingest and development activities, perhaps through the leadership of individual investigators associated with Constituent Databases or new data synthesis efforts. A second is to establish a voluntary dues model, perhaps through partnership with professional societies, which traditionally have supported other forms of scientific knowledge dissemination such as peer-reviewed journals. A third is to partner with journals by providing them with a high-quality data archival service that meets community data standards. A fourth is to partner with investigators on data mobilization campaigns that as part of the process include resources for preparing and vetting datasets for upload to Neotoma. A fifth is to work with home universities to establish long-term base support for scientific databases, in line with universities' mission of discovering and disseminating knowledge, in a role similar to that served by university presses. All these options are viable, and all are being explored.

While recognizing that uncertainties exist, we are fundamentally optimistic about the long-term persistence, growth, and evolution of Neotoma. Neotoma, through its constituent databases, has supported macro-scale research for decades, and it has coalesced and grown organically from the lab-scale data synthesis efforts of individual investigators, to the development of relational database systems, and now to the development of online client-server architectures and the rise of distributed and networked networks of developers and scientists. Neotoma originated in direct response to the scientific objectives of paleoecologists and allied disciplines, and the general challenge of pursuing broad-scale science with local-scale data. As data volumes grow, both inside Neotoma (Fig. 6) and outside, community-curated data resources such as Neotoma are, increasingly, foundational infrastructure for big data science.

CONCLUSIONS

The Neotoma Paleoecology Database seeks to advance large-scale paleoecological, biogeographic, and global change research by providing an open, high-quality, and community-

curated resource for paleoecological and associated paleoenvironmental data. Sedimentary paleoecological proxy data are expensive to collect, in time and money; Neotoma provides a low-cost solution to data sharing and access via a common platform for many different kinds of paleoecological and associated data. High data quality is achieved through open and distributed scientific governance, based on a distributed network of expert Data Stewards and associated Constituent Databases. Neotoma is in a growth stage, with open doors for membership, new Data Stewards being trained, development of new functional capabilities, extension of the data model to additional data types (e.g. organic biomarkers, stable isotopes), and more data uploaded. At the same time, much more work remains to be done, given the large volumes of paleoecological data worldwide that remain dark, trapped in unstructured publication supplements, spreadsheets on personal computers, or other inaccessible venues, and at high risk of permanent loss. We respectfully encourage other paleoecologists, paleontologists, paleoclimatologists, archaeologists, and allied disciplines to use Neotoma data and software resources as part of their research workflows, to contribute their paleoecological and associated data to Neotoma (or other community-curated resources), and to serve as members, Data Stewards, Taxonomic Experts, and on the Leadership Council. Such service can advance both personal research goals for one's own region, time, taxonomic group, and questions of interest and broader community goals of open data and enabling large-scale science. Gathering, structuring, and sharing our hard-won data into larger open resources is our community's big data challenge; community-curated resources such as Neotoma are an essential part of our community's solution.

ACKNOWLEDGMENTS

Neotoma is a community effort, and it relies on the voluntary data contributions made by individual scientists and research groups and on work by Data Stewards (http://bit.ly/2tzjEsZ) to clean and check these data during and after their submission to Neotoma. We thank them all. Neotoma has been supported by the NSF Geoinformatics (0948652, 0947459, 1550707, 1550717, 1550805, 1550728, 1550716, 1550700, 1550890, 1550721, 1550755) and EarthCube (1541002, 1540994, 1541015, 1540979, 1540977) programs and the Wisconsin Alumni Research Foundation (WARF). Additional support came from the Wisconsin Alumni Research Foundation. Any use of trade, firm, or product names is for descriptive purposes and does not imply endorsement by the US government. Much of this paper was written while JWW was a visiting fellow at Durham University, hosted by the Institute for Advanced Study and Dr. Brian

Huntley. This manuscript was improved by comments from Tom Webb, Simon Brewer, and several anonymous reviewers. Scott Farley assisted in figure design and drafting. Neotoma is indebted to the vision and work of early builders of community paleodatabases, including Pat Anderson, Anthony Barnosky, Patrick Bartlein, Richard Bradshaw, Simon Brewer, Paul Buckland, Rachid Cheddadi, Jacques-Louis de Beaulieu, Joel Guiot, Sheila Hicks, Geoff Hope, Brian Huntley, Anne-Marie Lézine, Anatoly Lozhkin, Ernie Lundelius, Vera Markgraf, Pierre Richard, Jon Sadler, and Tom Webb, as well as to all the database coordinators and contributors too numerous to mention. Their vision carries on today.

TABLES

Table 1: Constituent Databases in Neotoma and the number of datasets in each.

FIGURES

Figure 1: Papers citing Neotoma and its constituent databases.

Figure 2: Neotoma serves many communities and acts as a boundary organization (Guston, 2001) among these communities. Neotoma serves paleoecologists by providing a high-quality repository for their paleoecological data, with value added via digital object identifiers to facilitate data citation, data curation, and a flexible data model. Neotoma serves data users by providing a well-structured, open-access, and easy-to-use source of paleoecological data, specializing in timescales that bridge the boundary between global change ecology and geology (Betancourt, 2012; Dietl and Flessa, 2011; Jackson, in press; Jackson and Blois, 2015; Jackson and Hobbs, 2009; Kidwell, 2015). In return, these communities generate new questions and analytical approaches for paleoecological data. Neotoma serves educators, students, and the general public seeking to learn about the past distributions of charismatic species such as the Pleistocene megafauna and the effects of climate change on species distribution and diversity. Neotoma also serves as a boundary organization between geoscientists and computer scientists, passing data, new research questions, best practices and protocols, and geoscientific use cases & priorities.

Figure 3: Diagram of Neotoma's governance structure. Neotoma is governed by a Leadership Council, which is populated by elected members serving four year terms. The Executive working group coordinates day-to-day operations and reports to the Leadership Council. Other working groups coordinate education and outreach activities, build informatics and development activities, cultivate international partnerships, and handle membership requests and leadership elections. Constituent Databases and the Data Stewards within these databases are charged with uploading data to Neotoma, setting data standards and vocabularies, adopting and harmonizing taxonomies, and deciding default age models. These constituent databases are organized by taxonomic group or paleoecological proxy type and often are further subdivided by region or time period. The Neotoma governance system is extensible, such that new members can readily join and new Constituent Databases can form.

Figure 4: Diagram of the Neotoma software ecosystem. Data preparation and cleaning for upload to Neotoma are handled by the Tilia software (https://www.tiliait.com/), which has password-protected access for Data Stewards to upload datasets, update age models, and correct errors. Data are stored in the Neotoma relational database, which is deployed in SQL Server and currently hosted at Penn State's Center for Environmental Informatics. Neotoma data can be discovered, explored, viewed, and obtained through multiple platforms. Neotoma Explorer and its graphical map-based interface is designed for first-pass data explorations, new users, and educational and student groups. The APIs and neotoma R package are intended for programmatic access and for users who wish to do large-volume searches of Neotoma data holdings. Tilia can also download datasets from Neotoma, which is useful for data visualizations and for Data Stewards needing to update datasets or looking for examples of prepared Tilia files.

Figure 5: The Neotoma data model handles different kinds of sampling designs by paleoecologists through a flexible hierarchical system consisting of sites, collection units, analysis units, samples, and datasets. ***Sites*** are the field locations from which paleoecological data are obtained and can contain multiple collection units. ***Collection units*** are the specific point-level locations within sites from which data are obtained and can contain multiple analysis units. ***Analysis units*** are the specific depth horizons from which data are obtained and can contain multiple samples. A ***sample*** is a single piece of material extracted from an analysis unit, for which a single kind of measurement is made (e.g. analyzed for fossil pollen, stable

isotopic analyses, etc.). A ***dataset*** comprises all samples of a single dataset type in a single collection unit, e.g. all pollen samples from a single core.

Figure 6: History of data uploads to Neotoma, expressed as number of datasets (a) and observations (b). Neotoma launched in 2009 with a number of datasets already in it, mostly pollen and vertebrates, representing prior database building efforts from the Global Pollen Database and FAUNMAP efforts. Rate of data uploads accelerated after 2013, when the new Neotoma data model was established and Tilia's data upload and validation routines were written. The number of datasets is relatively even among several major dataset types (vertebrates, pollen, geochronological data) with recent rapid growth of ostracode and diatom datasets. The number of pollen observations (b) is large relative to the number of datasets (a) because pollen datasets often have many samples (e.g. many samples per core) and many variables per sample (i.e. dozens of taxa per sample). As other taxa- and sample-rich datasets are added to Neotoma (e.g. diatoms, ostracodes), their relative proportion will quickly increase.

Figure 7: Tilia's interface for stewards to add new taxonomic names to Neotoma's Taxa table. Names are placed within a taxonomic tree and each taxon name is assigned a unique identifier. Stewards can also upload a citation for the source of that taxonomic name.

REFERENCES

Ammann, B., van Leeuwen, J.F.N., van der Knaap, W.O., Lischke, H., Heiri, O., Tinner, W., 2013. Vegetation responses to rapid warming and to minor climatic fluctuations during the Late-Glacial Interstadial (GI-1) at Gerzensee (Switzerland). Palaeogeography, Palaeoclimatology, Palaeoecology 391, Part B, 40-59.

Arroyo-Cabrales, J., Polaco, O.J., Johnson, E., 2007. An overview of the Quaternary mammals of Mexico. Courier Forschunginstitut Senckenberg 259, 191-203.

Arroyo-Cabrales, J., Polaco, O.J., Johnson, E., 2009. Providing a national perspective on Quaternary mammals through a Mexican database. The SAA Archaeological Records 9, 21-23.

Barnosky, A.D., Hadly, E.A., Gonzalez, P., Head, J., Polly, P.D., Lawing, A.M., Eronen, J.T., Ackerly, D.D., Alex, K., Biber, E., Blois, J., Brashares, J., Ceballos, G., Davis, E., Dietl, G.P., Dirzo, R., Doremus, H., Fortelius, M., Greene, H.W., Hellmann, J., Hickler, T., Jackson, S.T., Kemp, M., Koch, P.L., Kremen, C., Lindsey, E.L., Looy, C., Marshall, C.R., Mendenhall, C., Mulch, A., Mychajliw, A.M., Nowak, C., Ramakrishnan, U., Schnitzler, J., Das Shrestha, K., Solari, K., Stegner, L., Stegner, M.A., Stenseth, N.C., Wake, M.H., Zhang, Z., 2017. Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. Science 355.

Barnosky, A.D., Lindsey, E.L., Villavicencio, N.A., Bostelmann, E., Hadly, E.A., Wanket, J., Marshall, C.R., 2016. Variable impact of late-Quaternary megafaunal extinction in causing ecological state shifts in North and South America. Proceedings of the National Academy of Sciences 113, 856-861.

Bartlein, P.J., Harrison, S.P., Brewer, S., Connor, S., Davis, B.A.S., Gajewski, K., Guiot, J., Harrison-Prentice, T.I., Henderson, A., Peyron, O., Prentice, I.C., Scholze, M., Seppä, H., Shuman, B., Sugita, S., Thompson, R.S., Viau, A.E., Williams, J., Wu, H., 2011. Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis. Clim. Dyn. 37, 775-802.

Bennett, C.R., Provan, J., 2008. What do we mean by 'refugia'? Quaternary Science Reviews 27, 2449-2455.

Bernabo, J.C., Webb, T., III, 1977. Changing patterns in the Holocene pollen record of northeastern North America: A mapped summary. Quaternary Research 8, 64-96.

Betancourt, J., 2012. Reflections on the relevance of history in a nonstationary world, in: Wiens, J., Hayward, G.D., Safford, H.D., Giffen, C.M. (Eds.), Historical Environmental Variation in

Conservation and Natural Resource Management, First edition. John Wiley & Sons, Ltd., pp. 307-317.

Betancourt, J.L., Van Devender, T.R., Martin, P.S., 1990. Packrat Middens: The Last 40,000 Years of Biotic Change. Tucson, The University of Arizona Press, p. 467.

Birks, H.H., 2015. South to north: Contrasting late-glacial and early-Holocene climate changes and vegetation responses between south and north Norway. Holocene 25, 37-52.

Birks, H.J.B., 1995. Quantitative paleoenvironmental reconstruction, in: Maddy, D., Brew, J.S. (Eds.), Statistical modelling of Quaternary Science Data. Technical Guide 5. Quaternary Research Association, Cambridge, pp. 116-254.

Birks, H.J.B., Line, J.M., 1992. The use of rarefaction analysis for estimating palynological richness from Quaternary pollen-analytical data. The Holocene 2, 1-10.

Blaauw, M., 2010. Methods and code for 'classical' age-modelling of radiocarbon sequences. Quaternary Geochronology 5, 512-518.

Blaauw, M., Christen, J.A., 2011. Flexible paleoclimate age-depth models using an autoregressive gamma process. Bayesian Analysis 6, 1-18.

Blarquez, O., Carcaillet, C., Frejaville, T., Bergeron, Y., 2014. Disentangling the trajectories of alpha, beta and gamma plant diversity of North American boreal ecoregions since 15,500 years. Frontiers in Ecology and Evolution 2.

Blois, J.L., Williams, J.W., Grimm, E.C., Jackson, S.T., Graham, R.W., 2011. A methodological framework for improved paleovegetation mapping from late-Quaternary pollen records. Quaternary Science Reviews 30, 1926-1939.

Blois, J.L., Zarnetske, P.L., Fitzpatrick, M.C., Finnegan, S., 2013. Climate change and the past, present, and future of biotic interactions. Science 341, 499-504.

Bocinsky, R.K., Kohler, T.A., 2014. A 2,000-year reconstruction of the rain-fed maize agricultural niche in the US Southwest. Nature Communications 5, 5618.

Boygle, J., 1999. Variability of tephra in lake and catchment sediments, Svínavatn, Iceland. Global and Planetary Change 21, 129-149.

Brewer, S., Jackson, S.T., Williams, J.W., 2012. Paleoecoinformatics: Applying geohistorical data to ecological questions. Trends Ecol. Evol. 27, 104-112.

Buckland, P.I., 2007. The Development and Implementation of Software for Palaeoenvironmental and Palaeoclimatological Research: The Bugs Coleopteran Ecology Package (BugsCEP), Department of Archaeology & Sámi Studies. University of Umeå, Umeå, Sweden, p. 236 pp + CD.

1188 Bush, R.T., McInerney, F.A., 2013. Leaf wax n-alkane distributions in and across modern plants:
1189 Implications for paleoecology and chemotaxonomy. Geochimica et Cosmochimica Acta 117,
1190 161-179.
1191 Carrasco, M.A., Barnosky, A.D., Kraatz, B.P., Davis, E.B., 2007. The Miocene mammal
1192 mapping project (MIOMAP): An online database of Arikareean throuh Hemphilian fossil
1193 mammals. Bulletin of Carnegie Museum of Natural History 39, 183-188.
1194 Cinget, B., de Lafontaine, G., Gérardi, S., Bousquet, J., 2015. Integrating phylogeography and
1195 paleoecology to investigate the origin and dynamics of hybrid zones: insights from two
1196 widespread North American firs. Molecular Ecology 24, 2856-2870.
1197 Clarke, S.J., Lynch, A.J.J., 2016. Palaeoecology to inform wetland conservation and
1198 management: some experiences and prospects. Marine and Freshwater Research
1199 http://dx.doi.org/10.1071/MF15031.
1200 Clement, B.M., 2004. Dependence of the duration of geomagnetic polarity reversals on site
1201 latitude. Nature 428, 637-640.
1202 CLIMAP Project Members, 1976. The surface of Ice-Age Earth: Quantitative geological
1203 evidence is used to reconstruct boundary conditions for the climate 18,000 years ago.
1204 Science 191, 1131-1137.
1205 Davis, M.B., 1976. Pleistocene biogeography of temperate deciduous forests. Geoscience and
1206 Man XIII, 13-26.
1207 Dawson, A., Paciorek, C.J., McLachlan, J.S., Goring, S., Williams, J.W., Jackson, S.T., 2016.
1208 Quantifying pollen-vegetation relationships to reconstruct forests using 19th-century forest
1209 composition and pollen data. Quaternary Science Reviews 137, 156-175.
1210 Dawson, T.P., Jackson, S.T., House, J.I., Prentice, I.C., Mace, G.M., 2011. Beyond predictions:
1211 Biodiversity conservation in a changing climate. Science 332, 53-58.
1212 De La Torre, A.R., Roberts, D.R., Aitken, S.N., 2014. Genome-wide admixture and ecological
1213 niche modelling reveal the maintenance of species boundaries despite long history of
1214 interspecific gene flow. Molecular Ecology 23, 2046-2059.
1215 deMenocal, P.B., 2001. Cultural responses to climate change during the late Holocene. Science
1216 292, 667-673.
1217 Dietl, G.P., Flessa, K.W., 2011. Conservation paleoecology: Putting the dead to work. Trends
1218 Ecol. Evol. 26, 30-37.
1219 Dietl, G.P., Kidwell, S.M., Brenner, M., Burney, D.A., Flessa, K.W., Jackson, S.T., Koch, P.L.,
1220 2015. Conservation paleobiology: Leveraging knowledge of the past to inform conservation
1221 and restoration. Annual Review of Earth and Planetary Sciences 43, 79-103.

1222 Dietze, M.C., 2017. Ecological Forecasting. Princeton University Press, Princeton, NJ.
1223 Dietze, M.C., Lebauer, D.S., Kooper, R., 2012. On improving the communication between
1224 models and data. Plant, Cell & Environment doi: 10.1111/pce.12043.
1225 Doughty, C.E., Wolf, A., Madhi, Y., 2013. The legacy of the Pleistocene megafauna extinctions
1226 on nutrient availability in Amazonia. Nature Geoscience 6, 761-764.
1227 ECMA International, 2013. The JSON data standard format. ECMA International, Geneva,
1228 Switzerland.
1229 Ellis, E.C., Kaplan, J.O., Fuller, D.Q., Vavrus, S., Goldewijk, K.K., Verburg, P.H., 2013. Used
1230 planet: A global history. Proceedings of the National Academy of Sciences 110, 7978-7985.
1231 Emery-Wetherell, M.M., McHorse, B.K., Davis, E.B., in press. Spatially explicit analysis sheds
1232 new light on the Pleistocene Megafaunal Extinction in North America. Paleobiology.
1233 Evans, M.N., Tolwinski-Ward, S.E., Thompson, D.M., Anchukaitis, K.J., 2013. Applications of
1234 proxy system modeling in high resolution paleoclimatology. Quaternary Science Reviews
1235 76, 16-28.
1236 FAUNMAP Working Group, 1994. FAUNMAP: A Database Documenting Late Quaternary
1237 Distributions of Mammal Species in the United States. Illinois State Museum, Springfield, IL.
1238 Ferguson, A.R., Nielson, J.L., Cragin, M.H., Bandrowski, A.E., Martone, M.E., 2014. Big data
1239 from small data: data-sharing in the 'long tail' of neuroscience. Nature Neuroscience 17,
1240 1442-1448.
1241 Finsinger, W., Giesecke, T., Brewer, S., Leydet, M., 2017. Emergence patterns of novelty in
1242 European vegetation assemblages over the past 15 000 years. Ecology Letters 20, 336–
1243 346.
1244 Flantua, S.G.A., Hooghiemstra, H., Grimm, E.C., Behling, H., Bush, M.B., González-Arango, C.,
1245 Gosling, W.D., Ledru, M.-P., Lozano-García, S., Maldonado, A., Prieto, A.R., Rull, V., Van
1246 Boxel, J.H., 2015. Updated site compilation of the Latin American Pollen Database. Rev.
1247 Palaeobot. Palynology 223, 104-115.
1248 Forester, R.M., Smith, A.J., Palmer, D.F., Curry, B.B., 2005. NANODe: North American
1249 Nonmarine Ostracode Database, version 1. Kent State University, Kent, OH, USA.
1250 Fritz, S.A., Schnitzler, J., Eronen, J.T., Hof, C., Böhning-Gaese, K., Graham, C.H., 2013.
1251 Diversity in time and space: wanted dead and alive. Trends Ecol. Evol. 28, 509-516.
1252 Fyfe, R.M., de Beaulieu, J.-L., Binney, H., Bradshaw, R.H.W., Brewer, S., Le Flao, A., Finsinger,
1253 W., Gaillard, M.-J., Giesecke, T., Gil-Romera, G., Grimm, E.C., Huntley, B., Kunes, P., Kühl,
1254 N., Leydet, M., Lotter, A.F., Tarasov, P.E., Tonkov, S., 2009. The European Pollen
1255 Database: past efforts and current activities. Veg. Hist. Archaeobot. 18, 417-424.

Gaillard, M.J., Sugita, S., Mazier, F., Trondman, A.K., Broström, A., Hickler, T., Kaplan, J.O., Kjellström, E., Kokfelt, U., Kuneš, P., Lemmen, C., Miller, P., Olofsson, J., Poska, A., Rundgren, M., Smith, B., Strandberg, G., Fyfe, R., Nielsen, A.B., Alenius, T., Balakauskas, L., Barnekow, L., Birks, H.J.B., Bjune, A., Björkman, L., Giesecke, T., Hjelle, K., Kalnina, L., Kangur, M., van der Knaap, W.O., Koff, T., Lagerås, P., Latałowa, M., Leydet, M., Lechterbeck, J., Lindbladh, M., Odgaard, B., Peglar, S., Segerström, U., von Stedingk, H., Seppä, H., 2010. Holocene land-cover reconstructions for studies on land cover-climate feedbacks. Climates of the Past 6, 483-499.

Gavin, D.G., Fitzpatrick, M.C., Gugger, P.F., Heath, K.D., Rodríguez-Sánchez, F., Dobrowski, S.Z., Hampe, A., Hu, F.S., Ashcroft, M.B., Bartlein, P.J., Blois, J.L., Carstens, B.C., Davis, E.B., de Lafontaine, G., Edwards, M.E., Fernandez, M., Henne, P.D., Herring, E.M., Holden, Z.A., Kong, W.-s., Liu, J., Magri, D., Matzke, N.J., McGlone, M.S., Saltré, F., Stigall, A.L., Tsai, Y.-H.E., Williams, J.W., 2014. Climate refugia: joint inference from fossil records, species distribution models and phylogeography. New Phytologist 204, 37-54.

Giesecke, T., Brewer, S., Finsinger, W., Leydet, M., Bradshaw, R.H.W., 2017. Patterns and dynamics of European vegetation change over the last 15,000 years. Journal of Biogeography 44, 1441-1456.

Giesecke, T., Davis, B., Brewer, S., Finsinger, W., Wolters, S., Blaauw, M., de Beaulieu, J.-L., Binney, H., Fyfe, R., Gaillard, M.-J., Gil-Romera, G., van der Knaap, W.O., Kuneš, P., Kühl, N., van Leeuwen, J.N., Leydet, M., Lotter, A., Ortu, E., Semmler, M., Bradshaw, R.W., 2014. Towards mapping the late Quaternary vegetation change of Europe. Veg. Hist. Archaeobot. 23, 75-86.

Goring, S., Dawson, A., Simpson, G., Ram, K., Graham, R.W., Grimm, E.C., Williams, J.W., 2015. *neotoma*: A Programmatic Interface to the Neotoma Paleoecological Database. Open Quaternary 1, 1-17.

Goring, S.J., Williams, J.W., 2017. Effect of historic land-use and climate change on tree-climate relationships in the upper Midwestern United States. Ecology Letters.

Goring, S.J., Williams, J.W., Mladenoff, D.J., Cogbill, C.V., Record, S., Paciorek, C.J., Jackson, S.J., Dietze, M.C., McLachlan, J.S., 2016. Novel and lost forests in the upper Midwestern United States, from new estimates of settlement-era composition, stem density, and biomass. PLoS One 11, e0151935.

Graham, R.W., Lundelius, E.L., Jr., Graham, M.A., Schroeder, E.K., Toomey, R.S., III, Anderson, E., Barnosky, A.D., Burns, J.A., Churcher, C.S., Grayson, D.K., Guthrie, R.D., Harington, C.R., Jefferson, G.T., Martin, L.D., McDonald, H.G., Morlan, R.E., Semken Jr.,

H.A., Webb, S.D., Werdelin, L., Wilson, M.C., 1996. Spatial response of mammals to Late Quaternary environmental fluctuations. Science 272, 1601:1606.

Grant, M.J., Stevens, C.J., Whitehouse, N.J., Norcott, D., Macphail, R.I., Langdon, C., Cameron, N., Barnett, C., Langdon, P.G., Crowder, J., Mulhall, N., Attree, K., Leivers, M., Greatorex, R., Ellis, C., 2014. A palaeoenvironmental context for Terminal Upper Palaeolithic and Mesolithic activity in the Colne Valley: Offsite records contemporary with occupation at Three Ways Wharf, Uxbridge. Environmental Archaeology 19, 131-152.

Grimm, E.C., 1988. Data analysis and display, in: Huntley, B., Webb, T., III (Eds.), Vegetation History. Kluwer Academic Publishers, Dordrecht, pp. 43-76.

Grimm, E.C., Blaauw, M., Buck, C.E., Williams, J.W., 2014. Age models, chronologies, and databases workshop: Complete report and recommendations. PAGES Workshop Report 22.

Grimm, E.C., Keltner, J., Cheddadi, R., Hicks, S., Lézine, A.-M., Berrio, J.C., Williams, J.W., 2013. Pollen databases and their application, in: Elias, S.A., Mock, C.J. (Eds.), Encyclopedia of Quaternary Science. Elsevier, pp. 831-838.

Guston, D.H., 2001. Boundary organizations in environmental policy and science: An introduction. Science, Technology, & Human Values 26, 399-408.

Gutiérrez-García, T.A., Vázquez-Domínguez, E., Arroyo-Cabrales, J., Kuch, M., Enk, J., King, C., Poinar, H.N., 2014. Ancient DNA and the tropics: a rodent's tale. Biology Letters 10, 20140224.

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., Porter, J.H., 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11, 156-162.

Heffernan, J.B., Soranno, P.A., Angilletta, M.J., Buckley, L.B., Gruner, D.S., Keitt, T.H., Kellner, J.R., Kominoski, J.S., Rocha, A.V., Xiao, J., Harms, T.K., Goring, S.J., Koenig, L.E., McDowell, W.H., Powell, H., Richardson, A.D., Stow, C.A., Vargas, R., Weathers, K.C., 2014. Macrosystems ecology: understanding ecological patterns and processes at continental scales. Frontiers in Ecology and the Environment 12, 5-14.

Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Library Trends 57, 280-299.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Yon Rhee, S., 2008. Big data: The future of biocuration. Nature 455, 47-50.

Huntley, B., Birks, H.J.B., 1983. An Atlas of Past and Present Pollen Maps for Europe: 0-13000 Years Ago. Cambridge University Press, Cambridge.

Jackson, S.T., in press. Late Quaternary biogeography: linking biotic responses to environmental variability across timescales, in: Lomolino, M., Heaney, L. (Eds.), Frontiers of Biogeography: New Directions in the Geography of Nature. Sinauer Associates, Inc., Sunderland, MA.

Jackson, S.T., Blois, J.L., 2015. Community ecology in a changing environment. Proceedings of the National Academy of Sciences 112, 4915-4921.

Jackson, S.T., Grimm, E.C., Thompson, R.S., 2000. Database resources in Quaternary paleobotany, in: Lipscomb, B., Pipoly, J., Sanders, R. (Eds.), Floristics in the New Millennium, pp. 113-120.

Jackson, S.T., Hobbs, R.J., 2009. Ecological restoration in the light of ecological history. Science 325, 567-569.

Jackson, S.T., Overpeck, J.T., Webb, T., III, Keattch, S.E., Anderson, K.H., 1997. Mapped plant-macrofossil and pollen records of late Quaternary vegetation change in eastern North America. Quaternary Science Reviews 16, 1-70.

Jezkova, T., Riddle, B.R., Card, D.C., Schield, D.R., Eckstut, M.E., Castoe, T.A., 2015. Genetic consequences of postglacial range expansion in two codistributed rodents (genus Dipodomys) depend on ecology and genetic locus. Molecular Ecology 24, 83-97.

Juggins, S., 2015. rioja: Analysis of Quaternary Science Data, R package version (0.9-5).

Kaplan, J.O., Krumhardt, K.M., Ellis, E.C., Ruddiman, W.F., Lemmen, C., Goldewijk, K.K., 2011. Holocene carbon emissions as a result of anthropogenic land cover change. The Holocene 21, 775-791.

Kaplan, J.O., Krumhardt, K.M., Zimmermann, N., 2009. The prehistoric and preindustrial deforestation of Europe. Quaternary Science Reviews 28, 3016-3034.

Kidwell, S.M., 2015. Biology in the Anthropocene: Challenges and insights from young fossil records. Proceedings of the National Academy of Sciences 12, 4922-4929.

Kujawa, E., Goring, S., Dawson, A., Calcote, R., Grimm, E.C., Hotchkiss, S.C., Jackson, S.T., Lynch, E.A., McLachlan, J., St. Jacques, J.-M., Umbanhowar, C., Jr., Williams, J.W., 2016. The effects of anthropogenic land cover change on pollen-vegetation relationships in the American Midwest. Anthropocene http://dx.doi.org/10.1016/j.ancene.2016.09.005.

Latorre, C., Moreno, P.I., Grimm, E.C., 2014. 1st Workshop on paleoecological databases in South America. PAGES Magazine 22, 52.

Lehnert, K., Hsu, L., 2015. The new paradigm of data publication. Elements 11, 368-369.

Li, C., 2004. Dynamics of late Quaternary mammal population inferred from geostatistical study of the Faunmap database and Its implications for conservation. Universität Trier, Trier, p. 172.

Lisiecki, L.E., Raymo, M.E., 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic $d^{18}O$ records. Paleoceanography 20, DOI: 10.1029/2004PA001071.

Loeffler, S., Ai, S., McEwan, R., Myrbo, A., 2015. Flyover Country: A plane ride could be to geoscience outreach what a planetarium is to astronomy outreach - the perfect venue for sharing big, awe inspiring ideas, with a view to match, American Geophysical Union, San Francisco, CA, pp. ED54A-07.

Lorenzen, E.D., Nogues-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K.A., Ugan, A., Borregaard, M.K., Gilbert, M.T.P., Nielsen, R., Ho, S.Y.W., Goebel, T., Graf, K.E., Byers, D., Stenderup, J.T., Rasmussen, M., Campos, P.F., Leonard, J.A., Koepfli, K.-P., Froese, D., Zazula, G., Stafford, T.W., Aaris-Sorensen, K., Batra, P., Haywood, A.M., Singarayer, J.S., Valdes, P.J., Boeskorov, G., Burns, J.A., Davydov, S.P., Haile, J., Jenkins, D.L., Kosintsev, P., Kuznetsova, T., Lai, X., Martin, L.D., McDonald, H.G., Mol, D., Meldgaard, M., Munch, K., Stephan, E., Sablin, M., Sommer, R.S., Sipko, T., Scott, E., Suchard, M.A., Tikhonov, A., Willerslev, R., Wayne, R.K., Cooper, A., Hofreiter, M., Sher, A., Shapiro, B., Rahbek, C., Willerslev, E., 2011. Species-specific responses of Late Quaternary megafauna to climate and humans. Nature 479, 359-364.

Lynch, C., 2008. How do your data grow? Nature 455, 28-29.

Lyons, S.K., 2001. A quantitative assessment of the community structure and dynamics of Pleistocene mammals. University of Chicago, Chicago, p. 230.

Marcott, S.A., Shakun, J.D., Clark, P.U., Mix, A.C., 2013. A reconstruction of regional and global temperature for the past 11,300 years. Science 339, 1198-1201.

MARGO Project Members, 2009. Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum. Nature Geosci 2, 127-132.

Martin, A.C., Harvey, W.J., 2017. The Global Pollen Project: a new tool for pollen identification and the dissemination of physical reference collections. Methods in Ecology and Evolution, n/a-n/a.

Moritz, C., Agudo, R., 2013. The future of species under climate change: Resilience or decline? Science 341, 504-508.

Muñoz, S.E., Gajewski, K., Peros, M.C., 2010. Synchronous environmental and cultural change in the prehistory of the northeastern United States. Proceedings of the National Academy of Sciences 107, 22008-22013.

Nature, 2017. Not-so-open data. Nature 546, 327.
Ogg, J.G., Smith, A.G., 2005. The geomagnetic polarity time scale., in: Gradstein, F.M., Ogg, J.G., Smith, A.G. (Eds.), A Geologic Time Scale 2004. Cambridge University Press, Cambridge, United Kingdom, pp. 63-86.
Ohri, A., 2014. R with cloud APIs, R for Cloud Computing. Springer, New York, pp. 217-235.
Olszewski, T.D., Kidwell, S.M., 2007. The preservational fidelity of evenness in molluscan death assemblages. Paleobiology 33, 1-23.
Ordonez, A., Williams, J.W., 2013. Climatic and biotic velocities for woody taxa distributions over the last 16 000 years in eastern North America. Ecology Letters 16, 773-781.
Panagiotakopulu, E., Buchan, A.L., 2015. Present and Norse Greenlandic hayfields – Insect assemblages and human impact in southern Greenland. The Holocene 25, 921-931.
Parnell, A.C., Haslett, J., Allen, J.R.M., Buck, C.E., Huntley, B., 2008. A flexible approach to assessing synchroneity of past events using Bayesian reconstructions of sedimentation history. Quaternary Science Reviews 27, 1872-1885.
Parnell, A.C., Haslett, J., Sweeney, J., Doan, T.K., Allen, J.R.M., Huntley, B., 2016. Joint palaeoclimate reconstruction from pollen data via forward models and climate histories. Quaternary Science Reviews 151, 111-126.
Peters, S.E., Husson, J.M., Wilcots, J., 2017. The rise and fall of stromatolites in shallow marine environments. Geology G38931.1.
Peters, S.E., Zhang, C., Livny, M., Ré, C., 2014. A machine reading system for assembling synthetic paleontological databases. PLoS ONE 9, e113523.
Radeloff, V.C., Williams, J.W., Bateman, B.L., Burke, K.D., Carter, S.K., Childress, E.S., Cromwell, K., Gratton, C., Hasley, A.O., Kraemer, B.M., Latzka, A.W., Marin-Spiotta, E., Meine, C.D., Munoz, S.E., Neeson, T.M., Pidgeon, A.M., Rissman, A.R., Rivera, R.J., Szymanski, L.M., Usinowicz, J., 2015. The rise of novelty in ecosystems. Ecol. Appl. 25, 2051-2068.
Richard, P.J.H., 1995. Le couvert végétal du Québec-Labrador il y a 6000 ans BP: essai. Geogr. Phys. Quat. 49, 117-140.
Roth, R.E., Donohue, R.G., Sack, C.M., Wallace, T.R., Buckingham, T.M.A., 2014. A process for keeping pace with evolving web mapping technologies. Cartographic Perspectives 25-52.
Sachse, D., Billault, I., Bowen, G.J., Chikaraishi, Y., Dawson, T.E., Feakins, S.J., Freeman, K.H., Magill, C.R., McInerney, F.A., van der Meer, M.T.J., Polissar, P., Robins, R.J., Sachs, J.P., Schmidt, H.-L., Sessions, A.L., White, J.W.C., West, J.B., Kahmen, A., 2012. Molecular

Paleohydrology: Interpreting the Hydrogen-Isotopic Composition of Lipid Biomarkers from Photosynthesizing Organisms. Annual Review of Earth and Planetary Sciences 40, 221-249.

Sadler, J.P., Buckland, P.C., Rains, M., 1992. BUGS: an entomological database. Antenna 16, 158-166.

Sandom, C.J., Ejrnæs, R., Hansen, M.D.D., Svenning, J.-C., 2014. High herbivore density associated with vegetation diversity in interglacial ecosystems. Proceedings of the National Academy of Sciences 111, 4162-4167.

Schmittner, A., Urban, N.M., Shakun, J.D., Mahowald, N.M., Clark, P.U., Bartlein, P.J., Mix, A.C., Rosell-Melé, A., 2011. Climate sensitivity estimated from temperature reconstructions of the Last Glacial Maximum. Science 334, 1385-1388.

Seddon, A.W., Macias-Fauria, M., Willis, K.J., 2015. Climate and abrupt vegetation change in Northern Europe since the last deglaciation. The Holocene 25, 25-36.

Shakun, J.D., Clark, P.U., He, F., Marcott, S.A., Mix, A.C., Liu, Z., Otto-Bliesner, B., Schmittner, A., Bard, E., 2012. Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. Nature 484, 49-54.

Shuman, B.N., Newby, P., Donnelly, J.P., 2009. Abrupt climate change as an important agent of ecological change in the Northeast U.S. throughout the past 15,000 years. Quaternary Science Reviews 28, 1693-1709.

Simpson, G.L., 2007. Analogue methods in palaeoecology: Using the analogue package Journal of Statistical Software 22, 1-29.

Simpson, G.L., Oksanen, J., 2015. analogue: Analogue matching and Modern Analogue Technique transfer function models. (R package version 0.16-3). .

Spaulding, W.G., Betancourt, J.L., Croft, L.K., Cole, K.L., 1990. Packrat middens: Their composition and methods of analysis, in: Betancourt, J.L., Van Devender, T.R., Martin, P.S. (Eds.), Packrat Middens: The Last 40,000 Years of Biotic Change. University of Arizona Press, Tucson, AZ, pp. 59-84.

Sullivan, T.J., Charles, D.F., 1994. The feasibility and utility of a paleolimnology/paleoclimate data cooperative for North America. J. Paleolimn. 10, 265-273.

Svenning, J.-C., Sandel, B., 2013. Disequilibrium vegetation dynamics under future climate change. American Journal of Botany 100, 1266-1286.

Trouet, V., Diaz, H.F., Wahl, E.R., Viau, A.E., Graham, R., Graham, N., Cook, E.R., 2013. A 1500-year reconstruction of annual mean temperature for temperate North America on decadal-to-multidecadal time scales. Environmental Research Letters 8, 024008.

1458 Uhen, M.D., Barnosky, A.D., Bills, B., Blois, J., Carrano, M.T., Carrasco, M.A., Erickson, G.M.,
1459 Eronen, J.T., Fortelius, M., Graham, R.W., Grimm, E.C., O'Leary, M.A., Mast, A., Piel, W.H.,
1460 Polly, P.D., Säilä, L.K., 2013. From card catalogs to computers: Databases in vertebrate
1461 paleontology. Journal of Vertebrate Paleontology 33, 13-28.
1462 Viau, A.E., Ladd, M., Gajewski, K., 2012. The climate of North America during the past 2000
1463 years reconstructed from pollen data. Global and Planetary Change 84–85, 75-83.
1464 Vickers, K., Buckland, P.I., 2015. Predicting island beetle faunas by their climate ranges: the
1465 tabula rasa/refugia theory in the North Atlantic. Journal of Biogeography 42, 2031-2048.
1466 Vincens, A., Lézine, A.-M., Buchet, G., Lewden, D., Le Thomas, A., 2007. African pollen
1467 database inventory of tree and shrub pollen types. Rev. Palaeobot. Palynology 145, 135-
1468 141.
1469 Webb, T., III, 1997. Spatial response of plant taxa to climate change: A palaeoecological
1470 perspective, in: Huntley, B., Cramer, W., Morgan, A.V., Prentice, H.C., Allen, J.R.M. (Eds.),
1471 Past and Future Rapid Environmental Changes: The Spatial and Evolutionary Responses
1472 of Terrestrial Biota. Springer-Verlag, Berlin, pp. 55-72.
1473 Weng, C., Hooghiemstra, H., Duivenvoorden, J.F., 2006. Challenges in estimating past plant
1474 diversity from fossil pollen data: statistical assessment, problems, and possible solutions.
1475 Divers. Distrib. 12, 310-318.
1476 Whitehouse, N.J., P.G., L., Bustin, R., Galsworthy, S., 2008. Fossil insects and ecosystem
1477 dynamics in wetlands: implications for biodiversity and conservation. Biodivers. Conserv. 17,
1478 2055–2078.
1479 Williams, J.W., Shuman, B.N., Webb, T., III, Bartlein, P.J., Leduc, P.L., 2004. Late Quaternary
1480 vegetation dynamics in North America: Scaling from taxa to biomes. Ecological
1481 Monographs 74, 309-334.
1482 Wright, H.E., Jr. , Kutzbach, J.E., Webb, T., III, Ruddiman, W.F., Street-Perrott, F.A., Bartlein,
1483 P.J., 1993. Global Climates since the Last Glacial Maximum. University of Minnesota Press,
1484 Minneapolis.
1485 Zazula, G.D., MacPhee, R.D.E., Metcalfe, J.Z., Reyes, A.V., Brock, F., Druckenmiller, P.S.,
1486 Groves, P., Harington, C.R., Hodgins, G.W.L., Kunz, M.L., Longstaffe, F.J., Mann, D.H.,
1487 McDonald, H.G., Nalawade-Chavan, S., Southon, J.R., 2014. American mastodon
1488 extirpation in the Arctic and Subarctic predates human colonization and terminal Pleistocene
1489 climate change. Proceedings of the National Academy of Sciences 111, 18460-18465.
1490 Zdanowicz, C.M., Zielinski, G.A., Germani, M.S., 1999. Mount Mazama eruption: calendrical
1491 age verified and atmospheric impact assessed. Geology 27, 621-624.

Zhang, Z., Zhao, M., Eglinton, G., Lu, H., Huang, C.-Y., 2006. Leaf wax lipids as paleovegetational and paleoenvironmental proxies for the Chinese Loess Plateau over the last 170kyr. Quaternary Science Reviews 25, 575-594.

SUPPLEMENTARY MATERIAL

Training and Help Resources

Neotoma PIs, Stewards, and developers have built and maintain a variety of resources designed to help orient new users, support educational and outreach activities, and to report issues. These are currently somewhat dispersed; in the future, we aim to better consolidate these resources. See Table S1 below for a full listing of Neotoma online resources.

***Training Workshops*.** We run several kinds of in-person training workshops and webinars. Some workshops are designed to support Neotoma data users and describe the tools available for exploring, viewing, and obtaining Neotoma data via Neotoma Explorer and the R *neotoma* package. Recent user-oriented workshops were held at the International Biogeography Society (2017), the American Quaternary Association (2016), the Society for Vertebrate Paleontology (2016), the European Pollen Database meeting (Giesecke et al. 2016), at PalEON Boot Camp (2014, 2016), the International Symposium on Ostracoda (2017), and in Chile for South American paleoecologists (Latorre et al., 2014). Other workshops train Data Stewards in how to prepare and upload data via Tilia. These Steward-oriented workshops have been provided as several-hour webinars, and as several-day workshops. We held Steward workshops at the American Quaternary Association (AmQua) 2016 biennial meeting and the European Pollen Database (2016). Instructional materials from these Steward- and user-oriented workshops and a template for developing new workshop materials are available on GitHub (https://github.com/NeotomaDB/Workshops). More workshops are planned.

Neotoma collaborated with the University of Wisconsin Cartography Lab to hold a day-long Design Challenge mapping workshop in February 2016, bringing together students in cartographic design and paleoecologists to explore new approaches for visualizing paleodata (https://www.earthcube.org/workspace/c4p/cartography-lab-design-challenge). Neotoma helped organize a Community Development Workshop (also called a Paleodata Hackathon) in June 2016, led by the Cyberinfrastructure for Paleogeosciences Research Coordination Network, in

which interested scientific users and developers convened to explore, use, and link data from Neotoma, the Paleobiology Database, Macrostrat, and other resources.

Manuals and Workbooks. The Neotoma Manual is available at http://neotoma-manual.readthedocs.org/en/latest/. A Tilia Manual is available at http://tilia-manual.readthedocs.io/en/latest/. Tilia help is also available from https://www.tiliait.com/help/. Sample workbooks are also available that provide users with standard code and fully realized workflows for common analyses of paleoecological data (http://neotomadb.github.io).

Educational Resources are housed on the Neotoma website (https://www.neotomadb.org/education/category/higher_ed/) and at the Carleton SERC website (http://serc.carleton.edu/neotoma/index.html). These are varied, but most are interactive lab-based exercises, primarily geared towards college-level or high school courses. We welcome development of Neotoma-related educational materials and are happy to host or link to such content on the Neotoma home website. Interested contributors should contact the Neotoma Education and Outreach Committee.

Several **Visualization Resources** are in development and are intended for a mixture of educational, outreach, and research purposes. The Flyover Country app (http://fc.umn.edu/) is intended as an outreach and research tool that makes geological data available to any traveler with a smartphone (Loeffler et al., 2015), and showcases Neotoma fossil data holdings. An interactive map-based visualization, called Ice Age Mapper, is in development and is intended to replace the now-defunct Pollen Viewer (Williams et al., 2004). Pollen Viewer is no longer compliant with Java security standards and relied upon static images; Ice Age Mapper will dynamically link to Neotoma data holdings.

Issue Tracking and Code Repositories. Most code for Neotoma is stored in GitHub, in the NeotomaDB organization (https://github.com/NeotomaDB/). Individual code repositories within this organizational domain include the repository for minting DOIs (https://github.com/NeotomaDB/AssignDOIs), the *neotoma* R package (https://github.com/NeotomaDB/neotoma, forked from ropensci/neotoma), and code for mapping Neotoma to the DarwinCore standard (https://github.com/NeotomaDB/DwC-Mapping). The Neotoma API code is not currently open, but will be moved to an open repository in the near future (https://github.com/NeotomaDB/Neotoma-API). Issue tracking for Neotoma's Stratigraphic Diagrammer within Neotoma Explorer is available at https://bitbucket.org/neotomadb/stratigraphicdiagram/issues but we plan to migrate this to GitHub.

Table S1: Summary of on-line resources

	URL	Description
Main Page	www.neotomadb.org	Main website for Neotoma Paleoecology Database
Data Access and Analysis	www.github.com/NeotomaDB	GitHub code repository for Neotoma DB and associated software
	http://apps.neotomadb.org/explorer/	Neotoma Explorer interface for finding, downloading, and quickly visualizing data
	http://api.neotomadb.org	APIs for programmatic access to Neotoma
	https://cran.r-project.org/web/packages/neotoma/index.html	CRAN repository and download source for *neotoma* R package
	https://github.com/ropensci/neotoma	GitHub code repository for neotoma R package
	http://www.neotomadb.org/snapshots	Archived snapshots of the Neotoma database
	https://dx.doi.org/10.6084/m9.figshare.3376393.v1	Example of an archived snapshot posted to figShare.
	https://data.neotomadb.org/datasets/1001/	Example of a digital object identifier (DOI) and associated landing page for an individual Neotoma dataset.
Data Contributions	https://www.neotomadb.org/data/category/contribution	Starting page for scientists interested in contributing data to Neotoma
	https://www.tiliait.com/	Website for downloading Tilia software and licenses.
Governance, Membership, Policy	http://www.neotomadb.org/about/category/governance	Neotoma's Governance Policy and By-Laws
	http://bit.ly/2tzjEsZ	List of Data Stewards (Google Docs)
	https://tinyurl.com/NeotomaMember	Webform to sign up as Neotoma member
	http://www.neotomadb.org/data/category/use	Data Use Policy

Help and Training Resources	http://www.neotomadb.org/uploads/NeotomaManual.pdf	Neotoma Database Manual, with information about relational database
	http://neotoma-manual.readthedocs.org/en/latest/	A port of the Neotoma Database Manual to the readthedocs.org website, which
	https://www.tiliait.com/help/	Help page for Tilia software
	http://tilia-manual.readthedocs.io/en/latest/	Port of the Tilia manual to readthedocs.org
	https://github.com/NeotomaDB/FAUNMAP_Import	Example of a batch script for importing data into the Tilia .tlx format
	http://neotomadb.github.io	Sample scripts and analytical workflows making use of Neotoma data
Teaching Resources	https://www.neotomadb.org/education/category/higher_ed/	Access location for contributed teaching exercises and other educational researchers
	http://serc.carleton.edu/neotoma/index.html	Educational resources for college and high school courses developed in partnership with the Science Education Research Center (SERC) at Carleton
Allied and Linked Resources	http://www.earthlifeconsortium.org/api_v1/ui/	EarthLifeConsortium APIs that simultaneously search Neotoma and the Paleobiology Database.
	http://fc.umn.edu/	Flyover Country, a mobile app for discovering geological data during travel.
	http://www.gbif.org/	Global Biodiversity Information Facility. We are in the process of exporting Neotoma data to GBIF.
	https://globalpollenproject.org/	Global Pollen Project, a community platform for pollen images and identification
	(https://www.ncdc.noaa.gov/data-access/paleoclimatology-data	NOAA Paleoclimatology Portal, searches will retrieve Neotoma data

1556

Figure 1

Click here to download Figure

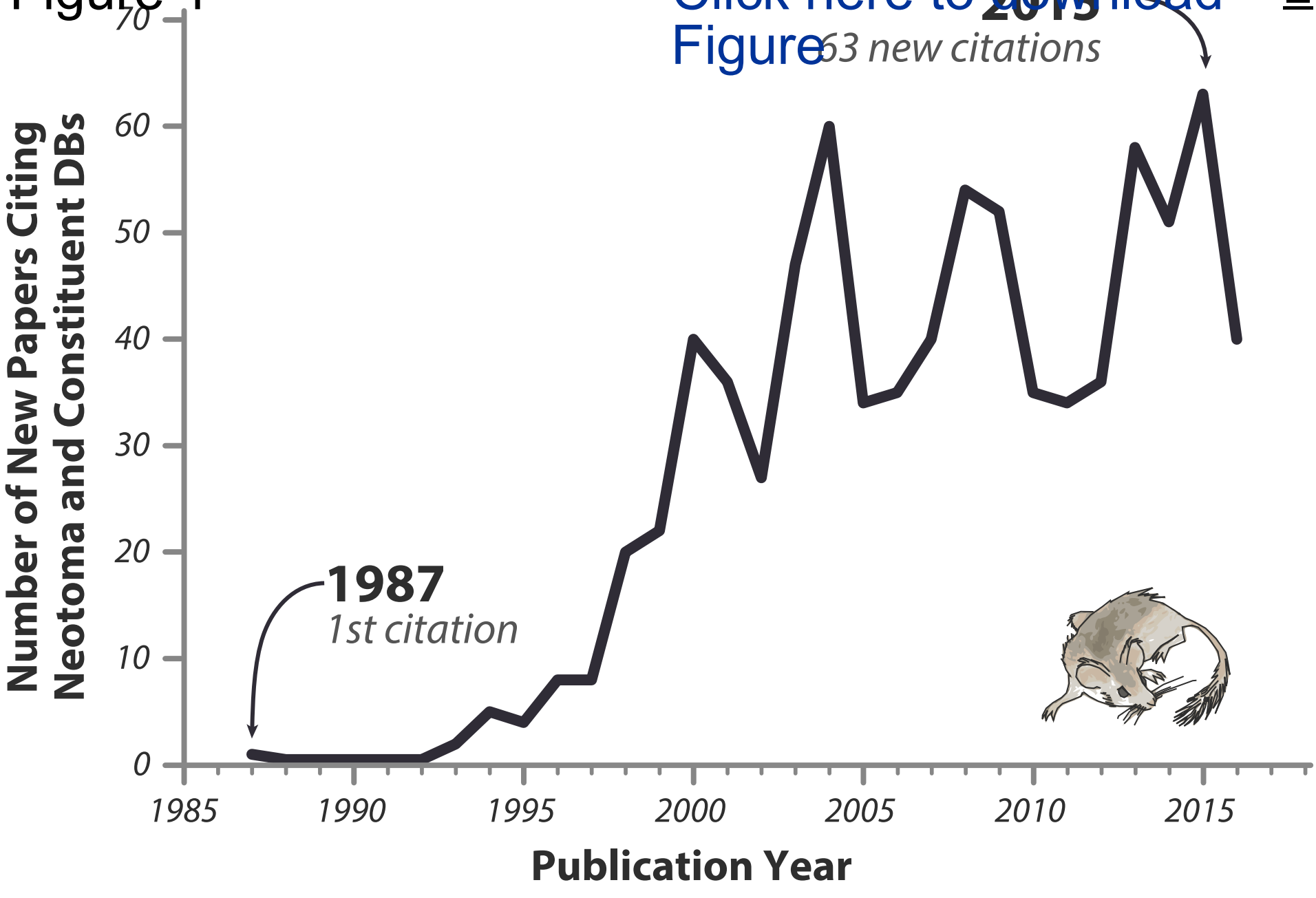


Figure 2

Click here to download Figure
FigFX2_NeotomaBoundaryOrganization_Rob_JWW.pdf

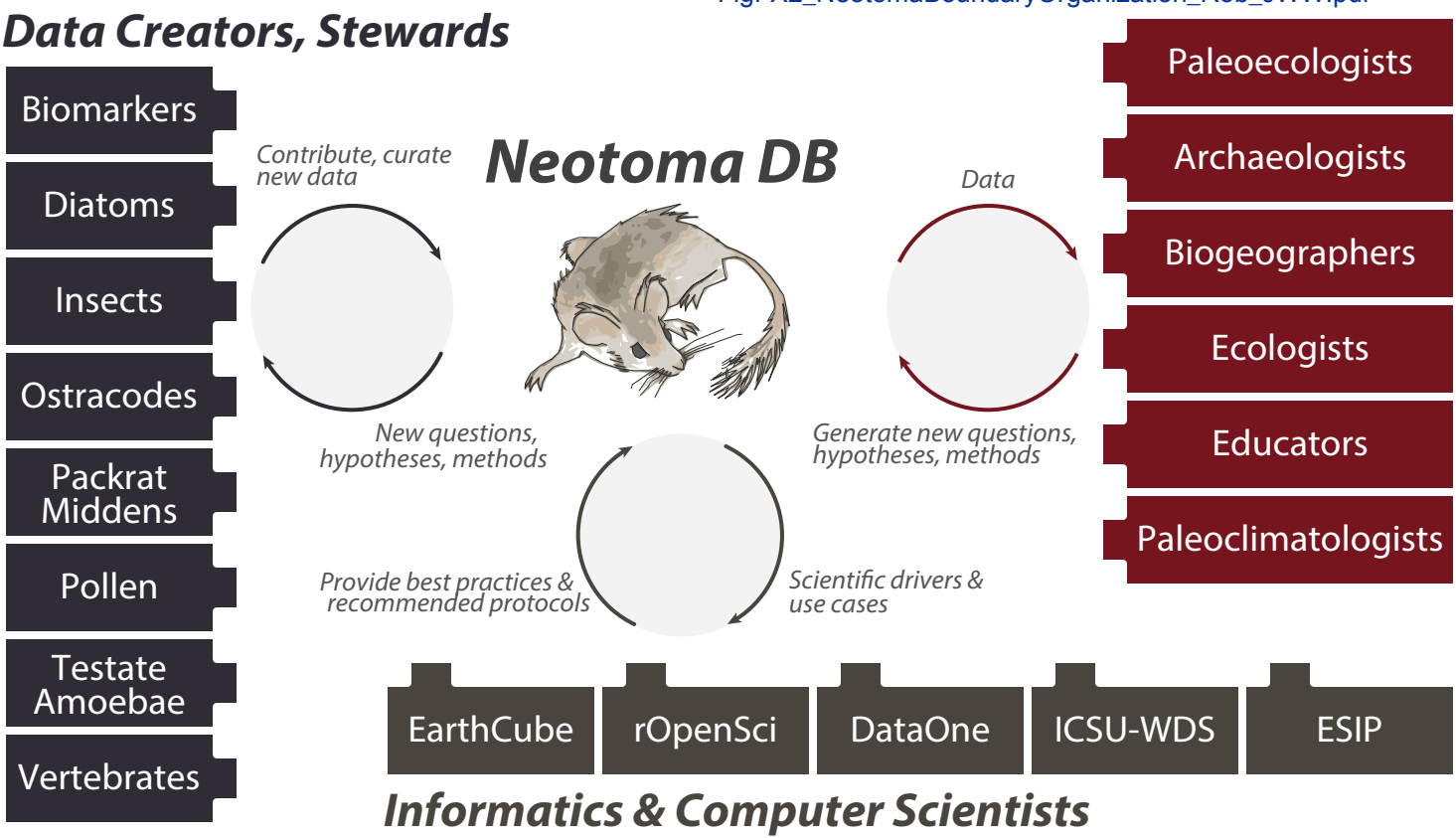


Figure 3

External Advisory Board

advises
networks
connects

Neotoma Leadership Council

external members

serves on

elects
serves on

Executive WG

Education & Outreach WG

Informatics & Technology WG

International Partnerships WG

Nominations & Membership WG

Neotoma Community

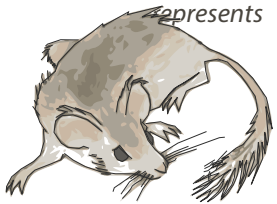
Data Contributors

Data Users

Educators

Researchers

represents



serves as

sets policy
appoints stewards
approves members

set upload priorities
vet & upload data
propose variable names
train data stewards
set default chronologies

Constituent Databases & Data Stewards

Biomarkers

Diatoms

Insects: BugsCEP

Insects: NA

Isotopes

Middens

Plant Macros

Pollen: APD

Pollen: EPD

Pollen: LAPD

Pollen: NAPD

Ostracodes

Taphonomy

Testate Amoebae

Vertebrates: Antigua

Vertebrates: FAUNMAP

Vertebrates: MIOMAP

...

Taxonomic Experts

approve names

Data Preparers

prepare data

Figure 4

Click here to download Figure
FigFX4_NeotomaSoftwareEcosystem_Rob_JWW.pdf

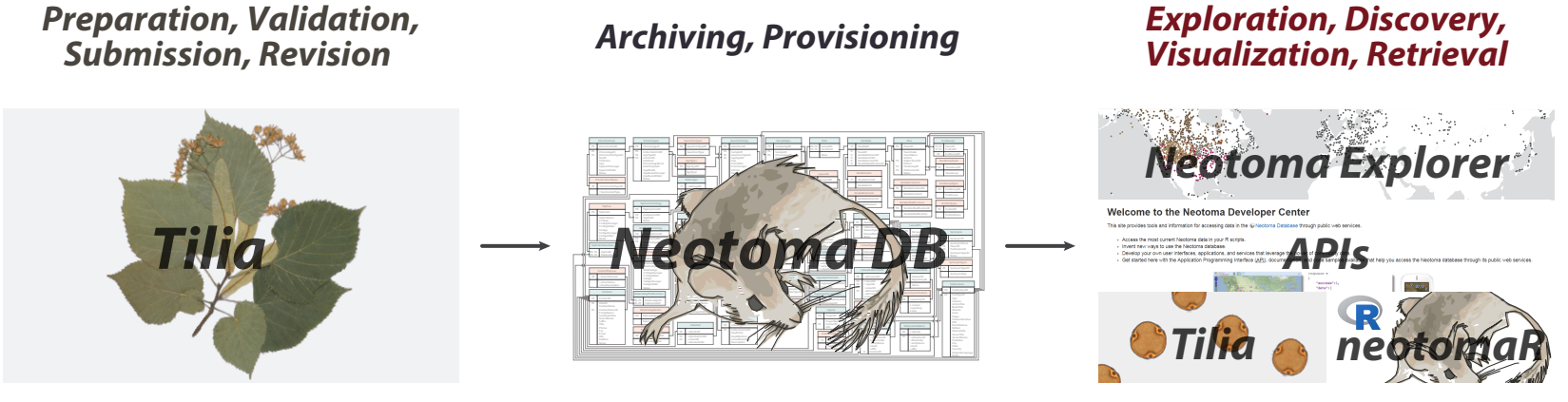


Figure 5

Click here to download Figure
Fig-X5_CoreComponents_Rob.pdf

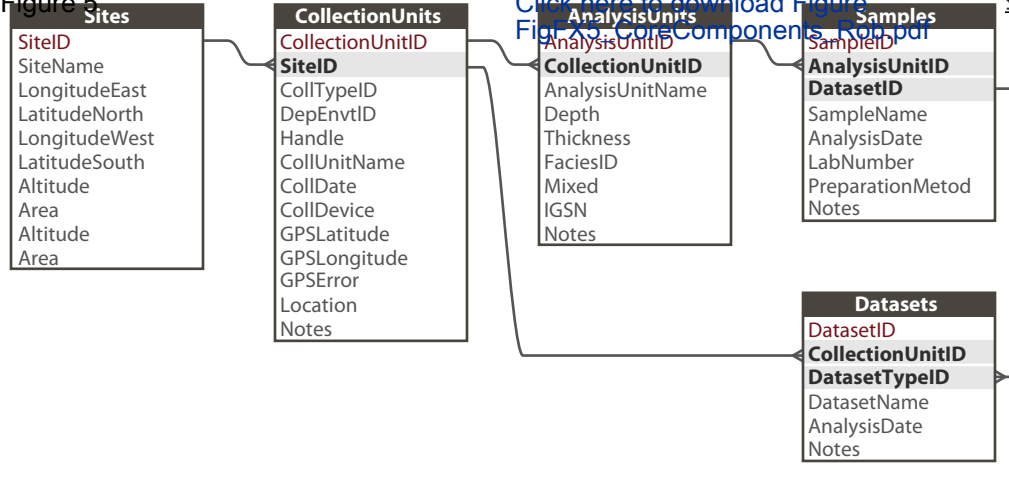


Figure 6

Click here to download Figure
FigFX6_CombinedGraphs_Rob.pdf

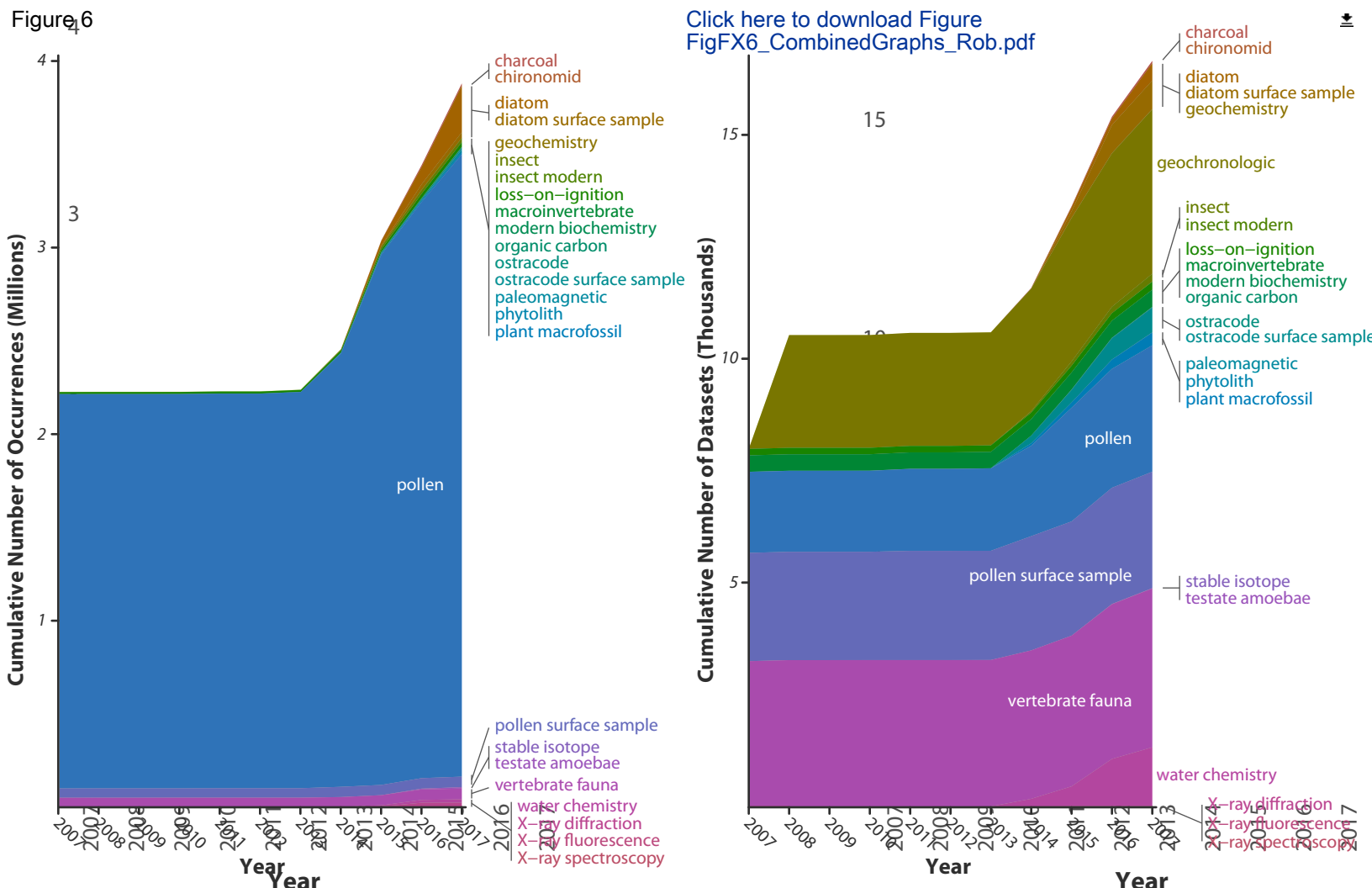


Figure 7

Click here to download Figure FigFX7_TiliaTaxonName.pdf