NOAA World Data Service for Paleoclimatology (WDS-Paleo) Data Stewardship Workflow

Contents

Changelog

2024-01-22 Creation of document

I. Purpose

The purpose of this document is to provide a public-facing, high-level overview of World Data Service for Paleoclimatology (WDS-Paleo) data stewardship operations including data appraisal, deposit, quality assurance, ingest, dataset identification, access, archival storage, and ongoing curation and workflow management.

II. Data Deposit and Appraisal

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

Data contributions are appraised against the NCEI Archive policy and in accordance with the NOAA Procedure for Scientific Records Appraisal and Archive Approval, which define submission

guidance for data providers, specify the type of data that will be acquired for long-term preservation, and state limitations that may affect acceptance of environmental data at NCEI.

Appraisal includes assessment of the completeness and understandability of data and metadata, as well as compliance with required data/metadata formats. Detailed contribution guidance for all paleo data types is available at: https://www.ncei.noaa.gov/products/paleoclimatology/contributing-data. Internal quality-control software, peer-reviewed journal manuscripts associated with the contribution, and visual inspection are used to assess data and metadata quality.

If data are contributed in non-standard formats, the contributor is requested to resubmit in standard format, however non-standard formats can be contributed as supplemental information.

III. Quality Assurance and Ingest

Ingest, quality control, and archival storage of all WDS-Paleo data types employ the NOAA WDS-Paleo Template defined at: https://www.ncei.noaa.gov/pub/data/paleo/templates/noaa-wds-paleo-template-instructions.txt. The NOAA WDS-Paleo Template describes metadata and data, and it is validated by the data manager as part of the appraisal and ingest processes.

Most contributions follow this detailed guidance for ingest workflow and quality control: https://www.ncei.noaa.gov/pub/data/paleo/data_management/Paleo-ingest-general-workflow.pdf

Data contributions to the International Tree-ring Data Bank (ITRDB) and International Mulit-proxy Paleo-fire Database (IMPD) require community-specific ingest and quality-control procedures. The IMPD includes two types of fire history data, based on tree-ring fire scars and sediment charcoal. For the ITRDB and both IMPD data types, the NOAA WDS-Paleo Template is required as an archival data format. The ITRDB and the IMPD tree-ring-based data also require community-specific standard formats, the Tucson decadal and Fire History Exchange (FHX) formats, respectively, for archival.

Detailed guidance for ingest and quality control of ITRDB data is available at: https://www.ncei.noaa.gov/pub/data/paleo/data_management/ITRDBworkflow.pdf

Detailed guidance for ingest and quality control of IMPD tree-ring fire scar data is available at: https://www.ncei.noaa.gov/pub/data/paleo/data_management/impd-tree-based-workflow.pdf

Detailed guidance for ingest and quality control of IMPD sediment charcoal data is available at: https://www.ncei.noaa.gov/pub/data/paleo/data_management/impd-charcoal-based-workflow.pdf

Data contributions are inspected by data managers prior to the start of the ingest process, and curated throughout the metadata/data long-term life-cycle. The data manager will work iteratively with the investigator to work through any issues in the formatting or description of the data set.

The NOAA WDS-Paleo Template is validated by an internal QC checker that identifies 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 Paleoenvironmental Standard Terms (PaST) thesaurus. The data manager and the original data contributor also complete a final review of the data set upon creation of a public-facing landing page for the data contribution.

After archival NOAA WDS-Paleo Templates have been created and validated, the Data Manager ingests the metadata into the WDS-Paleo Oracle database using either the internally-developed Paleo INGest MANager application (PINGMAN) or programmatically via SQL batch scripts. The execution status of the PINGMAN workflows are trackable via output to a metadata document in NASA DIF format representing the state of the study being ingested, and is accessible to the Data Manager at any time. Exception handling for PINGMAN workflows and programmatic workflows are facilitated via database constraints and integrity checks, rollback of the affected data to a known good state, and records of errors viewable by the Data Manager. The direct connection of PINGMAN to the WDS-Paleo Oracle database allows the ingest process to be guided by dynamically-generated lists of valid options and by the use of database constraints to check incoming entries. After metadata ingest, standard metadata record types (ISO 19115-2, NASA DIF, Dublin Core, JSON) are created from the WDS-Paleo Oracle database for each study.

For the final steps in the ingest process, an ISO-19139 record is generated for each study from the NOAA WDS-Paleo database. This ISO record is then quality controlled using NCEI's rubric for automated metadata checking. The NOAA WDS-Paleo Template is uploaded into a NOAA NCEI FTP directory for access by external users.

IV. Data Access

Once any embargo period (e.g., for publication of a journal article) is passed, the ingest process results in the data becoming accessible to the end user through the NOAA WDS-Paleo search API and graphical web-based capabilities (https://www.ncei.noaa.gov/paleo-search), and through mapping tools (e.g., https://www.ncei.noaa.gov/maps/paleo/?layers=1) available via individual data type links in the Products Section of the NOAA/WDS-Paleo home page (https://www.ncei.noaa.gov/products/paleoclimatology).

The usage and performance of access to the WDS-Paleo holdings is tracked through google analytics and database processes.

V. Dataset Persistent Identification

Upon ingest, datasets stewarded by NOAA WDS-Paleo receive the following persistent identifiers: (1) a permanent web address (URL), (2) a UUID, and (3) in the vast majority of cases, a dataset DOI. In this workflow, as per NOAA policy, datasets which have already been assigned a DOI by another organization are ineligible to receive another from NOAA.

VI. Archival Storage

On a monthly basis, all WDS-Paleo metadata and data are placed in the NCEI long-term archive. This process workflow is detailed in the NOAA WDS-Paleo Archive Submission Agreement, which contains NCEI IT protocols that have been modified to meet NOAA WDS-Paleo feedback about paleo data needs and community requirements.

Data are transferred to NCEI's Archive Branch and archived following NCEI's Archiving Workflow. As per the Archive Workflow, the Archive Branch follows the OAIS-RM standard as described by Rank (2007): https://ieeexplore.ieee.org/abstract/document/4423732.

During this workflow, technical decisions are made and executed regarding the limits of archive file sizes, aggregating files, and naming the files as per the Archive Submission agreement. Execution tracking and exception handling are performed through shell scripts that monitor integrity and report exceptions to data managers via automated email messages.

VII. Ongoing Curation

a. **Updating or correcting datasets:** Updates or corrections to metadata or data are identified either by the WDS-Paleo, the dataset contributor, or a user. Contributors and users share issues with the WDS-Paleo via email to paleo@noaa.gov. For issues identified by a user, we verify the proposed change with one of the dataset contributors. In cases where we are unable to verify the proposed change with any of the original contributors, we note the possible metadata or data quality issue (see section VIIb). Changes to metadata or data are applied directly to the impacted file(s) that are available to users. The NOAA WDS-Paleo Template file's date last modified is updated, and the rationale of the change is added both to the metadata of the template file and to the WDS-Paleo Oracle database. A copy of the original file is archived and available to data managers offline. All WDS-Paleo metadata records (currently including ISO, DIF, and JSON) are automatically exported from the database on a daily basis, facilitating updates to metadata records.

b. **Noting metadata or data quality issues:** When a metadata or data quality issue is brought to the attention of the WDS-Paleo but is unable to be verified with the dataset

contributor, the issue is documented in the dataset's notes field in both the NOAA WDS-Paleo Template and the NOAA WDS-Paleo Oracle database, from which all all dataset landing pages and metadata records (ISO, DIF, JSON) are automatically generated on a daily basis.

c. **Versioning datasets:** Different versions of a dataset are stored within an umbrella web-accessible folder for the dataset, in separate sub-folders labeled accordingly for each version. A description of major changes between versions is documented in the NOAA WDS-Paleo Template metadata and in the NOAA WDS-Paleo Oracle database, from which all dataset landing pages and metadata records (ISO, DIF, JSON) are automatically generated on a daily basis.

d. **Changing metadata and data standards:** Deprecation or replacement of metadata and data standards is based on consideration of factors such as reusability and interoperability. If any of the metadata or data formats in use by the WDS-Paleo are deprecated or replaced by later standards, the WDS-Paleo will do its best to transform these to modern replacements, but will still keep the original data available. Transformation of metadata or data into new formats is facilitated by the machine-readable/interoperable formats currently in use by the WDS-Paleo. Metadata for WDS-Paleo digital objects are periodically reappraised as new pieces of information become required. This information is then added to the NOAA WDS-Paleo Oracle database, within which complete metadata for every WDS-Paleo dataset are stored and quality controlled, and from which the information can be exported to enhanced metadata formats. As metadata and data files are transformed into new formats, essential properties of the versions are compared programmatically. New, replaced, and deprecated metadata and data formats are communicated to the community via the Paleoclimate Discussion List moderated by the WDS-Paleo (https://www.ncei.noaa.gov/products/paleoclimatology/discussion-list).

e. **Removing datasets:** Metadata about archived datasets are preserved for the long-term, even if the data referenced by the metadata are removed. In this case, a tombstone page linked to the persistent identifier (DOI or landing page) of the dataset will be created, informing potential users about the deletion.

VIII. Ongoing Workflow Management

NOAA WDS-Paleo has policies and procedures in place to cover the data stewardship lifecycle from the data appraisal phase to digital preservation. Workflows are used to manage all activities of data stewardship processes in order to reduce inefficiencies and to meet goals.

a. **Workflow Documentation:** For data ingest workflows, actions are codified based on the level of curation necessary. Public-facing workflow documents provide a high-level overview and/or product governance information for the user community. Workflow documents are available at: https://www.ncei.noaa.gov/pub/data/paleo/data_management/. Internal workflow documents not available to the public contain complete process information in the form of flowcharts, methods, and technology-specific details used by the WDS-Paleo data managers and the NCEI IT Department. The majority of these internal documents are maintained on a WDS-Paleo internal website.

b. **Workflow Monitoring:** Workflows for metadata and data quality assurance are carried out, monitored, and managed through automated repetitive tasks using the WDS-Paleo Oracle Database, scripting, and reporting of outcomes on a regular basis to data managers. These outcome reports are analyzed with respect to process completion and efficiency, and how enhancements would be beneficial.

c. **Change Management:** NOAA/WDS-Paleo's monitoring and management of the need for technical change ensures that metadata and data formats, metadata schemas, and persistent identifiers evolve with changes within the technical environment as well as the needs of the user community. New workflows or enhancements to existing workflows are designed, developed, and documented by NOAA WDS-Paleo data managers and scientists in ongoing collaboration with NOAA IT experts and the user community. The change process involves discussion (with careful consideration to decision points during the definition and documentation phase), consensus, and implementation. Additional oversight of tree ring data workflows is provided by the ITRDB Advisory Committee and additional oversight of fire history data workflows is provided by the IMPD Advisory Board. Each workflow document contains a changelog table, which documents changes to the workflow over time. Changes to database and programmatic quality-assurance workflow processes are tracked via the NCEI JIRA issue tracking system. Changes to the data contribution workflow are also reflected in the Contributing Data web page at: https://www.ncei.noaa.gov/products/paleoclimatology/contributing-data