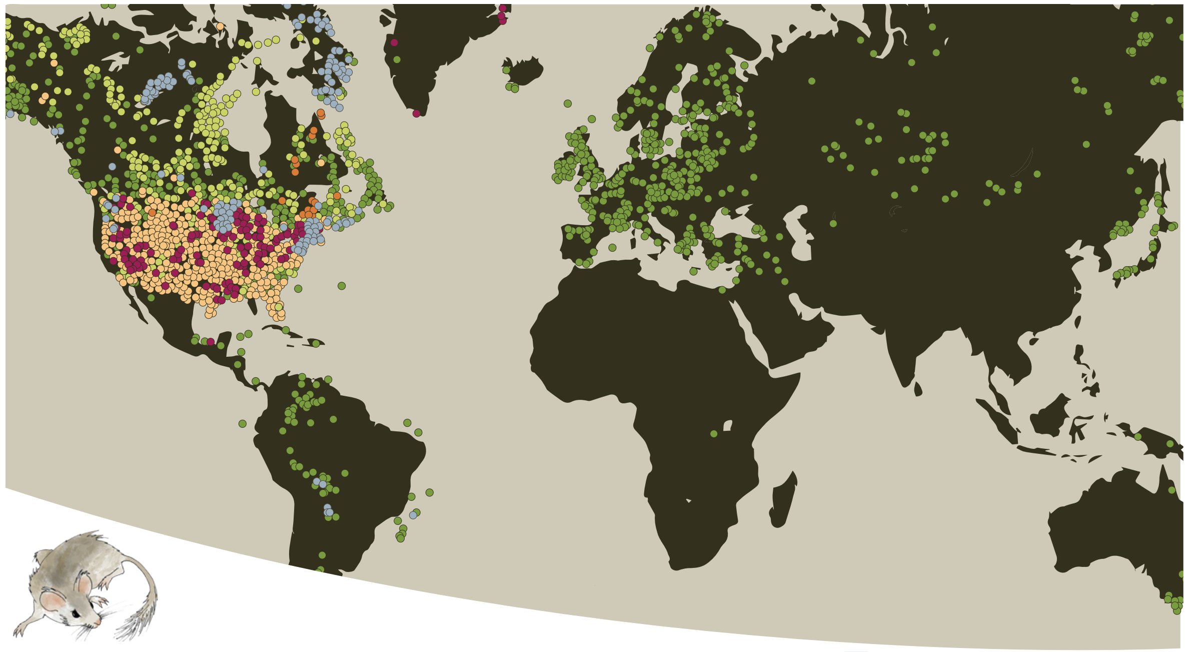
The Neotoma Paleoecology Database: Current Infrastructure, Ongoing Challenges, and Future Directions



NEOTOMA PALEOECOLOGY DATABASE
neotomadb.org

Jessica Blois
University of California, Merced

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

NSF-Geoinformatics,
NSF-EarthCube

@neotomadb
@jessicablois

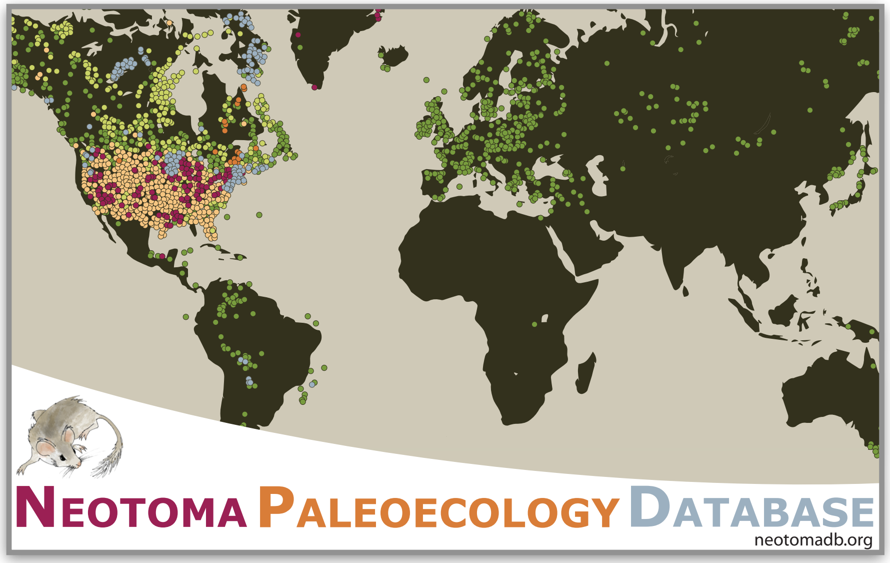
OVERVIEW

- Typical paleoecological data and data management approaches

- Neotoma: Current status

- Neotoma: Future directions

- Neotoma: Challenges and gaps



neotomadb.org

- 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

Bad news

Good news

KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

Bad news

- **'Long Tail':** Collected in the field & lab by many individuals and scientific teams.

Good news

KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

Bad news

- **'Long Tail':** Collected in the field & lab by many individuals and scientific teams.
- **Heterogeneity:** Many kinds of measurements & methods

Good news

KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

Bad news

- **'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

Good news

KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

Bad news

- **'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

Good news

KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

Bad news

- **'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

Good news

- **Commonality:** Most datasets involve measurements of proxies in various geological archives by depth, from which we estimate time.

KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

Bad news

- **'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

Good news

- **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

KEY CHARACTERISTICS OF PALEOECOLOGICAL DATA

Bad news

- **'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

Good news

- **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)

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)

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)

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…

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

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

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

MOVING SMALL DATASETS TOWARDS BIG DATA APPROACHES

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

MOVING SMALL DATASETS TOWARDS BIG DATA APPROACHES

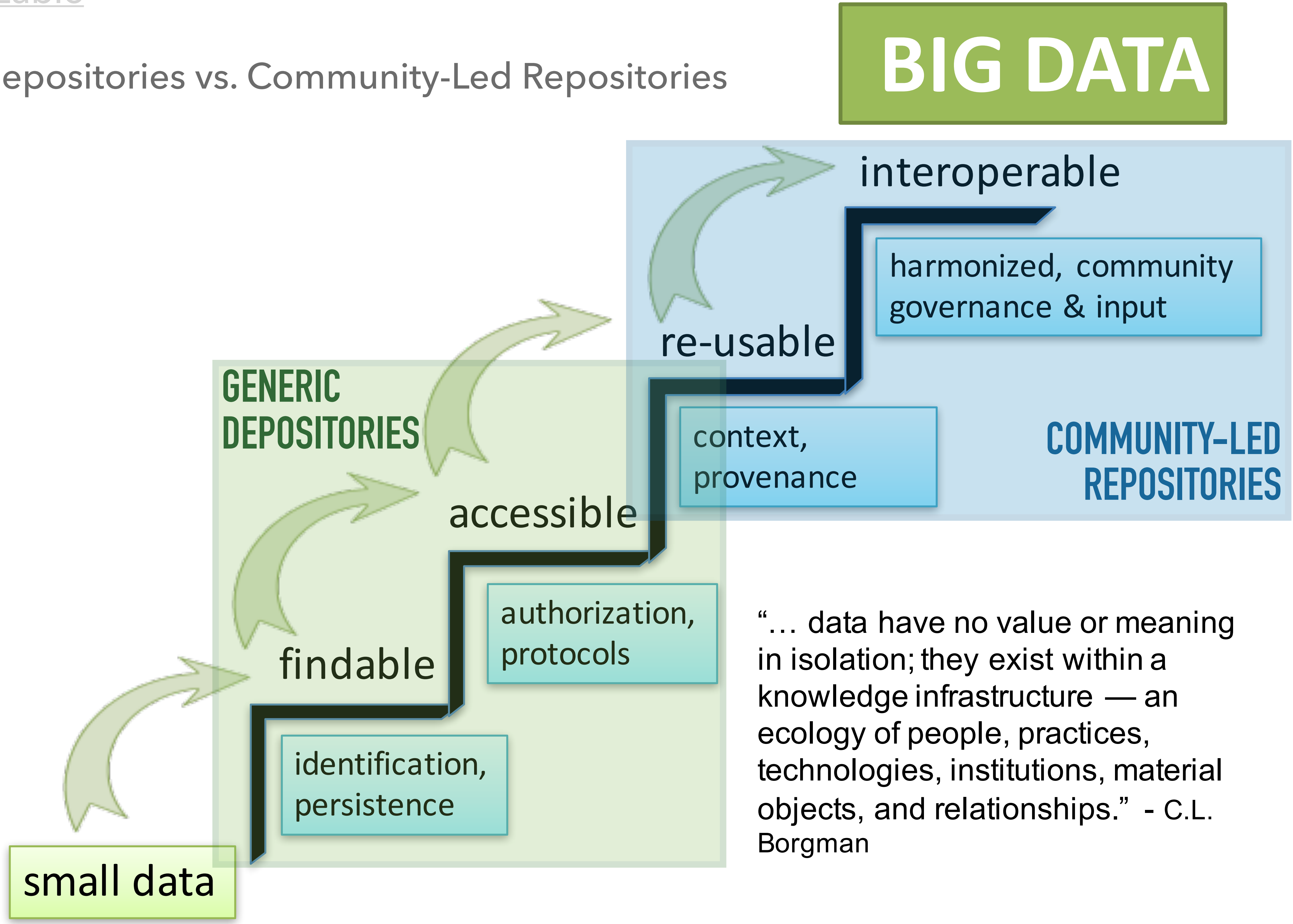
- What we want: paleobiological data that are easily discoverable, interpretable, accessible, and analyzable

- Generic Depositories vs. Community-Led Repositories

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

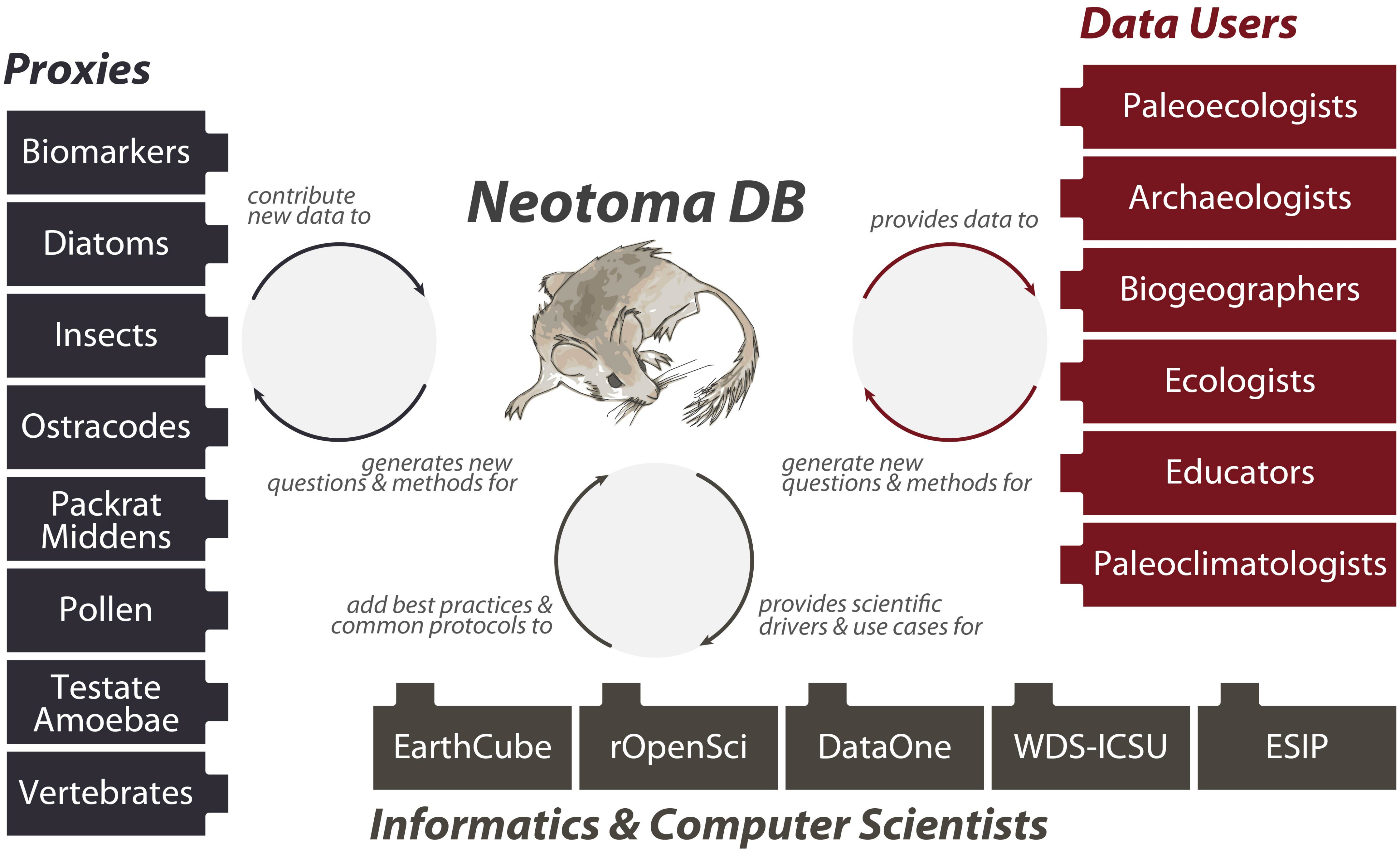
Modified from K. Lehnert

NEOTOMA PALEOECOLOGY DATABASE: ECOSYSTEM

Williams et al. 2018, Quaternary Research

NEOTOMA PALEOECOLOGY DATABASE: ECOSYSTEM

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



Williams et al. 2018, Quaternary Research

NEOTOMA: KEY FEATURES

- Spatiotemporal database: species occurrences & abundances in space and time

Williams et al. 2018, Quaternary Research

NEOTOMA: KEY FEATURES

- Spatiotemporal database: species occurrences & abundances in space and time
- Age controls and age models stored

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)

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

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

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, …

Williams et al. 2018, Quaternary Research

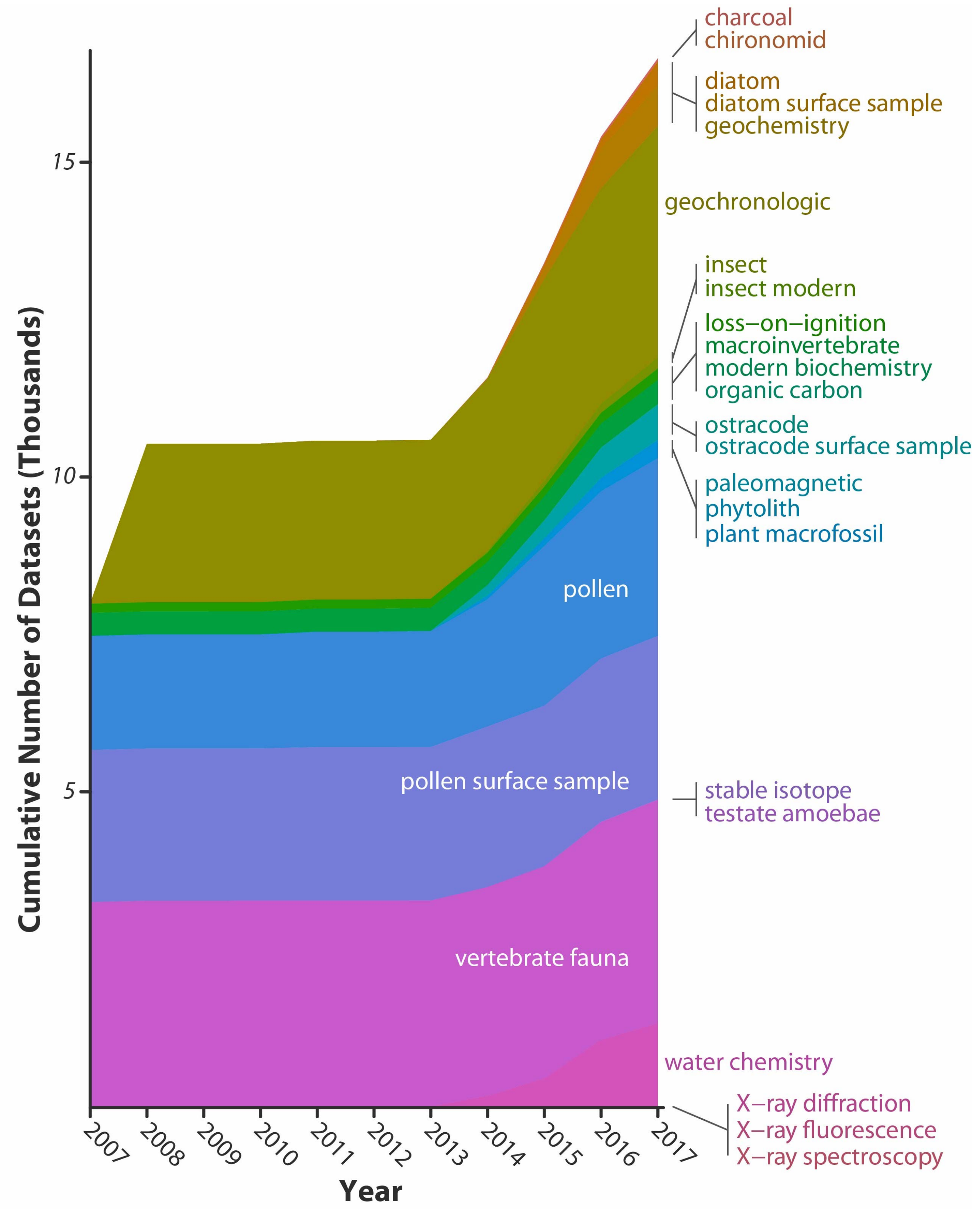
NEOTOMA: KEY FEATURES

- 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

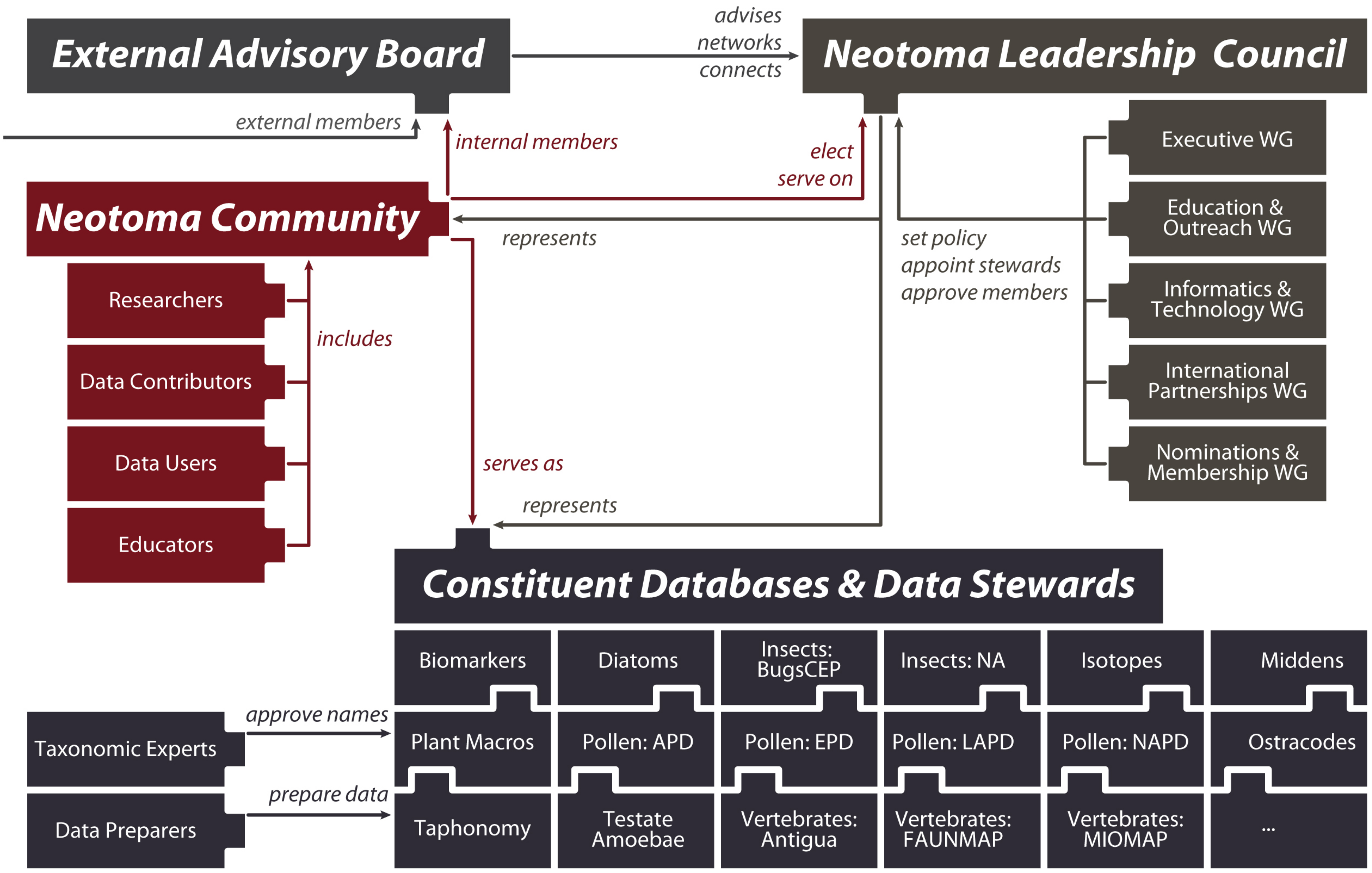
NEOTOMA: DATA MOBILIZATION CAMPAIGNS

- 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: GOVERNANCE



External Advisory Board

advises networks connects

Neotoma Leadership Council

external members

internal members

elect serve on

Executive WG

Education & Outreach WG

Informatics & Technology WG

International Partnerships WG

Nominations & Membership WG

Neotoma Community

represents

set policy appoint stewards approve members

Researchers

Data Contributors

Data Users

Educators

includes

serves as

represents

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

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.

earthlife
CONSORTIUM

NEOTOMA: KEY CHALLENGES

NEOTOMA: KEY CHALLENGES

- **Reducing data friction** along the pipeline from collection to final archiving

NEOTOMA: KEY CHALLENGES

- **Reducing data friction** along the pipeline from collection to final archiving
    - Science-driven data-mobilization or data-rescue campaigns

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

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

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

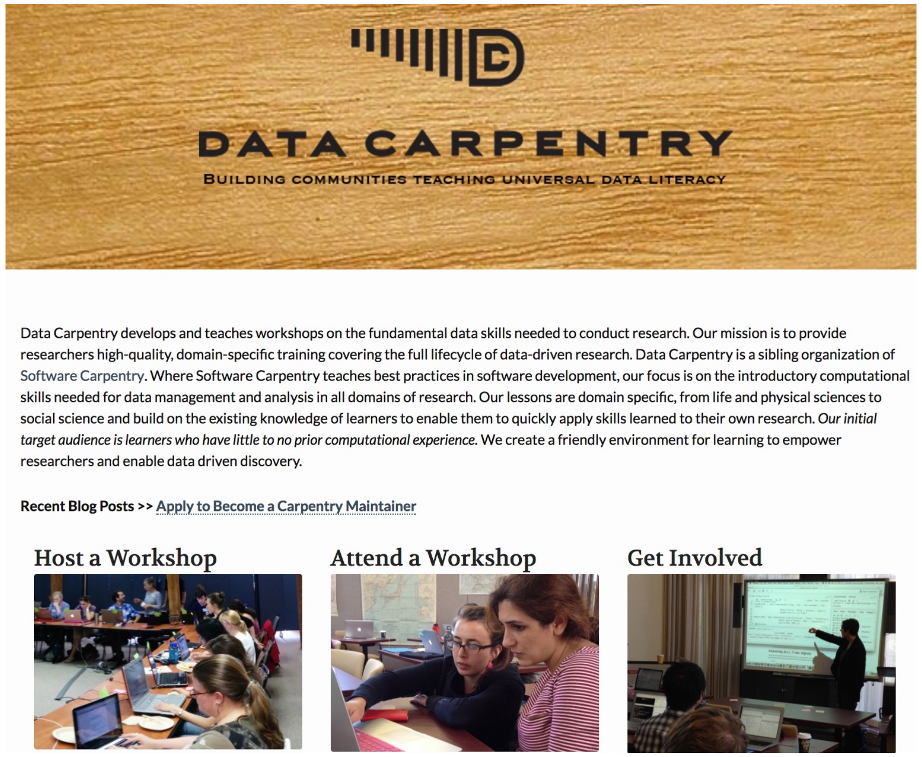
THE IDEAL!

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

Best Practices for Scientific Computing

Greg Wilson ✉, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

Published: January 7, 2014 • https://doi.org/10.1371/journal.pbio.1001745



DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

Data Carpentry develops and teaches workshops on the fundamental data skills needed to conduct research. Our mission is to provide researchers high-quality, domain-specific training covering the full lifecycle of data-driven research. Data Carpentry is a sibling organization of Software Carpentry. Where Software Carpentry teaches best practices in software development, our focus is on the introductory computational skills needed for data management and analysis in all domains of research. Our lessons are domain specific, from life and physical sciences to social science and build on the existing knowledge of learners to enable them to quickly apply skills learned to their own research. *Our initial target audience is learners who have little to no prior computational experience.* We create a friendly environment for learning to empower researchers and enable data driven discovery.

Recent Blog Posts >> Apply to Become a Carpentry Maintainer

Host a Workshop

Attend a Workshop

Get Involved

http://www.datacarpentry.org/

Our path to better science in less time using open data science tools

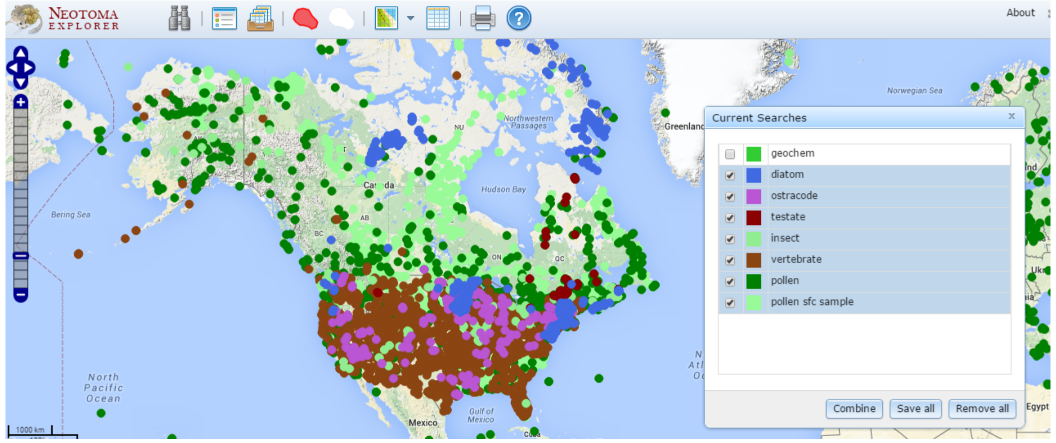
Julia S. Stewart Lowndes ✉, Benjamin D. Best, Courtney Scarborough, Jamie C. Afflerbach, Melanie R. Frazier, Casey C. O'Hara, Ning Jiang & Benjamin S. Halpern

Nature Ecology & Evolution **1**, Article number: 0160 (2017) doi:10.1038/s41559-017-0160

Received: 25 November 2016
Accepted: 07 April 2017
Published online: 23 May 2017

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/



Current Searches

□	🟩	geochem
☑	🟦	diatom
☑	🟪	ostracode
☑	🟥	testate
☑	🟩	insect
☑	🟫	vertebrate
☑	🟩	pollen
☑	🟩	pollen sfc sample

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

Become a member → More info: https://www.neotomadb.org/about/category/about

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

Become a member → More info: https://www.neotomadb.org/about/category/about

Use Neotoma Data → Explorer, APIs, R

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

Become a member → More info: https://www.neotomadb.org/about/category/about

Use Neotoma Data → Explorer, APIs, R

Cite Neotoma and Contributors

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

Become a member → More info: https://www.neotomadb.org/about/category/about

Use Neotoma Data → Explorer, APIs, R

Cite Neotoma and Contributors

Contribute data → More info: https://www.neotomadb.org/data/category/contribution

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

Become a member → More info: https://www.neotomadb.org/about/category/about

Use Neotoma Data → Explorer, APIs, R

Cite Neotoma and Contributors

Contribute data → More info: https://www.neotomadb.org/data/category/contribution

Become a Data Steward → Schedule a WebEx training session: neotoma-contact@googlegroups.com

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

Become a member → More info: https://www.neotomadb.org/about/category/about

Use Neotoma Data → Explorer, APIs, R

Cite Neotoma and Contributors

Contribute data → More info: https://www.neotomadb.org/data/category/contribution

Become a Data Steward → Schedule a WebEx training session: neotoma-contact@googlegroups.com

Launch a Constituent Database → Convene a working group

NEOTOMA: JOIN THE COMMUNITY

- Multiple ways to join the Neotoma community

Become a member → More info: https://www.neotomadb.org/about/category/about

Use Neotoma Data → Explorer, APIs, R

Cite Neotoma and Contributors

Contribute data → More info: https://www.neotomadb.org/data/category/contribution

Become a Data Steward → Schedule a WebEx training session: neotoma-contact@googlegroups.com

Launch a Constituent Database → Convene a working group

Spread the Word → Talk to your colleagues

NEOTOMA: EDUCATION

- Teaching Resources
  - SERC Carleton
    - http://serc.carleton.edu/neotoma/activities.html
  - Neotoma Webpage
    - https://www.neotomadb.org/education/category/higher_ed/