Interview date: January 3, 2019
Interviewee: Jack Williams (Neotoma)
Interviewer: Raleigh Martin (AAAS S&T Policy Fellow)
Document finalized: February 4, 2019

Disclaimer
All content presented here is the opinion of the interviewer and the interviewee alone and does not reflect an official position of the American Association for the Advancement of Science (AAAS), the U.S. National Science Foundation (NSF), or any NSF-funded institution.

About the data facility

The Neotoma Paleoecology Database ("Neotoma") is a facility for the curation, archiving, discovery, and reuse of paleoecological and paleoenvironmental data. Neotoma also provides a variety of tools for dataset access and processing, and it coordinates its activities with other paleobiological and geological data facilities.

Neotoma is housed at the Center for Environmental Informatics (CEI) at Pennsylvania State University. Operation of Neotoma is coordinated among Penn State, the University of Wisconsin-Madison, North Dakota State University, the University of Minnesota, Drexel University, Lehigh University, the University of California – Merced, the University of Illinois – Urbana-Champaign, and Kent State University. Snapshots of Neotoma data holdings are provided to Figshare, and plans are in progress to send full snapshots to NOAA's National Center for Environmental Information (NCEI) and customized subsetted snapshots to Neotoma's Constituent Databases. Neotoma is currently supported on a collection of four-year continuing grants through the NSF Geoinformatics program (award numbers: 1550707, 1550717, 1550805, 1550728, 1550716, 1550700, 1550890, 1550721, 1550755). Neotoma is a certified member of the International Science Council's World Data System and is endorsed by the American Quaternary Association and PAGES (Past Global Changes).

Neotoma is led by an elected 12-member Leadership Council and a team of PIs (Principal Investigators), collaborative institution PIs, and full-time staff. The lead PI and chair of the Neotoma Leadership Council is Jack Williams (Wisconsin), and the lead collaborative institution PIs are Jessica Blois (UC Merced), Robert Booth (Lehigh), Allan Ashworth (ND State), Allison Smith (Kent State), Donald Charles (Drexel), Russell Graham (Penn State), Eric Grimm (Minnesota), and Brandon Curry (Illinois). This large collection of PIs and councilors represents a range of expertise across paleoecology, with PIs acting as trusted ambassadors to their respective research communities. PI Eric Grimm designed and maintains the Neotoma data model and the associated Tilia software system for data ingest and quality control. In addition to these PIs, Neotoma employs one full-time software developer at Wisconsin (Simon Goring) and two part-time developers at Penn State. Undergraduate and masters students at some PI institutions receive hourly pay to help with data preparation processes. Otherwise, Neotoma relies on a network of trained volunteer data stewards (currently, about 50), who bring domain-specific expertise to the establishment of data standards and the curation of submitted data

objects. The work structure for Neotoma reflects a desire for maintaining a variety of expertise areas on a relative shoestring budget, while also encouraging sustainability via community buy-in and "ownership" of Neotoma.

Neotoma was launched in 2009 to bring together a variety of smaller paleoenvironmental data resources under a single open-source platform. Some of these constituent data resources have a long data curation history with strong community buy in (Grimm et al. 2018). For example, Eric Grimm has led operation of a pollen database for decades (Global Pollen Database), while Russ Graham led a similar initiative for vertebrate fossils (FAUNMAP). Other research communities have more recently become interested in merging their data resources, with Neotoma's support. Based on this background, Neotoma continues to support specialized curation of multiple sub-resources under the Neotoma umbrella, but all sub-resources are now built on a common data architecture for streamlined technical management. Much of the growth of Neotoma has been driven by specific data mobilization campaigns (each of which has been driven by a specific funding source), which have gradually built the data holdings and capabilities of the Neotoma resource. As of Nov. 2017, Neotoma contained over 3.8 million data records from 17,275 data sets and 9,269 sites. The history of Neotoma's formation and details of its operation are provided in Williams et al. (2018a) and Grimm et al. (2018).

Data facility services

Neotoma's mission is to build, maintain, and curate a resource for the submission, discovery, usage, and interoperability of paleoecology and associated paleoenvironmental data. Neotoma operates a centralized SQL-based relational database containing all its data holdings, which are organized around a common spatiotemporal paleoecological data model that stores information about species occurrences and abundances. Neotoma also stores primary measurements of geological age (e.g., radiocarbon dates) and derived age models. The age models are frequently updated by data curators and contributors, as age-depth models improve. Built on this common data model are a variety of "Constituent Databases" (e.g., North American Pollen Database, North American Non-marine Ostracode Database, FAUNMAP, Neotoma Biomarker Database), which specialize by data types and regions and are overseen by taxonomic experts. Though accommodating data from samples of all ages, Neotoma's focus is on data spanning the Neogene and Quaternary Periods (~23 Ma – present).

In support of this work, the services provided by Neotoma fall into five general areas: (1) data curation and preservation, (2) creation of data standards, (3) enabling data discovery and reuse, (4) training and education, and (5) governance. More specifically, the work performed by Neotoma staff includes basic maintenance of the data archival server, operation of a "Neotoma Explorer" map-based interface for data discovery, development of application program interfaces (APIs) and an R package to enable direct user interactions with data holdings, and partnerships with third parties (e.g., Flyover Country, the Global Pollen Project) to leverage the value of Neotoma holdings. In addition, Neotoma is a member of the EarthLife Consortium (ELC), a federation of paleobiological repositories (Neotoma, the Paleobiology Database, and

the Strategic Environmental Archaeology Database) working to build a resource for a single federated search across paleo science data facilities. The ELC was initiated with NSF EarthCube support (award numbers 1540997, 1541002, 1540994, 1541015, 1540979, 1540977) for development of federated search APIs, and it now continues as the independent non-profit EarthLife Consortium Foundation.

Underlying all its efforts, Neotoma is governed through a distributed model that centralizes technical architecture but allocates decisions over standards for Constituent Databases to specific domain research communities. Neotoma's concept of "Constituent Databases," each of which has one or more data stewards authorized to upload or modify data, enables distributed scientific governance and expert curation of different kinds of data housed in one resource.

Neotoma also serves as a data resource for NSF PIs through the proposal and award process. Its website contains a template with boilerplate text for researchers who are generating data management plans (DMPs) for NSF proposals, though this resource is not yet commonly used by NSF PIs. In addition, Neotoma is listed among the recommended data resources for NSF's Division of Earth Sciences. Neotoma encourages PIs to budget for data stewardship training and/or curation services in their project budgets, though they are still willing to help any researcher who approaches them for help. Jack estimated that Neotoma gets a few dozen "out of left field" data deposit requests per year, mainly from researchers needing to urgently meet funder or journal requirement. Neotoma's team of data stewards try their best to meet these unsolicited curation requests, though this requires a certain amount of volunteer "heroism" not dissimilar to the work of reviewers of academic journal articles.

Though Neotoma is focused on serving a specific set of subdisciplines within the paleosciences, the specific boundaries of what data types it serves are not strictly defined and continue to grow. Such growth often occurs via science-oriented grants with funding dedicated to data mobilization, in which data aggregators gather large volumes of "dark" but important data from the literature and their networks of colleagues, clean and upload these data to Neotoma, then run a large-scale analysis on these cleaned and uploaded data. Neotoma leadership is active in community outreach to bring new research subdomains under its umbrella. To accomplish this, Neotoma runs workshops at scientific meetings, both for data users (i.e., those trying to extract data from the Neotoma resource) and for data stewards (i.e., those helping with the process of data deposit in the Neotoma resource). To maintain data quality control and standardization, trained data stewards are the only people who can deposit new datasets into Neotoma. Other researchers can deposit their data in Neotoma by reaching out to an existing data steward in their research subdomain, or by themselves going through the data stewardship training. The process for data deposit and curation is detailed below.

Data deposit and curation

Direct deposit of research data into Neotoma is limited to trained data stewards. For researchers to deposit their data, they must either coordinate their data deposit with one of Neotoma's existing data stewards or go through the rigorous data stewardship training

program themselves. Neotoma data stewards are trained to provide accurate metadata on domain-specific elements of data entries, such as time-to-depth relationships in stratigraphic columns and standards for taxonomy and systematics. Limiting data deposit to these data stewards ensures a high quality and standardization for Neotoma data entries. Within data entries, all entities involved in the data lifecycle are acknowledged: the original data creators, the originators of the age models used for data processing, and the data stewards providing the curation. Such documentation helps in provenance tracking and for follow-up by data users. When deposited prior to publication of associated journal articles, users may choose to set an embargo period, which normally expires at the time of publication or after two years, whichever happens first. Data generators may ask for embargo extensions with proper justification. This embargo policy has been voted on and set by Neotoma's Leadership Council, and implementation is underway.

Currently, Neotoma does not assign persistent identifiers (PID) to its data holdings, partially over concern about the need for immutability of PIDs versus the need to continually update Neotoma datasets. However, work is now in progress to assign Digital Object Identifiers (DOIs) to static instances of datasets, while keeping open the possibility of continuing to update datasets after initial DOI assignment. Neotoma DOIs have been assigned in beta form to datasets, and operational release is scheduled for spring 2019.

Data discovery and access

All data contained in Neotoma is subject to an open license (CC-BY) and can be accessed through the Neotoma Explorer portal, which provides for faceted map-based search filtered by age range and taxon. Neotoma data are also discoverable through Flyover Country, NOAA/NCEI-Paleoclimatology, and other resources built on the Neotoma API. Through the EarthLife Consortium project, Neotoma data will soon also be available through federated search with the Paleobiology Database. As Neotoma has grown to incorporate other existing data resources, there has been continued debate over when it is best to fully subsume existing datasets directly into the Neotoma data holdings, and when it is more practical to federate through APIs (as Neotoma is currently doing with the Paleobiology Database). Neotoma is also exploring the adoption of open metadata standards for discovery by independent web-based search tools (e.g., Google Dataset Search). For example, in its current process of rolling out DOIs for its data holdings, Neotoma plans to design DOIs to be discoverable through EarthCube's ongoing Project 419.

User base

The primary users of Neotoma (both for data deposit and discovery) are paleoecologists and paleobiologists who deal with sample-based terrestrial records. However, the user base for utilization of Neotoma data holdings spans a wide range of scientific domains. For example, biogeographers, macroecologists, and conservation biologists are interested in using Neotoma data to track global biodiversity over time. Paleoarchaeologists and paleoclimatologists, such as those associated with the SKOPE (Synthesizing Knowledge of Past Environments) and PAGES

(Past Global Changes) projects, are interested in Neotoma data to help inform studies of past societies and climate change. A recent issue of *PAGES Magazine* (Williams et al., 2018b) highlights some of these efforts. Educators have developed open education modules based on Neotoma data holdings and its R programming tools (Goring et al. 2018).

Aside from this anecdotal understanding of the usage and value of Neotoma data holdings, Neotoma managers have prioritized development of new data services over detailed tracking of existing resources. Furthermore, user access to data holdings is anonymous, so rigorous tracking of data usage is difficult. Such tracking may be improved soon, when DOIs are implemented for Neotoma data holdings. The EarthCube-funded THROUGHPUT project (award numbers 1740697, 1740669, 1740667, 1740699, 1740663), which includes many Neotoma data stewards on its team, is working on development and adoption of data standards and API tools for community-curated data repositories like Neotoma. It is hoped that THROUGHPUT will facilitate more rigorous dataset provenance tracking, including through integration with the ORCID API for associating data products with individual researchers.

Prospects and challenges

Jack views Neotoma as filling an important "mesoscale" niche in the ecosystem of data resources for the geosciences: large enough to build a robust and sustained cyberinfrastructure for research data preservation and discovery, but small enough to support the domain-specific needs of individual long-tail research communities generating relatively small datasets. Its data model for storing and linking age controls, age models, and age inferences is advanced relative to similar resources in the field. Operating on a relative shoestring budget, its web services may not be the most polished and up-to-date relative to other communities, but its small size allows it to be relatively nimble and responsive to specific research community needs.

As it grows, Neotoma has generated a lot of interest from paleontologists and paleoecologists who traditionally have lacked data resources that serve their communities. Though certain subdomains have strong data management traditions, others are only now becoming aware of the need to document and preserve their data according to rigorous standards. In addition, there is a wide variety of paleo research communities with a variety of data curation traditions. For example, curation and preservation of macro-fossil samples (e.g., vertebrate bones) has typically been the preserve of museums, while micro-fossil samples are usually housed in individual research labs.

As with many other data facilities, a major concern for Neotoma is over sustainability of its operation and holdings. Given the specificity and academic nature of the resources curated by Neotoma, it is implausible to expect that they could ever depend on a private-sector funding model. Neotoma's existing funding model depends almost entirely on NSF support, much of which goes toward the growth of new features and data holdings rather than the management and continued modernization of the existing technical architecture. In partnership with the Paleobiology Database, Neotoma has launched a new non-profit, called the EarthLife Consortium Foundation, as an umbrella organization for private fund-raising and supporting electronic paleodata resources. The ELC Foundation, which is a follow-on to the two-year NSF

EarthLife Consortium award (see "Data Facility Services" above), has attracted seed funding from the Society for Vertebrate Paleontology and the Paleontological Society. Jack expressed interest in obtaining training and exposure to examples that would enable Neotoma to transition to a blend of support from professional societies, membership dues, fee-for-service, and traditional funding agencies (e.g., NSF). Jack was aware of EarthCube's Council for Data Facilities, but had the impression that this body primarily focuses on the needs of big data facilities rather than the mesoscale role of Neotoma and other community-curated data resources.

In general, Jack believes that the primarily project-driven funding that has supported Neotoma over the years has helped it to remain responsive to community needs while also building its back-end cyberinfrastructure. He also stressed the positive and growing community buy-in and the opportunities to cross-leverage data services supported by NSF funding with services provided by disciplinary experts (e.g. data contributions, error finding and correction, updates to age models and taxonomies, data use). However, he did note that some increased level of long-term background funding for the less glamorous tasks of upkeep and development of existing resources would provide a welcome complement to existing project-based lines of support. Despite these sustainability challenges, Jack noted that Neotoma has generally been well served by NSF support up until now.

References

Goring, S.J., Graham, R., Loeffler, S., Myrbo, A., Oliver, J.S., Ormond, C. and Williams, J.W. (2018) *The Neotoma Paleoecology Database: A Research Outreach Nexus*. Cambridge: Cambridge University Press (Elements of Paleontology). doi: 10.1017/9781108681582

Grimm, E.C., Blois, J.L., Giesecke, T., Graham, R., Smith, A.J., Williams, J.W., 2018. Constituent databases and data stewards in the Neotoma Paleoecology Database: History, growth, and new directions. *Past Global Changes Magazine 26*, 64–65. https://doi.org/10.22498/pages.26.2.64

Uhen, M.D., Goring, S., Jenkins, J., Williams, L., 2018. EarthLife Consortium: Supporting digital paleobiology. *Past Global Changes Magazine 26*, 78–79. https://doi.org/10.22498/pages.26.2.78

Williams, J.W., Grimm, E.C., Blois, J.L., Charles, D.F., Davis, E.B., Goring, S.J., Graham, R.W., Smith, A.J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A.C., Betancourt, J.L., Bills, B.W., Booth, R.K., Buckland, P.I., Curry, B.B., Giesecke, T., Jackson, S.T., Latorre, C., Nichols, J., Purdum, T., Roth, R.E., Stryker, M., Takahara, H., 2018a. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research 89*, 156–177. https://doi.org/10.1017/qua.2017.105

Williams, J.W., Newton, A.J., Kaufman, D.S., von Gunten, L. (Eds.), 2018b. Building and Harnessing Open Paleodata. *Past Global Changes Magazine 26*, 45–96. https://doi.org/10.22498/pages.26.2