Ingest Workflow and Quality Control for General Contributions to the World Data Service for Paleoclimatology (WDS-Paleo)

Contents

Changelog

2024-01-22: Edit for clarity and consistency; convert document to pdf.
2023-01-10: Update URLs.
2019-04-22: Initial creation of this document.

I. Purpose

The purpose of this file is to document the data ingest workflow and quality control for general contributions to the World Data Service for Paleoclimatology (WDS-Paleo). General contributions are all contributions except for those to the International Tree Ring Data Bank (ITRDB) and International Multiproxy Paleo-fire Database (IMPD), which both require community-specific data formats.

For ingest workflow and quality control for the ITRDB, visit:
https://www.ncei.noaa.gov/pub/data/paleo/data_management/ITRDBworkflow.pdf

For ingest workflow and quality control for the IMPD, visit:
https://www.ncei.noaa.gov/pub/data/paleo/data_management/impd-tree-based-workflow.pdf and https://www.ncei.noaa.gov/pub/data/paleo/data_management/impd-charcoal-based-workflow.pdf.

II. Data Contribution Files

NOAA WDS-Paleo provides a template for the uploading, providing and storage of data in a standardized format. For detailed contribution guidance, visit: https://www.ncei.noaa.gov/products/paleoclimatology/contributing-data.

Data contributions can be provided in either the NOAA WDS-Paleo Template (which is a UTF-8 text-based document) text template, which is the format used for data archiving, or by using the Excel contribution template provided by the NOAA WDS-Paleo. The Excel template is converted to the NOAA WDS-Paleo Template (UTF-8 text) document by the NOAA Paleo data managers through the use of an Excel macro. Documentation of the templating process is provided inside the template, as well as through templates filled out with examples to help guide contributors.

New data acquisitions arrive typically by email, but also possibly on media or via drop box download.

III. Quality Control, Curation, and Ingest

Data contributions are inspected by data managers at the start of the ingest process. The data manager will work iteratively with the investigator to work through any issues in the formatting or description of the data set. The text template is run through an internal QC checker that looks for missing or incorrectly filled out metadata, as well as confirming that all terms used to describe the variables in the template are valid terms in the Paleoenviromental Standard Terms (PaST) thesaurus. The data manager and the original data contributor will also do a final review of the data set upon creation of a public facing landing page for the data contribution.

Climate modeling data contributions, as well as some gridded and ensemble data, use the NOAA template format to archive the metadata about their study. However, the multi-dimensional nature of these data sets dictates that they are best archived in the netCDF format from Unidata (https://www.unidata.ucar.edu/software/netcdf), so the data is archived as netCDF files with pointers to these files stored in the NOAA text template. The metadata template for the modeling data is run through our QC checker to ensure that it is filled out correctly and contains valid information.

After the template has been validated, the metadata is entered into the NOAA WDS-Paleo Oracle database via the PINGMAN data entry application, or via batch ingest. Using an ingest tool that connects directly to the database allows for the ingest process to be guided by dynamically generated lists of valid options and through the use of database constraints to check incoming entries.

Lastly, the text template is put into a NOAA/NCEI FTP directory for access by external users.

IV. Identifiers, Access, and Archive

Each dataset ingested into the NOAA WDS-Paleo database receives a unique internal study identifier for tracking purposes, a DOI for permanent data location, and three metadata records in ISO, DIF, and JSON formats. The ISO-19139 record is quality controlled using NCEI's rubric for automated metadata checking.

This ingest process results in the data becoming accessible to the end user through the WDS-Paleo Dataset Search: https://www.ncei.noaa.gov/paleo-search/, as well as other Data Access tools located at product-specific pages listed at: https://www.ncei.noaa.gov/products/paleoclimatology

On a monthly basis, all WDS-Paleo data are placed in the NCEI long-term archive.