The Neotoma Paleoecology Database: Current Infrastructure, Ongoing Challenges, and Future Directions

Oh behalf of the Neotoma DB Consortium: John Williams, Eric Grimm, Don Charles, Ed Davis, Simon Goring, Russ Graham, Alison Smith, Mike Anderson, Allan Ashworth, Julio Betancourt, Brian Bills, Bob Booth, Phil Buckland, Brandon Curry, Thomas Giesecke, Sonja Hausmann, Steve Jackson, Claudio Latorre, Doug Miller, Jonathan Nichols, Timshel Purdum, Rob Roth, Hikaru Takahara, and many many others

Jessica Blois
University of California, Merced
OVERVIEW

- Typical paleoecological data and data management approaches
- Neotoma: Current status
- Neotoma: Future directions
- Neotoma: Challenges and gaps

See two resources for additional information:

- Williams et al. 2018, Quaternary Research 89: 156-177. *The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource*
- Williams et al. whitepaper posted to Authorea (https://goo.gl/ZopKco). *Cyberinfrastructure in the Paleosciences: Mobilizing Long-Tail Data, Building Distributed Community Infrastructure, Empowering Individual Geoscientists*
KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

- **‘Long Tail’**: Collected in the field & lab by many individuals and scientific teams.
- **Heterogeneity**: Many kinds of measurements & methods
- **Distributed Scientific Expertise**: By proxy type, archive type, region, time period, and/or taxonomic group
- **Uneven Workforce** training and interest in informatics
- **Commonality**: Most datasets involve measurements of proxies in various geological archives by depth, from which we estimate time.
- **Long Shelf Life**: Specimens & samples collected decades ago can be re-analyzed
- **Useful**: Increasingly assimilated with Earth System Models and conservation biology
TRADITIONAL DATA MANAGEMENT (ESP. FOR ‘SMALL’ DATA)

Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in the mundane labor of collecting and preparing data, before it can be explored for useful information. - NYTimes (2014)

- Files on individuals computers (may or may not be backed up)
- Details in field/lab notebooks, not captured, mis-transcribed, forgotten...
- Easy to view all/most data in one spreadsheet
- Files passed back and forth by email
- Potentially different versions floating around as revisions are made
MOVING SMALL DATASETS TOWARDS BIG DATA APPROACHES

- What we want: paleobiological data that are easily discoverable, interpretable, accessible, and analyzable

- Generic Depositories vs. Community-Led Repositories

“... data have no value or meaning in isolation; they exist within a knowledge infrastructure — an ecology of people, practices, technologies, institutions, material objects, and relationships.” - C.L. Borgman

Modified from K. Lehnert
NEOTOMA PALEOECOLOGY DATABASE: ECOSYSTEM

- Community-curated database consortium focused on Pliocene to Quaternary data from around the world.

**Proxies**
- Biomarkers
- Diatoms
- Insects
- Ostracodes
- Packrat Middens
- Pollen
- Testate Amoebae
- Vertebrates

**Neotoma DB**
- Contribute new data to
- Generates new questions & methods for
- Add best practices & common protocols to
- Provides scientific drivers & use cases for

**Data Users**
- Paleoecologists
- Archaeologists
- Biogeographers
- Ecologists
- Educators
- Paleoclimatologists

**Informatics & Computer Scientists**
- EarthCube
- rOpenSci
- DataOne
- WDS-ICSU
- ESIP

*Williams et al. 2018, Quaternary Research*
NEOTOMA: KEY FEATURES

- Spatiotemporal database: species occurrences & abundances in space and time
- Age controls and age models stored
- Composed of constituent databases (e.g. North American Pollen Database, European Pollen Database, FAUNMAP, NANODE)
  - ...but with centralized IT and distributed scientific governance
- Open data accessible via Explorer, APIs, an R package
- Broad user community: Paleoecologists, ecosystem modellers, paleoclimatologists, biogeographers, educators, ...
- Broad community support and partnerships: with NOAA, PaleobiologyDB, Linked Earth, etc.

Williams et al. 2018, Quaternary Research
- Pollen: NAPD, EPD, et al.
- Vertebrates: FAUNMAP2+, MIOMAP, ANTIGUA, MQMD
- Ostracodes: NANODE
- Diatoms: Drexel DB
- Testate Amoebae
- Packrat Middens
- Organic Biomarkers

- Current status: >3.8 million observations, >17,000 datasets, and >9,200 sites.

Williams et al. 2018, Quaternary Research
NEOTOMA: NEW DEVELOPMENTS

- Recently finished or coming down the pipeline
  - Specimen-level data (BETA RELEASE)
  - Stable isotopes as new data type (BETA RELEASE)
  - DOI assignments to datasets (ALMOST DONE)
  - Ice Age Mapper (UNDERWAY)
  - Embargoes for unpublished data (STARTED)
  - Webform uploader to contribute data (A TWINKLE IN THE EYE)

- **Earth-Life Consortium** (http://earthlifeconsortium.org/): seeks to make all *paleobiological* data easily *discoverable, accessible, and analyzable*, with the larger goal of understanding the interactions between the Earth’s biological and geophysical systems across all timescales of the Earth’s history.
NEOTOMA: KEY CHALLENGES

- **Reducing data friction** along the pipeline from collection to final archiving
  - Science-driven data-mobilization or data-rescue campaigns
  - Development of easy-to-use, scalable, extendable, multi-platform input and upload tools that support data validation and quality control
  - Ability to query across different data repositories

- **Funding sustainability**, particularly related to supporting the geoinformaticists necessary for database maintenance and development
THANKS!

- **Neotoma Executive Committee**
  - Chair: Jack Williams
  - Associate Chair: Jessica Blois
  - Alison Smith
  - Eric Grimm

- **Neotoma Leadership Council**
  - EC +
  - Allan Ashworth, International Working Group Chair, Steward, Insects
  - Suzanne Pilaar Birch, Steward, Isotopes
  - Phil Buckland, Steward, Insects
  - Don Charles, Steward, Diatoms
  - Thomas Giesecke, International Working Group, Steward, European Pollen Database
  - Simon Goring, IT Working Group Chair
  - Claudio Latorre, International Working Group, Steward, Packrat Middens
  - Hikaru Takahara, International Working Group, Steward, Japan Pollen Database

- **Neotoma database contributors**

- **Funding**
  - NSF EAR 1550700
  - NSF ICER 1540977
NEOTOMA: ACCESSING DATA

- Finding, Exploring, Downloading Data
  - Explorer
    - https://apps.neotomadb.org/Explorer/
  - APIs
    - https://api.neotomadb.org/
  - R
    - https://cran.r-project.org/web/packages/neotoma/index.html
    - https://github.com/ropensci/neotoma
  - DOIs & Landing Pages (coming soon)
    - http://data.neotomadb.org/datasets/1001/
NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

  - Become a member
  - Use Neotoma Data
  - Cite Neotoma and Contributors
  - Contribute data
  - Become a Data Steward
  - Launch a Constituent Database
  - Spread the Word

More info:
- Become a member: [https://www.neotomadb.org/about/category/about](https://www.neotomadb.org/about/category/about)
- Use Neotoma Data: [Explorer, APIs, R](https://www.neotomadb.org/about/category/about)
- Contribute data: [https://www.neotomadb.org/data/category/contribution](https://www.neotomadb.org/data/category/contribution)
- Become a Data Steward: [neotoma-contact@googlegroups.com](mailto:neotoma-contact@googlegroups.com)
- Launch a Constituent Database: [Convene a working group](https://www.neotomadb.org/about/category/about)
- Spread the Word: [Talk to your colleagues](https://www.neotomadb.org/about/category/about)
NEOTOMA: EDUCATION

- Teaching Resources
  - SERC Carleton
    - [http://serc.carleton.edu/neotoma/activities.html](http://serc.carleton.edu/neotoma/activities.html)

- Neotoma Webpage
  - [https://www.neotomadb.org/education/category/higher_ed/](https://www.neotomadb.org/education/category/higher_ed/)