Neotoma Data Sustainability & Management Plan

By: Simon Goring, Chair, Neotoma IT/Development Committee
Reviewed by: Neotoma Executive Committee
Originally drafted and reviewed: May 12, 2016
Last Modified: September 21, 2016

Preamble

The Neotoma Paleoecology Database (hereafter, 'Neotoma') recognizes that data sustainability is a critical component of any reproducible scientific workflow. Sustainability is particularly important for Neotoma because the organization is a community-curated data repository that serves as a long term data archive for data generated by the paleoecological and paleoenvironmental communities. Neotoma has established this Data Management & Sustainability (DMS) plan to describe how we archive data and maximize sustainability. The intended audience for this DMS plan includes Neotoma data contributors, funding agencies, and journals using Neotoma as an archive.

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.

Implementation

Neotoma operates at multiple levels of data redundancy.

- Individual researchers are expected to retain their original data files, in their original data format (XLS, CSV, TLX, and others).
- Data Stewards vet data submissions, ensure all required metadata fields are filled and upload the data to Neotoma. Data Stewards are domain experts and can often detect underlying issues with the data that automated checks cannot detect.
- During the process of upload, Tilia files are created, in XML format. Neotoma retains these original upload files, separately.
- Data uploaded to the Neotoma relational database are housed at the Center for Environmental Informatics at Penn State and are protected by data backup and archiving policy that facilitates swift backup and long-term archiving. Data are protected by multiple measures, including redundant disk storage, off-site mirroring, file-system snapshotting, regular tape backup, and duplication of the backup set. CEI is committed to continuing to host the Neotoma database on a long-term basis even if there are temporary gaps in funding. CEI is supported by multiple diverse funding sources and

partnerships that have enabled the development of a robust computing infrastructure. This infrastructure provides an economy of scale that CEI can leverage to preserve hosting through short-term funding gaps.

- All data in the Neotoma database are made publicly accessible through multiple portals, including APIs, Neotoma Explorer, and the *neotoma* R package. .
- Neotoma provides persistent data object identifiers (DOIs) for all Neotoma datasets and separate landing pages [to be implemented shortly], providing a layer of metadata redundancy and a permanent identifier for individual datasets.
- To further ensure long term sustainability, the Neotoma Database is replicated using quarterly snapshots (every three months) and stored on the fig**share** platform from the *Neotoma Paleoecological Database* account (e.g., the snapshot for October 2015: https://dx.doi.org/10.6084/m9.figshare.3376393.v1). fig**share** ensures long term data sustainability through its partnership with the Digital Preservation Network who obtain periodic snapshots of the entire fig**share** collection (including the Neotoma snapshots), and replicate it across at least two Replicating Nodes, including the Academic Preservation Trust (University of Virginia), DuraCloud Vault (University of California at San Diego & DuraSpace), Stanford Digital Repository, Texas Preservation Node and HathiTrust (University of Michigan).
- Neotoma also plans to send these quarterly snapshots to the Paleoclimatology branch of the NOAA National Center for Environmental Informatics [to be implemented shortly]

Adapting to Evolving Technologies

A key element to long term sustainability is ensuring that the technologies used within the repository are current and well supported. Adaptation and evolution of implemented technologies to current best practices and standards ensures that Neotoma will remain relevant and will be able to partner and maintain sustaining relationships with the geosciences community and beyond. To ensure that Neotoma is able to stay abreast of best practices we commit to releasing a biennial “State of the Database” report as part of the IT Committee, which will evaluate existing technologies within the database ecosystem, and evaluate the potential for adaptation to new technologies or systems.

Backwards Compatibility

Any changes that result in the loss of backwards compatibility for database elements will be accompanied by notification prior to such a change. Prior to upgrades that break backwards compatibility a complete backup of the database system will be undertaken and uploaded to a secondary repository such as fig**share**. Versions for database system elements such as the API will be clearly indicated, and in cases where these secondary elements interact with the core data holdings, the earlier versions will continue to be maintained, but will be clearly identified as deprecated elements.

Change In Support

This section describes plans for Neotoma should the database need to change hosting institutions. This is a contingency plan only; as described above, CEI is committed to hosting Neotoma on a long-term basis.
In the event that Neotoma changes hosting institutions, three key components would need to be moved - the database, the API and the interface.

Use Cases

- A researcher contributes data to Neotoma but loses raw data as a result of fire, server crash, or other lab-scale data loss.
    - The data for sample counts, taxa, chronological data, site information and dataset-specific information can be recovered from Neotoma.
- A researcher requests data from a project that has published in the scientific literature.
    - The data can be obtained directly from Neotoma. Neotoma also includes data contributor contact information, and this provides a secondary pathway for data acquisition
- A journal requires access to data through a “live” data link in perpetuity
    - Neotoma’s partnership with the University of Wisconsin Library System and DataCite allow us to provide Digital Object Identifiers, which can serve as a link from any location on the web directly to a pertinent dataset. DOIs can be revised if necessary, and can be used to support data versioning using DataCite’s metadata fields.
- Neotoma funding ceases.
    - The Penn State CEI will continue to support Neotoma’s web services for some time in the event of funding gaps.
    - The database snapshots on fig**share** will be available from fig**share** as long as the organization exists
    - DataCite metadata persists indefinitely in the absence of the database.
    - Should fig**share** cease operations, snapshots can become available through the Digital Preservation Network’s snapshots of fig**share**.
- Penn State CEI ceases to support the Neotoma Paleoecological Database as a host institution.
    - The key components are moved to a new Windows Server
    - The Database is denormalized at the level of the dataset and transferred to the Penn State Scholarsphere (https://scholarsphere.psu.edu/).

Procedure for Modifying the Data & Sustainability Plan

Neotoma's Data Sustainability policy is a living document that represents current best practices. These practices and this document will be updated as best practices change. The IT/Development Working Group is responsible for maintaining and updating this document, and for developing and implementing technical solutions to data management and sustainability. The Neotoma Leadership Council is responsible for setting policies to maximize sustainability.