Paleoceanography and Paleoclimatology

AGU ADVANCING EARTH AND SPACE SCIENCE

RESEARCH ARTICLE

10.1029/2020PA004193

Key Points:

- The new Paleoenvironmental Standard Terms (PaST) thesaurus standardizes names of measured and inferred variables that are relevant to paleoenvironment and paleoclimate
- Paleoenvironmental Standard Terms (PaST) is now operational at World Data Service for Paleoclimatology, enabling more effective data searches and retrievals
- Semantic interoperability is supported by cross-linkages to other vocabularies and by Simple Knowledge Organization System formatting

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Morrill,
carrie.morrill@colorado.edu

Citation:

Morrill, C., Thrasher, B., Lockshin, S. N., Gille, E. P., McNeill, S., Shepherd, E., et al. (2021). The paleoenvironmental standard terms (PaST) thesaurus: Standardizing heterogeneous variables in paleoscience. *Paleoceanography and Paleoclimatology*, *36*, e2020PA004193. https://doi.org/10.1029/2020PA004193

Received 14 DEC 2020
Accepted 28 MAY 2021

The Paleoenvironmental Standard Terms (PaST) Thesaurus: Standardizing Heterogeneous Variables in Paleoscience

Carrie Morrill[1,2], Bridget Thrasher[1,2,5], Samuel N. Lockshin[1,2], Edward P. Gille[1,2], Shelley McNeill[3,4], Ethan Shepherd[3,4], Wendy S. Gross[2], and Bruce A. Bauer[2]

[1]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, US, [2]NOAA's National Centers for Environmental Information, Boulder, CO, US, [3]Riverside Technology, Inc., Fort Collins, CO, US, [4]NOAA's National Centers for Environmental Information, Asheville, NC, US, [5]Now at: National Center for Atmospheric Research, Boulder, CO, US

Abstract Paleoscience data are extremely heterogeneous; hundreds of different types of measurements and reconstructions are routinely made by scientists on a variety of types of physical samples. This heterogeneity is one of the biggest barriers to finding paleoclimatic records, to building large-scale data products, and to the use of paleoscience data beyond the community of specialists. Here, we document the Paleoenvironmental Standard Terms (PaST) thesaurus, the first authoritative vocabulary of standardized variable names for paleoclimatic and paleoenvironmental data developed in a formal knowledge organization structure. This structure is designed to improve data set discovery, support automated processing of data, and provide connectivity to other vocabularies. PaST is now used operationally at the World Data Service for Paleoclimatology (WDS-Paleo), one of the largest repositories of paleoscience information. Terms from the PaST thesaurus standardize a broad array of paleoenvironmental and paleoclimatic measured and inferred variables, providing enough detail for accurate and precise data discovery and thereby promoting data reuse. We describe the main design decisions and features of the thesaurus, the governance structure for ongoing maintenance, and WDS-Paleo services that now employ PaST. These services include an advanced search by variable name, an interface for thesaurus navigation, and a machine-readable representation in the Simple Knowledge Organization System (SKOS) standard. This overview is designed for developers of thesauri, data contributors, and users of the WDS-Paleo, and serves as a building block for future efforts within the broader paleoscience community to improve how data are described for long-term findability, accessibility, interoperability, and reusability.

1. Introduction

A classic example of the "long tail of science" (Heidorn, 2008), paleoclimate data consist of many small studies completed by individual investigators in separate labs using a multitude of methods and techniques. The heterogeneity of paleoscience data makes standardized metadata on variables (i.e., describing what was measured and how) essential in order to maximize reuse of the data (e.g., Wilkinson et al., 2016), and also creates a daunting task to completely and precisely describe potentially thousands of different types of measurements and reconstructions. Without standards for describing measured and inferred variables, the same measurements or methods are often given different names (e.g., sulfate, sulphate, SO_4, SO_4^{2-}) and details critical for interpretation are not archived (e.g., whether data are raw or have been transformed, the material on which measurements were made). Issues with the quantity and quality of metadata are not unique to paleoclimatology; scientists in many fields often unintentionally omit critical details about their methods and analysis in their published papers (Bowker, 2005) and are undertrained in sharing their data or workflows in standardized and structured ways (Borgman, 2012; Cragin et al., 2010).

Metadata standardization, including systematic variable definitions, has been identified by the paleoclimate community as a key improvement area (e.g., Diepenbroek et al., 2017; Emile-Geay & Eshleman, 2013; Jansma et al., 2011; Khider et al., 2019; McKay & Emile-Geay, 2016; Williams, Grimm, et al., 2018). Studies that synthesize previous results often require immense amounts of time, sorting through paleoclimate records manually and querying either the published literature or data contributors to gather basic metadata on

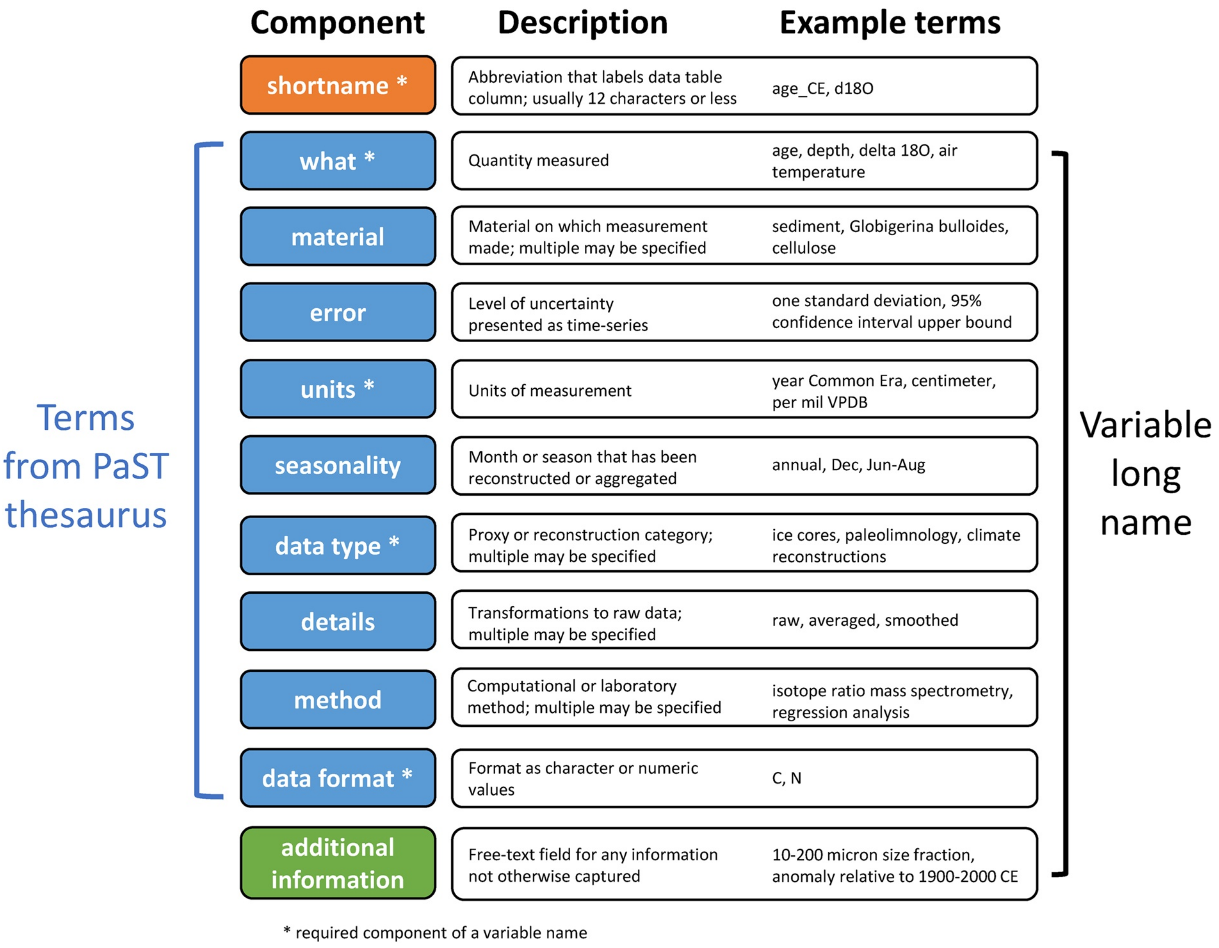
Table 1
WDS-Paleo Proxy and Reconstruction Data Types

WDS-Paleo data type	# Of studies as of April 1, 2021	% Of studies standardized with PaST thesaurus	Comments
Borehole	1,017	100	University of Michigan borehole database standardized to PaST
Climate forcing	64	100	
Climate reconstructions	608	100	
Corals and sclerosponges	237	100	
Fire History	829	100	International Multiproxy Paleofire Database standardized to PaST
Historical	31	100	
Ice Cores	347	100	
Insect	11	100	
Lake levels	22	100	
Loess	19	100	
Other collections	49	0	Not standardized, studies too heterogeneous
Paleoceanography	930	100	
Paleoclimatic modeling	38	100	
Paleolimnology	409	100	
Plant macrofossils	296	0	Not standardized to PaST, separate controlled vocabularies maintained by the Neotoma Paleoecology Database
Pollen	9	0	Not standardized to PaST, separate controlled vocabularies maintained by the Neotoma Paleoecology Database
Speleothems	229	100	
Tree rings	5,268	100	International Tree Ring Data Bank standardized to PaST

Abbreviations: PaST, Paleoenvironmental Standard Terms; WDS-Paleo, World Data Service for Paleoclimatology.

what was measured and how for each individual study (e.g., Jonkers et al., 2020). Yet, this work is essential since paleoclimatic reconstructions gain power when they combine many sites and different proxies sampling different components of the earth system (e.g., Kaufman et al., 2020; Marsicek et al., 2018; Neukom et al., 2019; Routson et al., 2019; Tardif et al., 2019; Tierney et al., 2020). The lack of systematic variable metadata is also a barrier to ensuring reproducibility of results. Without information about how a measurement was made, it becomes impossible to determine whether divergent results may be due simply to differences in measurement or definition. Additionally, there is a need to update records through time as new age and proxy calibrations become available, however this goal can be realized only if one knows adequately what was measured. Lastly, standardizing variable names contributes to building cyberinfrastructure within geosciences, allowing computer applications to accelerate data-intensive science (e.g., Williams, Kaufman, et al., 2018). A recent embodiment of this idea, the FAIR principles emphasize data set findability, accessibility, interoperability and reusability, all of which are promoted by controlled terminology (Wilkinson et al., 2016).

The need for standardized variable metadata is critical at the World Data Service for Paleoclimatology (WDS-Paleo), one of the largest and most heterogeneous repositories of paleoscience data. At the WDS-Paleo, early efforts in the late 1990s and early 2000s to develop databases for a few specific proxy types (e.g., tree rings, fire history) led to the establishment of some variable naming conventions in these sub-disciplines. In the decade of the 2000s, the WDS-Paleo combined separate proxy databases into one overarching database, with parallel metadata development that allowed searching across 18 proxy and reconstruction data types (Table 1). However, systematic nomenclature and metadata for measured and inferred variables was beyond the scope of the project at the time. In 2013, the WDS-Paleo developed a variable naming scheme that involves multiple components (e.g., what was measured, units, method; Figure 1) and began to use it within machine-readable data templates (Figure 2) for new contributions across all proxy types. This new approach



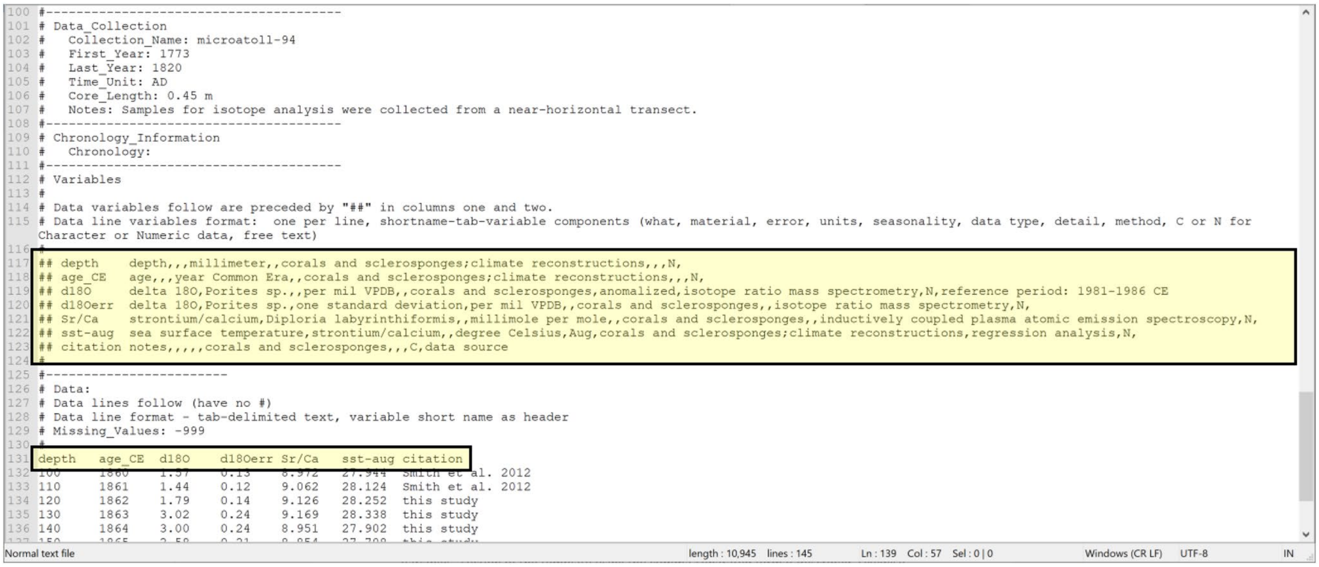
Component	Description	Example terms
shortname *	Abbreviation that labels data table column; usually 12 characters or less	age_CE, d18O
what *	Quantity measured	age, depth, delta 18O, air temperature
material	Material on which measurement made; multiple may be specified	sediment, Globigerina bulloides, cellulose
error	Level of uncertainty presented as time-series	one standard deviation, 95% confidence interval upper bound
units *	Units of measurement	year Common Era, centimeter, per mil VPDB
seasonality	Month or season that has been reconstructed or aggregated	annual, Dec, Jun-Aug
data type *	Proxy or reconstruction category; multiple may be specified	ice cores, paleolimnology, climate reconstructions
details	Transformations to raw data; multiple may be specified	raw, averaged, smoothed
method	Computational or laboratory method; multiple may be specified	isotope ratio mass spectrometry, regression analysis
data format *	Format as character or numeric values	C, N
additional information	Free-text field for any information not otherwise captured	10-200 micron size fraction, anomaly relative to 1900-2000 CE

Terms from PaST thesaurus
Variable long name
* required component of a variable name

Figure 1. Components of the World Data Service for Paleoclimatology (WDS-Paleo) naming scheme for measured or inferred variables. Components requiring terminology standardized by the Paleoenvironmental Standard Terms (PaST) thesaurus are shown in blue. Variable short names label columns of a data table, while the comma-separated components of the variable long name (as displayed in Figure 2) provide more detailed information about the variable.

separated essential metadata into categories, providing an organizational structure that is general enough to handle variable names across the many different proxy and reconstruction types archived, and is extensible enough to capture important proxy-specific metadata. However, the specific terms used within this structure were not standardized. This lack of standardized terms, in turn, limited variable-related data discovery at the WDS-Paleo to a free-text search and to general parameters based on the NASA Global Change Master Directory (GCMD) science keywords (e.g., geochemistry, physical properties, population abundance; Global Change Master Directory, 2020). These searches often fail to return relevant datasets to interested users. The free-text search, which searches several metadata text fields (e.g., study notes, study publication, and abstract), often yields results for variables that have low recall due to non-standardized terminology or incomplete metadata, or else have low precision due to variables that were not measured being included in metadata text fields (e.g., in an abstract). GCMD parameter keywords tend to be general and limit targeted searches for specific variables.

Here, we document the Paleoenvironmental Standard Terms (PaST) thesaurus, the first vocabulary of variable names for paleoclimatic and paleoenvironmental data developed in a formal knowledge organization structure (Zeng, 2008) in which terms are related and defined. This structure enhances data set discovery, automated processing of data, and connectivity to other vocabularies. We also describe an advanced search by variable name, an interface for thesaurus navigation that displays definitions for all terms and visualizations of their relationships, and a machine-readable representation of the thesaurus in the Simple Knowledge Organization System (SKOS) standard, all at the WDS-Paleo. The PaST thesaurus now standardizes



```
#---------------------------------------
# Data_Collection
#   Collection_Name: microatoll-94
#   First_Year: 1773
#   Last_Year: 1820
#   Time_Unit: AD
#   Core_Length: 0.45 m
#   Notes: Samples for isotope analysis were collected from a near-horizontal transect.
#---------------------------------------
# Chronology_Information
#   Chronology:
#---------------------------------------
# Variables
#
# Data variables follow are preceded by "##" in columns one and two.
# Data line variables format:  one per line, shortname-tab-variable components (what, material, error, units, seasonality, data type, detail, method, C or N for
Character or Numeric data, free text)
#
## depth    depth,,,millimeter,,corals and sclerosponges;climate reconstructions,,,N,
## age_CE   age,,,year Common Era,,corals and sclerosponges;climate reconstructions,,,N,
## d18O     delta 18O,Porites sp.,,per mil VPDB,,corals and sclerosponges,anomalized,isotope ratio mass spectrometry,N,reference period: 1981-1986 CE
## d18Oerr  delta 18O,Porites sp.,one standard deviation,per mil VPDB,,corals and sclerosponges,,isotope ratio mass spectrometry,N,
## Sr/Ca    strontium/calcium,Diploria labyrinthiformis,,millimole per mole,,corals and sclerosponges,,inductively coupled plasma atomic emission spectroscopy,N,
## sst-aug  sea surface temperature,strontium/calcium,,degree Celsius,Aug,corals and sclerosponges;climate reconstructions,regression analysis,N,
## citation notes,,,,,corals and sclerosponges,,C,data source
#
#-----------------------
# Data:
# Data lines follow (have no #)
# Data line format - tab-delimited text, variable short name as header
# Missing_Values: -999
#
depth   age_CE  d18O    d18Oerr Sr/Ca   sst-aug citation
100     1860    1.57    0.13    8.972   27.944  Smith et al. 2012
110     1861    1.44    0.12    9.062   28.124  Smith et al. 2012
120     1862    1.79    0.14    9.126   28.252  this study
130     1863    3.02    0.24    9.169   28.338  this study
140     1864    3.00    0.24    8.951   27.902  this study
150     1865    2.58    0.21    8.854   27.708  this study
```

Normal text file
length : 10,945   lines : 145
Ln : 139   Col : 57   Sel : 0 | 0
Windows (CR LF)   UTF-8
IN

Figure 2. Snippet of World Data Service for Paleoclimatology (WDS-Paleo) machine-readable data template, showing metadata for measured and inferred variables in yellow highlight. The upper yellow box shows short names followed by variable long names using the comma-separated components detailed in Figure 1, while the lower yellow box contains variable short names to label columns of the data table. Terms used in this example follow specifications of the Paleoenvironmental Standard Terms (PaST) thesaurus. While nearly all data sets at the WDS-Paleo are searchable using PaST (Table 1), only text templates created since March 2020 display PaST-compliant variable names.

variable names for 15 out of 18 proxy and reconstruction data types, including nearly 97% of existing data sets, archived at the WDS-Paleo (Table 1). New contributions to the WDS-Paleo are required to conform to standard naming as defined by PaST.

Development of the thesaurus followed five steps, which are discussed in more detail in the next five sections of this paper: (a) collect requirements for the vocabulary, (b) review existing paleodata resources and vocabularies, (c) select desired structure and format, (d) identify specific terms to use as well as their relationships, and (e) ensure conformance with relevant standards and community requirements. After documenting these aspects of the PaST thesaurus, we detail current access to the thesaurus and operational uses of the thesaurus within the WDS-Paleo web service, as well as the governance structure for undertaking future modifications.

This overview of PaST is intended to support several groups of readers. First, PaST enables a new advanced search of data sets by variable name at the WDS-Paleo. To support this new searching capability, contributors to the WDS-Paleo are now required to use PaST to describe variables in their data sets. Because PaST has more features than a simple controlled vocabulary, our discussion of PaST structure (Section 4), terminology (Section 5), and user interfaces (Section 7) is designed to help contributors and users select appropriate terminology for their data sets and searches. Users of the WDS-Paleo who construct data syntheses and process data sets in automated ways may benefit from the descriptions of thesaurus quality control (Section 6) and versioning (Section 8). Lastly, information about requirements (Section 2) and existing paleoscience resources and vocabularies (Section 3) are relevant to developers of thesauri and other vocabularies.

2. Vocabulary Requirements

We gathered requirements for PaST from users of the WDS-Paleo through several channels. These include feedback submitted to the WDS-Paleo by email, interactions with community members at national and international conferences and workshops, and advisory panels consisting of data contributors and users. These advisory panels also helped to guide vocabulary structure and terminology during development and will be discussed further in Section 5. These sources gave four main requirements for PaST:

1. Broad in scope: The collection of terms must be broad enough to capture the heterogeneity of measured and inferred variables in the WDS-Paleo collections.

2. Specific yet sustainable: Terms should be specific enough to capture details critical for discovery and for re-use. Rather than high-level keywords, terms should capture the specific quantity measured in most cases as well as important ancillary information (material on which measurements were made, units, data transformations, etc). However, it is important to avoid terms that are too specific and that are likely to be used only by a few data sets. In these cases, a more general term is sufficient.
3. Extensible: The paleoclimate community regularly develops new proxies and reconstruction targets, as well as new methods and analytical techniques. A governance process should be established to handle ongoing additions or modifications to terminology. The vocabulary structure should also be scalable and flexible for capturing new categories of metadata for measured and inferred variables, if desired by users in the future.
4. Improves data discovery: In addition to having a broad scope and appropriately specific terms, terms should be harmonized across data types to support cross-disciplinary searches and should be organized into a hierarchical structure to enable searches at different levels of specificity.

The multiple-part variable framework in use by the WDS-Paleo is the starting structure for the PaST controlled vocabulary, which provides terms to use within nine components of the variable long name (Figure 1). This variable framework is general enough for use with all of the eighteen different proxy and reconstruction data types archived by the WDS-Paleo (Table 1), thus allowing more uniformity to be applied to its holdings and allowing metadata to be stored and searched across data types in a single database structure.

3. Existing Paleoscience Data Resources and Vocabularies

As a second step in developing PaST, we reviewed other existing vocabularies and standards with relevance to measured and inferred variables in paleoscience. The purpose of this review was to avoid duplicating past efforts and also to assess other structures and approaches, particularly from the standpoint of maximizing interoperability between PaST and other resources. Controlled vocabularies can take several different forms (e.g., Harpring, 2010; Hedden, 2010). At their most basic, controlled vocabularies are lists of terms without any complex structure or relationships. Next in complexity, a taxonomy is a controlled vocabulary with a hierarchical structure that reflects parent/child (broader/narrower) relationships. Taxonomies are often visualized as a tree with terms as nodes. A thesaurus is a taxonomy with three sorts of relationships: hierarchical (broader/narrower), associative (related to), and equivalent (same as). Other sorts of metadata are also often included, including definitions, notes on term usage, and notes on term history. An ontology is a sort of taxonomy in which relationships can be more precisely specified and can extend beyond the three kinds represented in a thesaurus. As vocabulary forms become more sophisticated, they can support more complex queries and greater automation by analysis software.

Given the highly interdisciplinary nature of paleoscience, many existing vocabularies could be considered relevant to the WDS-Paleo. In this section, we summarize several that are primarily focused on paleoclimate, paleoecology, paleoceanography, or climate science. There are other domain-specific vocabularies that are useful to a subset of the WDS-Paleo holdings (Table 2). In Section 5, we further discuss these additional vocabularies and efforts to cross-walk PaST to them.

The WDS-Paleo tags data sets with science keywords from the NASA GCMD (Global Change Master Directory, 2020), and this information is used by the WDS-Paleo keyword search. The GCMD keywords are hierarchical and controlled (e.g., earth science > paleoclimate > ice core records > isotopes). Additional, uncontrolled levels provide for more complete description. For example, "isotopes" could have a sublevel titled "oxygen." While the GCMD keywords have long been useful to the WDS-Paleo, neither the structure nor the specificity meets current WDS-Paleo requirements to describe measured and inferred variables. The GCMD keywords lack discrete fields, for example, details such as units must be included in the keyword (e.g., isotopes > oxygen per mil PDB), and other metadata such as material used or data transformations are not easily recorded. Also, GCMD's scope covers all of Earth Sciences, and paleoclimate variables are described with about 100 keywords. Nonetheless, the GCMD keywords are useful at the data type level and offer an opportunity for cross-walking this component (Table 2).

The World Data Center PANGAEA archives data from earth and life sciences and has a large collection of paleo-relevant data sets with an emphasis on paleoceanography (https://pangaea.de; Diepenbroek

Table 2
Online Databases and Vocabularies Referenced in PaST Thesaurus

Name	Reference	# Mapping links
Chemical Entities of Biological Interest (ChEBI)	Degtyarenko et al. (2008); Hastings et al. (2016)	55
Chemical Methods Ontology (CHMO)	Batchelor (2020)	54
Climate Forecast (CF) Standard Names	Climate and Forecast Metadata (2020); Hankin et al. (2010)	98
Global Change Master Directory (GCMD)	Global Change Master Directory (2020)	27
Integrated Taxonomic Information System (ITIS)	Integrated Taxonomic Information System (2020)	33
MinDat	Hudson Institute of Mineralogy (2020)	85
PubChem	Kim et al. (2019)	176
Units of Measurement ontology (UO)	Gkoutos (2020); Gkoutos et al. (2012)	97
World Register of Marine Species (WoRMS)	Horton et al. (2017); Vandepitte et al. (2018); WoRMS Editorial Board (2020)	128

et al., 2002). To describe variables, PANGAEA employs a parameter naming scheme with several parts: variable name, short name, units, method/device, and comment (Diepenbroek et al., 2017). The variable name is complex, meaning that several different pieces of metadata are combined into a compound concept that is stored as a string with components separated by commas. These metadata components are similar to some of those used by the WDS-Paleo (Figure 1) and include what was measured, the material on which it was measured, the error, and details related to data transformations and the seasonality of a reconstruction. These similarities in the types and specificity of metadata between PANGAEA and WDS-Paleo offer a future opportunity for some degree of harmonization and interoperability, particularly for paleoceanographic data sets.

The Neotoma Paleoecology Database houses data sets related to paleoenvironments, including pollen, plant macrofossil, diatoms, ostracodes, and fossil mammal data (https://www.neotomadb.org). Neotoma variable names consist of a required taxon name and the optional properties of element, units, and context, all of which are populated with controlled terms (Williams, Grimm, et al., 2018). The taxon name can be an actual taxon or, more generally, what was measured (e.g., a geochemical quantity). The element is the organ or part of the organism that was identified in the case of a biological measurement, or a modifier in the case of a physical measurement. This usage is roughly equivalent to the WDS-Paleo "material" field (Figure 1). The context term refers to the depositional context that may influence the interpretation of the data values. The majority of data archived in the Neotoma database are taxonomic and generally do not overlap with WDS-Paleo data types, so there is currently little opportunity for re-use of controlled terms. However, the scope of the Neotoma database is expanding into data types routinely archived by the WDS-Paleo, raising the potential for future collaboration to establish interoperability in terminologies between the two databases.

The Climate and Forecast (CF) metadata conventions have been developed by the atmospheric science community, specifically to promote sharing of data in NetCDF format (Eaton et al., 2020). CF variables are defined by standard names and canonical units in a list that is tightly controlled and maintained by the community (Climate and Forecast Metadata, 2020). These variable names contain a level of detail sufficient for sharing and comparison of data sets across many different applications, perhaps most notably climate model intercomparisons in support of IPCC assessment reports (Hankin et al., 2010; Williams et al., 2009). Given the focus on modern climatic and oceanographic measurements, the CF standard names cannot describe the majority of WDS-Paleo data holdings, but they do offer guidance and cross-walking opportunities in terminology for quantities reconstructed from raw proxy measurements (Table 2).

Recently, the Linked Paleo Data (LiPD) framework has grown as a standard data format for paleoclimatic data, but LiPD does not include any controlled vocabularies (McKay & Emile-Geay, 2016). LiPD is part of a broader effort called the Linked Earth Ontology (LEO), which structures paleoclimate metadata and data using linked data concepts (Emile-Geay & Eshleman, 2013). The LEO structure, formalized as an ontology, is general enough to accommodate any sort of paleo data and includes several modules to describe variables. These modules include the Proxy Archive, Proxy Sensor, Proxy Observation, Inferred Variable, and Instrument. Each of these modules has a general counterpart within the WDS-Paleo variables framework

(Figure 1). Within each of these modules are additional properties such as units, calibration method, and climate interpretations. Developing a detailed controlled vocabulary for variables for use within the ontology structure was beyond the scope of LEO, and there are opportunities for the PaST vocabulary to be re-used for LEO.

To summarize, none of the existing paleo-relevant vocabularies meets all of the requirements for the WDS-Paleo. Differences in scope, either related to specificity or to subject matter, limit our ability to re-use some of the vocabularies. It is desirable, however, to re-use what we can from these other vocabularies and to develop a terminology that maximizes our ability to map structure and terms across systems. Our review points to several cases where this should be feasible. Adherence to published standards for vocabulary structure and sharing, as described more in the next section, is critical for promoting this interoperability.

4. Choices for PaST Structure: Thesaurus and SKOS

Of the vocabulary forms described in Section 3, the thesaurus best fits requirements for PaST. The hierarchical structure provides organization to ease navigation, context to provide meaning, and logic to allow searches at varying levels of specificity. Associative relationships can bridge the different components of a variable name (Figure 1). For example, these relationships can identify which materials are associated with a particular data type. Likewise, equivalence relationships are necessary for tracking synonyms and connections to other vocabularies, a key component of interoperability. These three relationships capture most of the connections necessary to meet our requirements, obviating the need for a more complex ontology.

In designing the structure for our thesaurus, we consulted two relevant standards: ISO-25964 (International Organization for Standardization, 2011, 2013) and SKOS (World Wide Web Consortium, 2009). These standards provide guidelines for naming concepts clearly, establishing relationships between concepts, and ensuring that logic is maintained. Compliance with these guidelines enables machine-to-machine communication (interoperability) and semantic reasoning. The ISO-25964 standard describes the development and maintenance of thesauri, and the management of interoperability among thesauri. SKOS, on the other hand, provides recommendations for sharing and linking thesauri or other simple knowledge systems on the Web. Interactions between the groups developing SKOS and ISO-25964 harmonized their individual recommendations (Dextre Clarke & Zeng, 2011).

Both ISO-25964 and SKOS provide data models; the ISO model is specific to thesauri and the SKOS model applies more generally to different sorts of knowledge systems. For the PaST thesaurus, we adopt the SKOS data model since it has become the de facto standard for representing various classification schemes on the Web and promotes interoperability among vocabularies. The SKOS data model is expressed as Resource Description Framework (RDF; World Wide Web Consortium, 2014) triples in the form of subject-predicate-object that can be understood by computers. Unlike more formal ontologies, SKOS is lightly specified using thesaurus-like relationships rather than logical semantic ones. This is by design in order to make SKOS broadly applicable to thesauri and controlled vocabularies, which are not always easily converted to formal logical semantics. SKOS is also designed to be compatible with ISO-25964 standards, although it does not meet all of the standards (Baker et al., 2013), as will be discussed further in Section 6.

SKOS (as well as ISO-25964) organizes vocabularies around concepts, which have labels and relationships. Figure 3 shows the SKOS data model for one concept, using the element carbon as an example from PaST. Concepts can have multiple labels, including the preferred term and the non-preferred term (prefLabel, altLabel). Examples of non-preferred terms are alternative spellings, abbreviations, and obsolete terminology. Concepts have semantic relations that are hierarchical (narrower, broader) or associative (related). Likewise, concepts can be linked across thesauri using mapping relations (exactMatch, closeMatch). Concepts can also have additional metadata to document their meaning or use cases (definition, scopeNote).

Each concept in PaST belongs to one of the top level components shown in Figure 1. Concepts are linked to these top level components via hierarchical relations. For example, carbon's broader term “element” has the broader term “chemical composition,” which in turn has the broader term “what.” Cross-component linkages are captured by associative relationships (related). The only cross-component relationships currently in PaST consist of concepts belonging to the “what” and “material” components linked to specific concepts

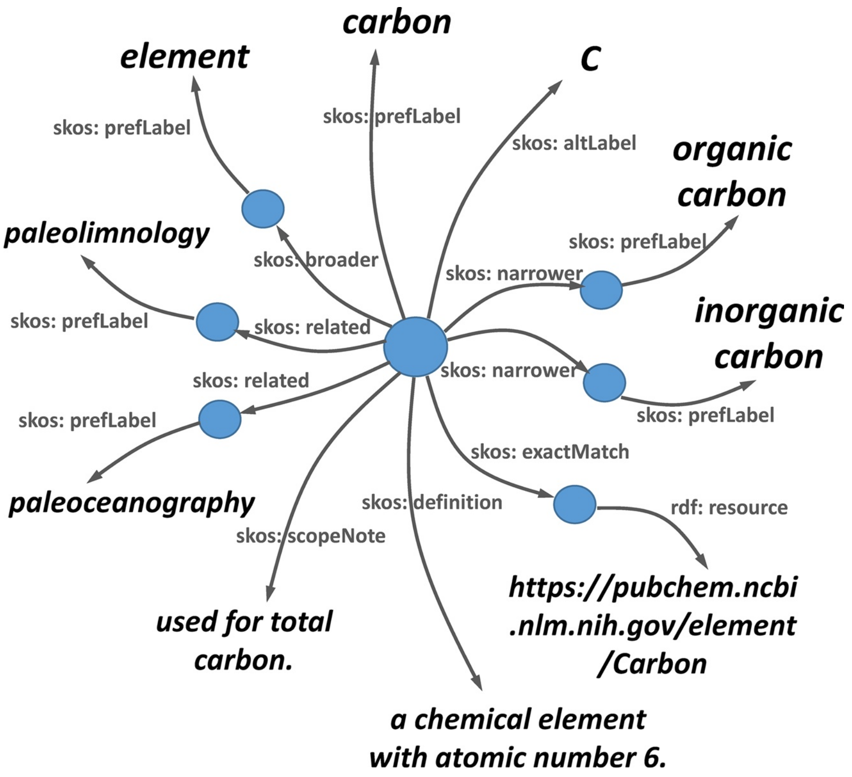


Figure 3. Simple Knowledge Organization System (SKOS) data model employed by the Paleoenvironmental Standard Terms (PaST) thesaurus using the concept "carbon" as an example. Concepts are represented by circles. SKOS and Resource Description Framework (RDF) relationships and tags are denoted with the skos and rdf prefixes, respectively. Labels, notes, and resources associated with concepts are shown in bold font.

within the "data type" component. For example, Figure 3 shows that "carbon" is related to the "paleoceanography" and "paleolimnology" data types. Since the list of relevant "what" and "material" concepts vary widely between data types, these associative relationships aid contributors in locating appropriate terms to describe their data. Other types of associative relationships, for example relating particular "units" or "methods" concepts to relevant "what" concepts, could be a future refinement to PaST.

5. Populating PaST: Concepts, Terminology, and Relationships

We selected concepts, labels, and relationships through examination of more than 8,600 data sets housed by the WDS-Paleo as of March 2018. About 6,500 of these data sets belong to one of three proxy-specific databases in which a known set of standard measurements are reported (i.e., tree ring, fire history, borehole; Table 1). To represent these measurements in the PaST thesaurus, we chose concept labels in line with vocabulary used by the science community and reported appropriate relationships. For the other approximately 2,100 data sets, we collected existing variable metadata from data set files and manually standardized names across studies. Through this process, we both identified concepts for the PaST thesaurus, populated preferred and non-preferred labels, and created standardized variable names for nearly all legacy data sets at the WDS-Paleo. One strength of this approach is that the WDS-Paleo houses three decades worth of paleoscience research and the resulting PaST vocabulary is reasonably comprehensive. Since the initial development and application of PaST, all newly contributed data sets also follow PaST terminology and nearly all WDS-Paleo data sets are now PaST-compliant (Table 1).

While building the thesaurus, we consulted and cross-referenced several existing databases and vocabularies (Table 2). These vocabularies include controlled lists, taxonomies, and ontologies whose scope overlaps with part of the PaST thesaurus. This cross-referencing served to quality control our selection of labels (e.g., choosing most up-to-date taxonomic names), usage of terms, and specification of hierarchical relations.

Through this process, we also established mapping relations between our concepts and those in other vocabularies (Table 2), contributing to the semantic web and providing ways for users of PaST to tap into the rich metadata, reference material, and imagery provided by these additional sources.

We next gathered feedback on draft versions of sections of the thesaurus from advisory panels composed of subject matter experts in the paleo research community. Invited panelists had all contributed to and used the WDS-Paleo archive. We convened six panels via videoconference, one for each of the major data types without standardized variable names (Table 1: corals and sclerosponges, climate reconstructions, ice cores, paleolimnology, paleoceanography, and speleothems). Each panel met twice for a total period of three hours, providing guidance and feedback that were incorporated into the final version of the thesaurus.

These advisory panels provided particularly important feedback for decisions related to the specificity of concepts. Our goal here was to reach a level of specificity that enables searches with a useful degree of discrimination and that provides the information necessary for reproducibility and reusability, while avoiding populating the thesaurus with concepts that will be used only rarely. As subject matter experts, the panelists advised us on which concepts were critical and which were not. For example, the methods component contains only high-level categories, and not details such as specific models of laboratory instruments. The thesaurus provides general concepts (e.g., mineral index, grain size class) to be used instead of customized measures without wide applicability (e.g., ratios of two minerals, non-standard grain size classes). Recognizing that in some cases, a general concept may be unsatisfactory for explaining a measured variable fully, we have adopted an additional free-text field in which explanatory information may be added about any of the nine controlled fields in a variable name ("additional information" in Figure 1).

Based on guidance from the panels, the thesaurus includes limited taxonomies, with genus and species names for corals, coralline algae, and selected foraminifera only. The panel recommended particular foraminiferal species for inclusion based primarily on their usage for geochemical, rather than assemblage, measurements. Species that do not have their own PaST entry are assigned a less specific "what" term (e.g., "identified foraminifer" or "identified diatom") with the genus or species name provided in the "additional information" field. This approach allows for important materials metadata to be captured for geochemical measurements, while recognizing that tracking complex and evolving taxonomies is best done by a team of subject matter experts. Given the general nature of taxonomic tracking in PaST, species terms are assigned based on a name match rather than more specific taxonomic definitions or references. Lastly, tree species taxonomies of the International Tree Ring Data Bank (ITRDB) predate the development of PaST and have their own metadata structure and specialized searching capability in the WDS-Paleo advanced data search (https://www.ncdc.noaa.gov/paleo-search).

We built hierarchical relationships between concepts using primarily an inductive (bottom-up) approach from the lists of concepts identified from WDS-Paleo studies. The resulting relationships are generally either generic or instantive, that is, they follow the "all-some" or "is a" test in which all instances of a narrower concept are members of the broader concept but only some instances of a broader concept are members of the narrower concept (e.g., Mazzocchi et al., 2007). In the generic case, both broader and narrower terms are concepts (e.g., a hydrocarbon is an organic compound), while in the instantive case the narrower term is an instance of the broader term (e.g., carbon is an element). However, some hierarchical relationships in PaST are partitive (e.g., sediment is a part of bulk geological material). While all of these types of relationships are classified as hierarchical in the PaST thesaurus, an ontology would be able to more clearly differentiate them. In our case, we prioritize the "local practical" goal of making an efficient local system with improved searching capabilities over the "big picture logical" goal of deviating as little as possible from the most logical relationships (Mazzocchi et al., 2007). If these goals shift, an expansion of PaST to include more ontological relationships is feasible, particularly because both use the same underlying structure of RDF triples.

Regarding the grammatical form of terms, we follow recommendations from ISO-25964-1 when possible. These include using nouns and noun phrases, minimizing non-alphabetic characters, maintaining a consistent capitalization style, expressing count nouns as plural and non-count nouns as singular, adopting the most widely used spelling for preferred terms, and generally avoiding abbreviations and acronyms.

Table 3
Statistics for Past Thesaurus Version 1 by Variable Name Component

Component	# Concepts	# Non-metaterms (assignable terms)	# Mapping relations
what	1,185	1,102	397
material	333	314	178
error	67	61	0
units	370	305	97
seasonality	166	153	0
data type	20	19	27
detail	16	15	0
method	122	114	54
data format	3	2	0
TOTAL	2282	2085	753

Version 1.0 of the PaST thesaurus (https://www.ncdc.noaa.gov/paleo/study/31892) contains 2,282 concepts. Each has a preferred label, and there are also 733 non-preferred labels (altLabel). These non-preferred terms are not used to describe variables in data sets, but are maintained in PaST to aid users in locating preferred terms (see Section 7). Non-preferred terms come from legacy data sets; we did not seek out additional synonyms, under the assumption that the large collection of legacy files at the WDS-Paleo (Table 1) provide a reasonably comprehensive collection of terms commonly used by the community. Of the 2,282 concepts, 197 are metaterms (i.e., they are broader terms that are not used to describe a measured variable, but instead have an organizational purpose) and the remaining 2,085 non-metaterms are assignable in variable names. There are 753 mapping relations (Table 2), which cross-reference a PaST concept to a concept or term in another vocabulary. For example, the PaST concept "carbon" is mapped to the same concept in PubChem, an open chemistry database at the National Institutes of Health (Figure 3; Table 2). Thus, nearly a third of non-metaterm concepts have a mapping relation. Counts of concepts, metaterms, and mapping relations for each of the nine controlled components in WDS-Paleo variable names (Figure 1) are provided in Table 3.

We built the PaST thesaurus through the process of standardizing variable metadata in nearly all studies archived at the WDS-Paleo (Table 1). When complete, both the thesaurus terms and the standardized file-level metadata for measured and inferred variables were ingested into the relational database that houses all WDS-Paleo metadata for data discovery. Variable metadata are outputted from the database to Directory Interchange Format (DIF; Olsen & Chiddo, 2008), JavaScript Object Notation (JSON; Ecma International, 2017), and Dublin Core (DCMI Usage Board, 2020) metadata records (https://www.ncei.noaa.gov/pub/data/metadata/published/paleo/). Currently, studies contributed to the WDS-Paleo after March 2020 contain PaST-compliant variables in their text data files, study landing pages, and DIF, JSON, and Dublin Core metadata records. Data sets contributed prior to this time contain PaST-compliant variables only in study landing pages and metadata records. Future work will incorporate updated variable names in legacy text data files.

6. Quality Control

We have validated the PaST thesaurus using the PoolParty SKOS quality checker based on qSKOS (Mader et al., 2012; Schandl & Blumaer, 2010; Suominen & Mader, 2014). This checker tests for 27 quality issues. The SKOS specification, with its emphasis on flexibility across different vocabulary types, has only six integrity checks for consistency with the SKOS data model (World Wide Web Consortium, 2009). In addition to consistency checks for the SKOS specification, the PoolParty qSKOS checker identifies issues in three other general areas: labeling and documentation issues, structural issues, and issues specific to linked data. These additional checks are more specific to thesauri and include integrity checks recommended by the ISO-25964 standard (Martínez-González & Alvite-Díez, 2020). Examples of labeling and documentation issues are empty or missing labels, missing definitions or other documentation, and unprintable characters. Structural issues include orphaned concepts (i.e., having no relationships) and cyclic hierarchical relationships. Constraints specific to linked data include checks for links to other vocabularies and for use of the current SKOS name space.

The PoolParty qSKOS validator focuses on aspects of thesauri for which automated checking can be straightforwardly implemented. Kless and Milton (2010) review criteria for assessing a thesaurus more generally, including aspects that are not easily integrated into automated processing. Their quality measures include the degree to which concepts are relevant and exhaustive, have appropriate granularity or specificity, are conceptually clear, have correct relationships and complete documentation, and are easy to navigate. Many of these criteria are subjective and require subject matter expertize for evaluation. Advisory panels such as those we convened can provide feedback on some of these criteria, and these are an important aspect of planned PaST governance (Section 8). Other researchers have proposed automated comparison of different vocabularies or ontologies to assess these more difficult syntactic or semantic aspects, and this is an area of active research (Lacasta et al., 2016).

7. Interfaces for Data Set Discovery Using PaST and Thesaurus Term Searching

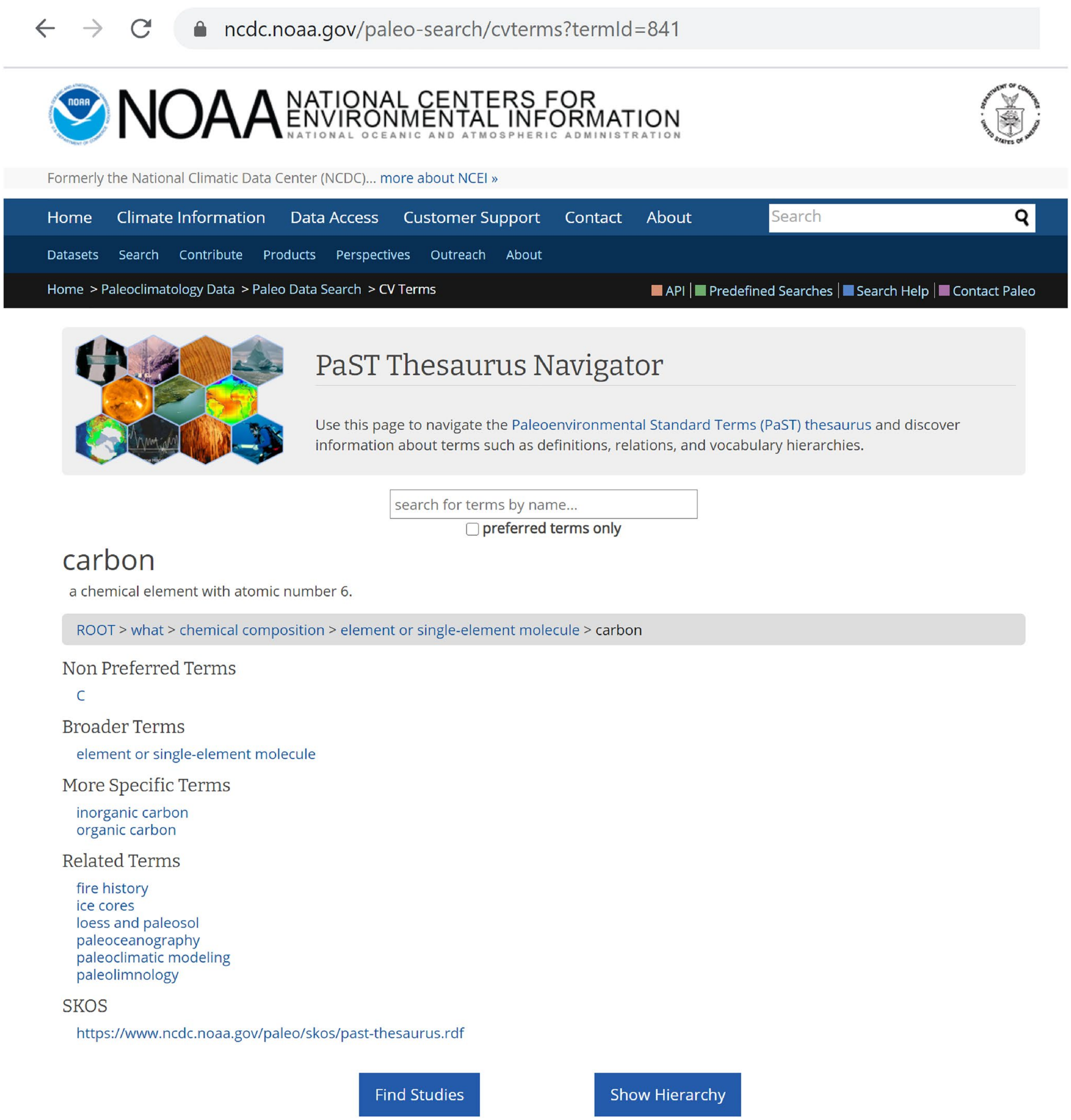
The PaST thesaurus is the basis of a new search for measured and inferred variables among WDS-Paleo data holdings (https://www.ncdc.noaa.gov/paleo-search). This search capability accesses nearly all data sets housed at the WDS-Paleo (Table 1). Previously, users of the WDS-Paleo searched for data sets with a particular measured variable by using a free-text search or by selecting GCMD parameter keywords. As discussed in Section 1, both of these methods have significant limitations.

The new advanced search by variable, which searches the what, material, data type, and seasonality components of PaST, has greatly improved data discovery. For some common searches, the new fielded variable search yields three to five times more results than the previous search options. For example, when the PaST-based variables search was first released in March 2018, this search returned 296 results for “sea surface temperature” while the GCMD keyword-based search returned 32 studies. This difference results largely from the fact that the WDS-Paleo adopted GCMD parameter keywords about a decade ago and not all legacy data sets have been tagged with them. Furthermore, some of the keyword-tagged records relate to sea surface temperature but do not have this actual variable in their data table. A second example is a search for delta 13C values measured on benthic foraminifers. Prior to the PaST-based variables search, the WDS-Paleo free-text general search was the main option for finding this variable. In March 2018, the free-text search returned 43 results, of which only 26 had the desired variable. Because the free-text search matches strings in fields such as publication abstracts and study notes, it is not guaranteed that the entered text string is a variable in the data table. The PaST-based advanced search, however, returned 137 studies, all with the desired variable. This example also highlights the power of PaST's hierarchical structure. A search with “benthic foraminifer” specified as the material yields all studies using either that term or a narrower term (i.e., a particular benthic foraminifer species).

While the WDS-Paleo offers a cross-repository search for PANGAEA and Neotoma data sets (https://www.ncdc.noaa.gov/paleo-search), PaST cannot yet be used for cross-repository variable searches because the different vocabularies used by each data provider have not been cross-walked (Section 3). Currently, cross-repository searching options include data set type, investigators, locations, and free text. Integration of variables vocabularies across these data providers is an important future direction.

Terms in the PaST thesaurus can be accessed and visualized in several ways. From the PaST Navigator (https://www.ncdc.noaa.gov/paleo-search/cvterms), users can search and browse terms, their definitions, non-preferred labels, and relationships (Figure 4). A search function in the Navigator returns terms matching a text string, and uses text matches to non-preferred terms to point to preferred terms. Browsing between terms is accomplished by selecting broader, narrower (more specific), or related terms. A visualization tool in the Navigator displays dendrograms showing the hierarchical relationships between terms (“Show Hierarchy”; Figure 5). The Navigator also allows users to find data sets that have specified what, material, or seasonality terms (“Find Studies”). Lastly, the current thesaurus including changes made since the last released version may be downloaded from the Navigator in SKOS format. All previous thesaurus versions in SKOS format are available from the PaST thesaurus landing page (https://www.ncdc.noaa.gov/paleo/study/31892). Given that SKOS is an RDF format, it is easily queried by tools such as RDFLib in Python, enhancing the capabilities of PaST for users compiling large amounts of paleoclimate data.

In addition to the graphical user interface, the WDS-Paleo provides an Application Programming Interface (API) for data discovery (https://www.ncdc.noaa.gov/paleo-search/api). Users can search via the API for data sets with particular “what,” “material,” “data type,” and “seasonality” terms. Additionally, the API allows queries on individual PaST concepts, either by term name or term ID, and returns the relevant labels, relationships, and definitions for that concept in JSON format. We anticipate that this API may be of interest to many users as a comprehensive source of well-defined paleoclimatic terms.



ncdc.noaa.gov/paleo-search/cvterms?termId=841

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Formerly the National Climatic Data Center (NCDC)... more about NCEI »

Home Climate Information Data Access Customer Support Contact About
Search
Datasets Search Contribute Products Perspectives Outreach About
Home > Paleoclimatology Data > Paleo Data Search > CV Terms
API | Predefined Searches | Search Help | Contact Paleo

PaST Thesaurus Navigator

Use this page to navigate the Paleoenvironmental Standard Terms (PaST) thesaurus and discover information about terms such as definitions, relations, and vocabulary hierarchies.

search for terms by name...

□ preferred terms only

carbon

a chemical element with atomic number 6.

ROOT > what > chemical composition > element or single-element molecule > carbon

Non Preferred Terms

C

Broader Terms

element or single-element molecule

More Specific Terms

inorganic carbon
organic carbon

Related Terms

fire history
ice cores
loess and paleosol
paleoceanography
paleoclimatic modeling
paleolimnology

SKOS

https://www.ncdc.noaa.gov/paleo/skos/past-thesaurus.rdf

Figure 4. Interface of the Paleoenvironmental Standard Terms (PaST) Thesaurus Navigator, showing the entry for "carbon" as an example. The text search matches strings to either preferred terms only or to both preferred and non-preferred terms. All terms displayed are clickable links to separate entries in the thesaurus. The "Find Studies" button performs a search for all studies containing the displayed term, and the "Show Hierarchy" button produces a dendrogram for the displayed term.

8. Governance and Versioning

To manage future changes, the PaST thesaurus has a governance structure that draws upon the recommendations of ISO-25964 (International Organization for Standardization, 2011) and the practices of NASA GCMD (Global Change Master Directory, 2020). This governance structure describes how terms or relationships may be added or modified, how frequently the thesaurus should undergo expert review, and how

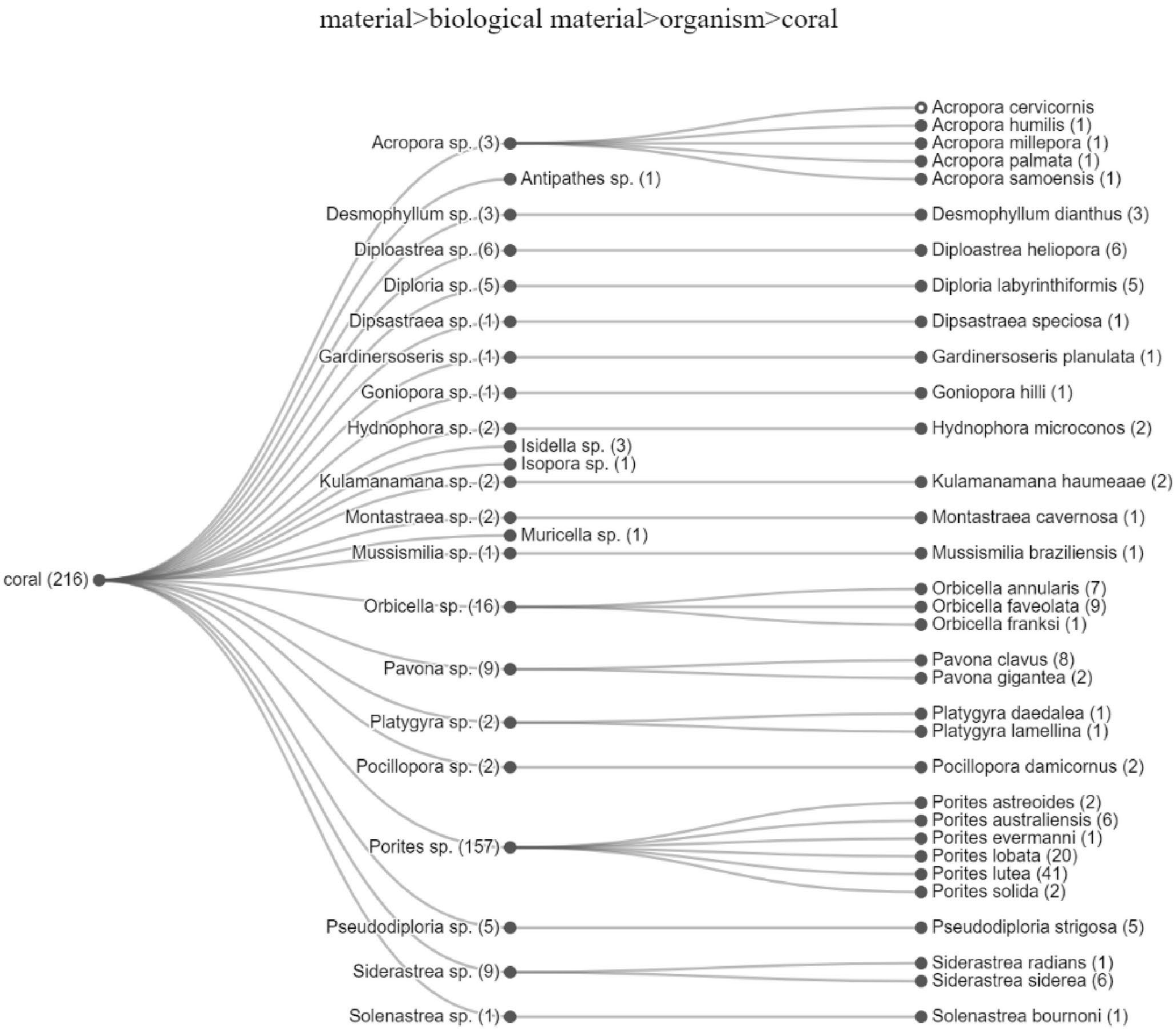


Figure 5. A dendrogram for "coral" branch of the "material" component, as produced by the Paleoenvironmental Standard Terms (PaST) Thesaurus Navigator. The number in parentheses following each term is a count of the number of studies using that term. An unfilled circle indicates a term that has not yet been assigned to any World Data Service for Paleoclimatology (WDS-Paleo) studies.

the thesaurus will be versioned. Version 1.0 of PaST's governance document is included in the Supporting Information; all current and past versions of the governance document are publicly accessible from the WDS-Paleo (https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/past-thesaurus; https://www.ncdc.noaa.gov/paleo/study/31892).

In order to maintain backward compatibility, no term will be deleted from the thesaurus. In the case that an existing term is to be renamed (e.g., species name to be changed), the original term will be preserved as a non-preferred term. The relational database structure of PaST then propagates this change in nomenclature to the landing pages and the DIF, JSON, and Dublin Core metadata records of all impacted studies, as well as to the WDS-Paleo searching capabilities. This process enforces continued standardization of variable metadata. Maintaining the original term as a non-preferred term is essential to (a) direct users to the correct preferred term in the PaST Navigator and (b) maintain backward compatibility for legacy data files, which we are not currently able to update programmatically. New terms may be added to the thesaurus to describe measurements being contributed to the WDS-Paleo. Data set contributors may suggest new terminology at the time of data submission. WDS-Paleo staff review all contributions for compliance with PaST terminology and to ensure consistent usage and application of terms. In the case of a proposed new term, WDS-Paleo staff will review the suggestion and add the new term, its definition, any non-preferred terms, and all relevant relationships to the thesaurus. Modifications to terms, definitions, or relationships may be made in response to a user comment or request, or following identification of a necessary update or correction by WDS-Paleo staff.

Proposed modifications to the thesaurus that are small (i.e., having little impact on other users) may be made on an as-needed basis and without external review. More significant modification (i.e., having greater impacts to other users), either by adding a new set of terms or by modifying existing entries, will be reviewed externally. When a significant subset of terms needs to be re-examined or added, the WDS-Paleo will convene a virtual advisory panel of approximately 2–4 subject matter experts to ensure that the proposed changes align with community standards.

Incremental changes occur to the thesaurus through the processes described above. The frequency of the incremental updates depends on the frequency of new contributions and of user feedback. Minor version numbers of PaST (e.g., version 2.1) represent a snapshot in time of the thesaurus and will be given following a number of incremental changes. Minor versions will be released periodically on an as-needed basis, but generally not more than once per year. Major version numbers (e.g., version 2.0), which also represent a snapshot in time of the thesaurus, will be given following significant changes to multiple subcomponents of the thesaurus. WDS-Paleo staff will convene a virtual advisory panel before each new major version release. This panel does not need to review the thesaurus as a whole, but will review the components of the thesaurus that have undergone significant change since the last major version release.

9. Conclusions

We developed the PaST thesaurus in order to: (a) provide consistent, complete, and accurate terms for measured and inferred variables, (b) enhance data set interoperability, and (c) improve data discovery. The thesaurus is now used to describe measured and inferred variables in more than 10,400 data sets housed at the WDS-Paleo and to perform advanced searches by variable name. The thesaurus meets requirements gathered from the WDS-Paleo user community, as well as recommendations from international standards organizations (i.e., International Organization for Standardization, World Wide Web Consortium).

Even though the PaST thesaurus was designed for operational use at the WDS-Paleo, it may be useful in other ways. By following recommendations for data set interoperability (e.g., using a standard data model and storing mapping relationships), integration with other vocabularies is feasible, enabling cross-repository metadata harmonization and data discovery. In particular, opportunities exist for integration with related domain repositories (e.g., PANGAEA and Neotoma) and domain-specific data resources (e.g., Linked Earth Ontology). Because the WDS-Paleo archive is large and heterogeneous, the thesaurus also provides a useful vocabulary for paleoscience more generally. Paleoscience is a highly interdisciplinary field, and using and understanding paleo data requires much specialized knowledge. By providing a standard terminology, complete with definitions and term relationships, the PaST thesaurus offers an entry point to the field for non-specialists.

Future work related to PaST is likely to focus on two areas. First, the PaST thesaurus will continue to respond to evolving community requirements. For example, the recently released PaCTS standard identifies additional proxy-specific metadata that the paleo research community wishes to track (Khider et al., 2019). Some of these metadata (e.g., calibration information for reconstructions) were not included in PaST since the primary goal was to unite different data types in a cross-proxy search. PaST is extensible, however, and finer levels of metadata detail could be added in a structured way. Second, coordination with other data providers and data centers is critical for building and maintaining metadata interoperability. This coordination first requires identifying topic areas of overlap within separate vocabularies, and then within these areas comparing specific terms or concepts to assign, for example, exactMatch or closeMatch relationships. Since paleo-relevant data centers often have different scopes and goals, it is unrealistic to expect that all will use the same vocabulary. Instead, reusing terminology where appropriate and otherwise cross-walking between terms will be important strategies to harmonize metadata, enabling federated searches and interoperability of data sets.

PaST contributes to growing efforts in the paleoscience community to establish data standards that promote findability, interoperability, accessibility, reproducibility and reusability of data sets (e.g., Khider et al., 2019; McKay & Emile-Geay, 2016; Williams, Kaufman, et al., 2018). Variables, in particular, serve as the "currency" between disparate data sets, allowing them to be discovered, reused, and combined (Qin et al., 2014). PaST will expand as new types of proxy measurements or reconstruction methodologies are developed, and can be refined as the community defines new data standards. The governance structure of the thesaurus specifies ways in which anyone in the paleoscience community can contribute to setting these standards to improve sharing and reuse of paleodata.

Several areas of research will particularly benefit from PaST. First, improved standardization of variable metadata will accelerate the reuse of paleo data in the sorts of synthesis studies that are advancing our field, as well as in creative, unanticipated avenues of research. By permitting more specialized queries on variables and by providing the information necessary to compare data sets, PaST improves the efficiency and effectiveness of data discovery for data compilations. Additionally, PaST's adoption of the SKOS model and its integration with the WDS-Paleo API enables researchers to automate some of the workflows associated with building large data compilations. These compilations permit science discovery that would not be possible from any one study alone and have recently made considerable contributions to understanding climate variability and climate sensitivity. In addition to PaST, reformatting legacy datasets to be machine-readable is critical for further automation. Second, metadata on variables also facilitate the reuse of data in interdisciplinary ways. PaST's mapping relations to vocabularies in other disciplines are an important first step in this process. Variable metadata standardized by PaST makes studies more useful to researchers across domain areas and across disciplines, promoting new uses for these studies and more interdisciplinary collaborations and citations.

Data Availability Statement

PaST, Version 1.0, is available from the World Data Service for Paleoclimatology at: https://www.ncdc.noaa.gov/paleo/study/31892.

Acknowledgments
Development of the thesaurus was supported by funding from the NSF Paleoclimate Program (AGS-1545762) and NOAA's Climate Program Office (Cooperative Agreement #NA17OAR4320101). The authors thank John Williams and two anonymous reviewers for their helpful reviews, and thank the scientists who provided feedback on drafts of the thesaurus through virtual advisory panels: Christo Buizert, Emilie Dassie, Kristine DeLong, Sarah Feakins, T. J. Fudge, Ben Hardt, Sharon Hoffmann, Bronwen Konecky, Shaun Marcott, Nick McKay, Anders Noren, Alexandra Noronha, Ian Orland, David Porinchu, Julie Richey, Jeffery Stone, Jessica Tierney, Amy Wagner, Eugene Wahl, Branwen Williams, Erika Wise, and Corinne Wong.

References

Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013). Key choices in the design of simple knowledge organization system (SKOS). *Journal of Web Semantics*, *20*, 35–49. https://doi.org/10.1016/j.websem.2013.05.001

Batchelor, C. (2020). *Chemical methods ontology*. Retrieved from http://purl.obolibrary.org/obo/chmo.owl

Borgman, C. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*, 1059–1078. https://doi.org/10.1002/asi.22634

Bowker, G. (2005). *Memory practices in the sciences*. MIT Press.

Climate and Forecast Metadata. (2020). *Standard name table v76*. Retrieved from http://cfconventions.org/standard-names.html

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*, 4023–4038. https://doi.org/10.1098/rsta.2010.0165

DCMI Usage Board. (2020). *Dublin core metadata initiative (DCMI) metadata terms*. Retrieved from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Degtyarenko, K., Matos, de, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2008). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, *36*, D344–D350. https://doi.org/10.1093/nar/gkm791

Dextre Clarke, S. G., & Zeng, M. L. (2011). From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modeling. *Information Standards Quarterly*, *24*(1), 20–26. https://doi.org/10.3789/isqv24n1.2012.04

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., et al. (2002). PANGAEA – an information system for environmental sciences. *Computers and Geosciences*, *28*, 1201–1210. https://doi.org/10.1016/S0098-3004(02)00039-0

Diepenbroek, M., Schindler, U., Huber, R., Pesant, S., Stocker, M., Felden, J., et al. (2017). Terminology supported archiving and publication of environmental science data in PANGAEA. *Journal of Biotechnology*, *261*, 177–186. https://doi.org/10.1016/j.jbiotec.2017.07.016

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., et al. (2020). *NetCDF climate and Forecast (CF) metadata conventions*. Version 1.8. Retrieved from https://cfconventions.org/documents.html

Ecma International (2017). *ECMA-404: The JSON data Interchange syntax*. Retrieved from https://www.ecma-international.org/publications/standards/Ecma-404.htm

Emile-Geay, J., & Eshleman, J. A. (2013). Toward a semantic web of paleoclimatology. *Geochemistry, Geophysics, Geosystems*, *14*, 457–469. https://doi.org/10.1002/ggge.20067

Gkoutos, G. V. (2020). *Units of measurement ontology*. Retrieved from http://purl.obolibrary.org/obo/uo.owl

Gkoutos, G. V., Schofield, P. N., & Hoehndorf, R. (2012). The units ontology: A tool for integrating units of measurement in science. *Database*. https://doi.org/10.1093/database/bas033

Global Change Master Directory (2020). *GCMD keywords*. version 9.1. Retrieved from https://earthdata.nasa.gov/gcmd-forum

Hankin, S., Blower, J. D., Carval, T., Casey, K. S., Donlon, C., Lauret, O., et al. (2010). NetCDF-CF-OPeNDAP: Standards for ocean data interoperability and object lessons for community data standards processes. *Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society*, *2*. https://doi.org/10.5270/OceanObs09.cwp.41

Harpring, P. (2010). *Introduction to controlled vocabularies: Terminology for art, architecture, and other cultural works*. Getty Research Institute.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., et al. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, *44*(D1), D1214–D1219. https://doi.org/10.1093/nar/gkv1031

Hedden, H. (2010). Taxonomies and controlled vocabularies best practices for metadata. *Journal of Digital Asset Management*, *6*, 279–284. https://doi.org/10.1057/dam.2010.29

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*, 280–299. https://doi.org/10.1353/lib.0.0036

Horton, T., Gofas, S., Kroh, A., Poore, G. C. B., Read, G., Rosenberg, G., et al. (2017). Improving nomenclatural consistency: A decade of experience in the world register of marine species. *European Journal of Taxonomy*, *389*, 1–24. https://doi.org/10.5852/ejt.2017.389

Hudson Institute of Mineralogy (2020). *Mindat.org*. Retrieved from https://www.mindat.org/

Integrated Taxonomic Information System (2020). *ITIS online database*. Retrieved from http://www.itis.gov

International Organization for Standardization (2011). *ISO 25964-1:2011 information and documentation - thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval*. Retrieved from https://www.niso.org/schemas/iso25964

International Organization for Standardization (2013). *ISO 25964-2:2013 information and documentation - thesauri and interoperability with other vocabularies - Part 2: Interoperability with other vocabularies*. Retrieved from https://www.niso.org/schemas/iso25964

Jansma, E., Brewer, P., & Zandhuis, I. (2011). TRiDaS 1.1: The tree-ring data standard. *Dendrochronologia*, *28*, 99–130. https://doi.org/10.1016/j.dendro.2009.06.009

Jonkers, L., Cartapanis, O., Langner, M., McKay, N., Mulitza, S., Strack, A., et al. (2020). Integrating palaeoclimate time series with rich metadata for uncertainty modeling: Strategy and documentation of the PalMod 130k marine palaeoclimate data synthesis. *Earth System Science Data*, *12*, 1053–1081. https://doi.org/10.5194/essd-12-1053-2020

Kaufman, D., McKay, N., Routson, C., Erb, M., Davis, B., Heiri, O., et al. (2020). A global database of Holocene paleotemperature records. *Scientific Data*, *7*, 115. https://doi.org/10.1038/s41597-020-0445-3

Khider, D., Emile-Geay, J., McKay, N. P., Gil, Y., Garijo, D., Ratnakar, V., et al. (2019). PaCTS 1.0: A crowdsourced reporting standard for paleoclimate data. *Paleoceanography and Paleoclimatology*, *34*, 1570–1596. https://doi.org/10.1029/2019PA003632

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research*, *47*(D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033

Kless, D., & Milton, S. (2010). *Towards quality measures for evaluating thesauri*. In 4th International conference on metadata and semantic research (Vol. 108, pp. 312–319). Springer.

Lacasta, J., Falquet, G., Zarazaga-Soria, F. J., & Nogueras-Iso, J. (2016). An automatic method for reporting the quality of thesauri. *Data and Knowledge Engineering*, *104*, 1–14. https://doi.org/10.1016/j.datak.2016.05.002

Mader, C., Haslhofer, B., & Isaac, A. (2012). *Finding quality issues in SKOS vocabularies*. In P. Zaphiris, G. Buchanan, E. Rasmussen, & F. Loizides (Eds.), TPDL 2012 theory and practice of digital libraries (Vol. 7489, pp. 222–233). Springer.

Marsicek, J., Shuman, B. N., Bartlein, P. J., Shafer, S. L., & Brewer, S. (2018). Reconciling divergent trends and millennial variations in Holocene temperatures. *Nature*, *554*, 92–96. https://doi.org/10.1038/nature25464

Martínez-González, M. M., & Alvite-Díez, M.-L. (2020). A semantic web methodological framework to evalute the support of integrity in thesaurus tools. *Journal of Information Science*, *46*, 378–391. https://doi.org/10.1177/0165551519837195

Mazzocchi, F., Tiberi, M., Santis, De, B., & Plini, P. (2007). Relational semantics in thesauri: Some remarks at theoretical and practical levels. *Knowledge Organization*, *34*(4), 197–214. https://doi.org/10.5771/0943-7444-2007-4-197

McKay, N. P., & Emile-Geay, J. (2016). Technical note: The linked paleo data framework for paleoclimatology. *Climate of the Past*, *12*, 1093–1100. https://doi.org/10.5194/cp-12-1093-2016

Neukom, R., Steiger, N., Gomez-Navarro, J. J., Wang, J., & Werner, J. P. (2019). No evidence for globally coherent warm and cold periods over the preindustrial Common Era. *Nature*, *571*, 550–554. https://doi.org/10.1038/s41586-019-1401-2

Olsen, L., & Chiddo, C. (2008). *Directory Interchange format (DIF) standard*. Retrieved from https://earthdata.nasa.gov/esdis/eso/standards-and-references/directory-interchange-format-dif-standard

Qin, H., Davis, L., Mayernik, M., Romero Lankao, P., D'Ignazio, J., & Alston, P. (2014). Variables as currency: Linking meta-analysis research and data paths in sciences. *Data Science Journal*, *13*, 158–171. https://doi.org/10.2481/dsj.14-030

Routson, C., McKay, N. P., Kaufman, D. S., Erb, M. P., Goosse, H., Shuman, B. N., et al. (2019). Mid-latitude net precipitation decreased with Arctic warming during the Holocene. *Nature*, *568*, 83–87. https://doi.org/10.1038/s41586-019-1060-3

Schandl, T., & Blumaer, A. (2010). *PoolParty: SKOS thesaurus management utilizing linked data*. In L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, & T. Tudorache (Eds.), 7th extended semantic web conference lecture notes in computer science (Vol. 6089, pp. 421–425). Springer.

Suominen, O., & Mader, C. (2014). Assessing and improving the quality of SKOS vocabularies. *Journal on Data Semantics*, *3*, 47–73. https://doi.org/10.1007/s13740-013-0026-0

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., et al. (2019). Last millennium reanalysis with an expanded proxy database and seasonal proxy modeling. *Climate of the Past*, *15*, 1251–1273. https://doi.org/10.5194/cp-15-1251-2019

Tierney, J. E., Zhu, J., King, J., Malevich, S. B., Hakim, G. J., & Poulsen, C. J. (2020). Glacial cooling and climate sensitivity revisited. *Nature*, *584*, 569–573. https://doi.org/10.1038/s41586-020-2617-x

Vandepitte, L., Vanhoorne, B., Decock, W., Vranken, S., Lanseens, T., Dekeyzer, S., et al. (2018). A decade of the world register of marine species – general insights and experiences from the data management team: Where are we, what have we learned and how can we continue? *PloS One*, *13*(4), e0194599. https://doi.org/10.1371/journal.pone.0194599

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Williams, D. N., Ananthakrishnan, R., Bernholdt, D. E., Bharathi, S., Brown, D., Chen, M., et al. (2009). The earth system grid: Enabling access to multimodel climate simulation data. *Bulletin of the American Meteorological Society*, *90*, 195–205. https://doi.org/10.1175/2008BAMS2459.1

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goring, S. J., et al. (2018). The neotoma paleoecology database, a multiproxy, international, community-curated data resource. *Quaternary Research*, *89*, 156–177. https://doi.org/10.1017/qua.2017.105

Williams, J. W., Kaufman, D. S., Newton, A., & Gunten, von, L. (2018). Editorial: Building and harnessing open paleodata. *Past Global Changes Magazine*, *26*, 45–96. https://doi.org/10.22498/pages.26.2

World Wide Web Consortium (2009). *SKOS simple knowledge organization system reference*. Retrieved from http://www.w3.org/TR/2009/REC-skos-reference-20090818/

World Wide Web Consortium (2014). *RDF 1.1 concepts and abstract syntax*. Retrieved from http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

WoRMS Editorial Board (2020). *World register of marine species*. Retrieved from http://www.marinespecies.org

Zeng, M. L. (2008). Knowledge organization systems (KOS). *Knowledge Organization*, *35*(2–3), 160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160