

The extending Ocean Drilling Pursuits (eODP) Project: Synthesizing Scientific Ocean Drilling Data

Jocelyn Sessa¹, Andrew J. Fraass², Leah J. LeVay³, Shanan E Peters⁴, and Katie Marie Jamson²

¹Academy of Natural Sciences of Drexel University

²University of Victoria

³International Ocean Discovery Program, Texas A&M University

⁴University of Wisconsin-Madison

November 24, 2022

Abstract

For over fifty years, cores recovered from ocean basins have generated extensive fossil, lithologic, and chemical archives that have revolutionized the fields of plate tectonics and oceanography, and significantly improved our understanding of climate change. Although scientific ocean drilling (SOD) data are openly available after each expedition, formats for these data are heterogeneous. Furthermore, lithological, chronological, and paleobiological data are typically separated into different repositories, limiting researchers' abilities to discover and analyze integrated SOD data sets. Emphasis within Earth Sciences on adhering to FAIR Data Principles and the establishment of community-lead databases provide a pathway to unite SOD data and further harness the scientific potential of the investments made in offshore drilling. Here, we describe a workflow for compiling, cleaning, and standardizing key SOD records, and importing them into the Paleobiology Database (PBDB) and Macrostrat, systems with versatile, open data distribution mechanisms. These efforts are being carried out by the extending Ocean Drilling Pursuits (eODP) project. eODP has processed all of the lithological, chronological, and paleobiological data from one SOD repository, along with numerous other datasets that were never deposited in a database; these were manually transcribed from original reports. This compiled dataset contains over 78,000 lithological units from 1,048 drilling holes from 390 sites. Over 26,000 fossil-bearing samples, with 5,280 taxonomic entries from 13 biological groups, are placed within this lithologic spatiotemporal framework. Information is available via the PBDB and Macrostrat application programming interfaces, which render data retrievable by a variety of parameters, including age, taxon, site, and lithology.

Hosted file

essoar.10512156.1.docx available at <https://authorea.com/users/551035/articles/604206-the-extending-ocean-drilling-pursuits-eodp-project-synthesizing-scientific-ocean-drilling-data>

The extending Ocean Drilling Pursuits (eODP) Project: Synthesizing Scientific Ocean Drilling Data

Jocelyn A. Sessa¹, Andrew J. Fraass^{1,2}, Leah J. LeVay³, Shanan E. Peters⁴, Katie M. Jamson²

¹*Academy of Natural Sciences of Drexel University, Philadelphia, PA, United States.* ²*School of Earth and Ocean Sciences, University of Victoria, Victoria, B.C., Canada.* ³*International Ocean Discovery Program, Texas A&M University, College Station, TX, United States.* ⁴*Dept. of Geoscience, University of Wisconsin-Madison, Madison, WI, USA 53706*

Corresponding authors: Jocelyn A. Sessa; Andrew J. Fraass (jsessa@drexel.edu; andyfraass@uvic.ca)

Key Points:

- Scientific ocean drilling has produced vast amounts of data; however, they are not archived in a way that meets FAIR data principles.
- The extending Ocean Drilling Pursuits project standardizes lithology, paleontology, and age data across decades of drilling programs.
- This project has migrated datasets to existing, open-access, searchable databases to enable scientific research.

Abstract

For over fifty years, cores recovered from ocean basins have generated extensive fossil, lithologic, and chemical archives that have revolutionized the fields of plate tectonics and oceanography, and significantly improved our understanding of climate change. Although scientific ocean drilling (SOD) data are openly available after each expedition, formats for these data are heterogeneous. Furthermore, lithological, chronological, and paleobiological data are typically separated into different repositories, limiting researchers' abilities to discover and analyze integrated SOD data sets. Emphasis within Earth Sciences on adhering to FAIR Data Principles and the establishment of community-lead databases provide a pathway to unite SOD data and further harness the scientific potential of the investments made in offshore drilling. Here, we describe a workflow for compiling, cleaning, and standardizing key SOD records, and importing them into the Paleobiology Database (PBDB) and Macrostrat, systems with versatile, open data distribution mechanisms. These efforts are being carried out by the extending Ocean Drilling Pursuits (eODP) project. eODP has processed all of the lithological, chronological, and paleobiological data from one SOD repository, along with numerous other datasets that were never deposited in a database; these were manually transcribed from original reports. This compiled dataset contains over 78,000 lithological units from 1,048 drilling holes from 390 sites. Over 26,000 fossil-bearing samples, with 5,279 taxonomic entries from 13 biological groups, are placed within this lithologic spatiotemporal

framework. All information is available via GitHub and Macrostrat’s application programming interface, which renders data retrievable by a variety of parameters, including age, taxon, site, and lithology.

1 Introduction

Scientific ocean drilling (SOD), through the International Ocean Discovery Program (IODP) and its predecessors, has a far-reaching legacy. Since its inception in the 1960s, SOD has produced vast quantities of marine data, the results of which have revolutionized many geoscience subdisciplines (e.g., O’Connell, 2019). For example, SOD supplied conclusive evidence of plate tectonics via seafloor spreading and demonstrated that both abrupt and gradual changes in climate are driven by variability in Earth’s orbit. Meta-analytical studies from SOD efforts exist for paleontology (e.g., Lazarus, 1994; Bown et al., 2004; Bown, 2005; Fraass et al., 2015; Fenton et al., 2016; Trubovitz et al., 2020; Lowery et al., 2020; Jamson et al., 2022a), paleotemperature (e.g., Zachos et al., 2001; 2008; Dunkley Jones et al., 2013), and marine sedimentation (e.g., Lyle, 2003; Pälke et al., 2012; Peters et al., 2013; Wade et al., 2020) but they are few due to the decentralized nature of the data. Each study of sedimentation, for example, requires another synthesis of data from numerous sources; a slow, difficult process that limits reproducibility and is largely a redundant effort.

Parallels can be drawn with the study of Phanerozoic marine diversity, which began with individual researchers each assembling datasets (e.g., Raup, 1976; Sepkoski, 1981) that necessitated considerable time (see Sepkoski’s 1993 paper entitled “Ten Years in the Library...”). Such studies (e.g., Alroy et al. 2001, 2008; Peters, 2008; Alroy, 2010a,b) are now much more easily accomplished and are reproducible because of the Paleobiology Database (PBDB), which enables the quantification of the fossil record across both space and time. Since its inception in 1998, the PBDB has enabled groundbreaking studies on the patterns and causes of biodiversity, the origins and development of biological communities and their complexity, and has helped inform efforts to characterize and mitigate biodiversity loss (e.g., Bowen et al., 2002; Villier and Korn, 2004; Kiessling and Kocsis, 2016; Ivany et al., 2018). SOD data could be but are not commonly used in a similar fashion because the data are not housed in an easily accessible database. Most unmodified (that is, the original interpretations of taxonomy and age from the scientists aboard the drilling ships) SOD data are housed in three distinct online repositories that are not readily searchable. This means that users must already know what they are looking for and where to go to find it. Furthermore, some SOD data, particularly age models and lithologic records, are decoupled from related datasets and not available online. All of these issues hinder the investigation of large-scale temporal and geographic patterns because researchers must create both the datasets and analytical tools for each study, in isolation, in contrast to the group of paleobiologists who generate open data as well as analysis and visualization tools built around those data (see the PBDB website’s “Resources” section). The lack of a shared and

integrated SOD database is a known issue in paleoceanography (e.g., Greene and Thirumalai, 2019) and has been a source of recent work (e.g., Khider et al., 2019). Establishing a community-led, open-source ecosystem of SOD fossil and stratigraphic data is vital for achieving many paleoceanography and marine sedimentary geology research goals, such as quantifying regional to global biodiversity (including the effects of mass extinction), food web interactions, and marine sedimentation and sediment subduction trends (i.e., Müller et al., 2022).

In this paper, we introduce the extending Ocean Drilling Pursuits (eODP; <https://eodp.github.io/>) project, which is building capabilities for the improved use and reuse of SOD data via existing databases. The nexus for creating a unified SOD database system is stratigraphy; the age and environment of deposition is the foundation upon which all sedimentary research is built. In current SOD databases, stratigraphic data are stored unconnected to other data, including the fossil occurrences found within these layers (Fig. 1; Table 1). SOD has generated large amounts of data, but these data are of limited value without their meta-stratigraphic context. The eODP project is facilitating the curation, access, analysis, refinement, and visualization of comprehensive and integrated marine fossil and stratigraphic datasets by adapting several established databases and tools. Macrostrat, a stratigraphic database that stores ages, sediment thicknesses, and lithologies in easily accessible and flexible formats, provides the spatiotemporal stratigraphic scaffolding that a unified SOD ecosystem requires, while the PBDB stores paleontological records and has considerable taxonomic capabilities to deal with more than 50 years of evolving taxonomic concepts for SOD fossils. The goal of eODP is to make SOD data easily accessible and manipulatable by geoscientists, oceanographers, and biologists by adhering to Findable, Accessible, Interoperable, and Reusable (FAIR) Data Principles (Wilkinson et al., 2016).

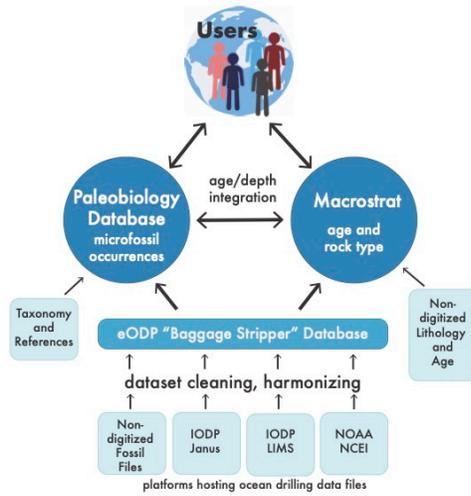


Figure 1. Schematic of the eODP ecosystem, displaying data sources, data types, databases, and connectivity amongst the various components. The three data types that eODP is focusing on are age models, fossil occurrence and abundance data, and lithology. NOAA = National Oceanic and Atmospheric Administration; NCEI = National Centers for Environmental Information; IODP = International Ocean Discovery Program; LIMS = Laboratory Information Management System.

Scientific Ocean Drilling data sources by governing program & years of operation, & method of access

Data Type:	Deep
fossil occurrences & abundances	NOA
age model	NOA
lithology	NOA

Table 1. Data types, websites currently housing data, and method of data access for phases of Scientific Ocean Drilling. Several datasets were manually transcribed, including all of the Janus-era lithology, some microfossil files, and most age model information. OCR = optical character recognition; NOAA = National Oceanic and Atmospheric Administration; NCEI = National Centers for Environmental Information; LIMS = Laboratory Information Management System.

1.1 SOD data sources, data types, and data management

The International Ocean Discovery Program (IODP) is a multi-country collaboration to study Earth Science, primarily using deep-coring vessels (see OConnell, 2019 and references within for a detailed history of SOD programs and for the revolutionary effects SOD has had on understanding Earth processes). IODP has had several predecessor programs dating back over 50 years: Deep Sea Drilling Project (DSDP) from 1968-1983; Ocean Drilling Program (ODP) from 1983-2003; and Integrated Ocean Drilling Program from 2003-2013 - each with its own way of formatting and archiving data (Table 1). During each expedition (formerly known as legs), a drillship visits multiple sites or localities and drills or cores one or more holes at each site. The shipboard scientists gather information on the rocks, sediments, and fossils recovered from the cores. This information includes, but is not limited to, detailed macroscopic and microscopic lithologic descriptions, physical properties measurements, geochemistry, magnetic properties, and paleontology. Age-depth relationship interpretations are constructed shipboard using indicator fossils, magnetic polarity, and occasionally using well-dated and described marker beds, such as volcanic tephras. These shipboard data are published in Initial Reports or Proceedings Volumes (hereafter termed shipboard reports) and are stored in three online sources: the National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Information (NCEI) International Ocean Drilling Data Archive houses all DSDP data and ODP data from Leg 100 to 129; the SOD database 'Janus' stores some data from ODP Leg 129 through Integrated Ocean Drilling Phase I Expedition 312; and the IODP database 'LIMS' (Laboratory Information Management System) stores data from Integrated Ocean Drilling Phase II and IODP Expeditions 317 to present (Table 1). These sources do not include data from sites cored by the Chikyu vessel or Mission Specific Platform Expeditions (for example, Expeditions 313-316).

There are several datasets not available in any online repository (Table 1), such as lithologic descriptions from the Janus era, which are only available from the core description forms (standardized lithological/stratigraphic columns of individual cores that are published within the shipboard reports). Neither general geological ages of cored material nor detailed age models are available online for most SOD programs. Additionally, some fossil data were not incorporated into any database and were instead stored as a large table within the shipboard reports. Due to the fractured nature of SOD data, it is currently difficult to even estimate the total magnitude of the data store. As of April 2022, there have been 282 completed expeditions that visited a total of 1601 sites (https://www.iodp.tamu.edu/publicinfo/ship_stats.html). The Janus database alone contains over 1 million fossil occurrences, while to date the PBDB in total contains roughly 1.5 million, most of which are derived from continental outcrops of marine and terrestrial rock units, i.e., environments not typically recorded in offshore ocean basins.

The fossil remains of animals, plants, bacteria, protists, and fungi are all found

within SOD samples and are generally lumped under the descriptive (rather than taxonomic) terms ‘micropaleontology’ and ‘microfossil’ because their small size necessitates a microscope to study them. The most common taxonomic groups found within SOD samples are listed in Table 2. Both the preservation and abundance of all microfossil taxa within a SOD sample are recorded via terms that are fairly standardized across expeditions but vary between fossil groups because of the differing sampling processing and counting methodologies required for each group. Microfossils have a robust species-level record (Ezard et al. 2011; Fraass et al. 2015; Jamson et al., 2022a), a true novelty in paleobiology, and thus form a rich dataset for addressing many of the questions highlighted in ‘Grand Challenges in Paleobiology’ 2017 EarthRates’ Report (see the ‘Grand Challenges’: earthrates.org/news/earthrates-community-news-2/) at unprecedented levels of specificity.

There are ongoing efforts to mobilize key SOD data. Notably, the Neptune Sandbox (NSB) is a database of microfossil occurrences and age-depth relationships, largely constructed with postcruise age models (Renaudie et al., 2020). NSB is a tremendous resource for the paleoceanographic community, as it has focused on key sites with highly resolved chronologies and has been a source of important work for decades (e.g., Spencer-Cervato et al., 1994; Trubovitz et al., 2020). However, NSB does not include all sites, all shipboard microfossil data, nor does it include lithology logs. The targeted approach taken by NSB is complementary to the eODP project goals and the two projects are actively aligning efforts. For example, all of the age models from NSB are incorporated into Macrostrat as alternative age models for sites currently included in NSB.

Bolboformids	Rick McCourt (Academy of Natural Sciences); Marina Potapova (Academy of Natural Sciences)
Calcareous Nannoplankton	Leah LeVay (International Ocean Discovery Program; Texas A&M University)
Chrysophyte cysts	
Diatoms	Beth Caissie (USGS Menlo Park); Marina Potapova (Academy of Natural Sciences)
Dinoflagellates	Vera Pospelova (University of Minnesota)
Ebridians	

Bolboformids	Rick McCourt (Academy of Natural Sciences); Marina Potapova (Academy of Natural Sciences)	
Benthic Foraminifera	Ellen Thomas (Wesleyan University)	
Planktic Foraminifera	Andy Fraass (University of Victoria)	
Ostracods		
Pollen/Palynology	Vera Korasidis (palynology; University of Melbourne); Chelsea Smith (plants; Academy of Natural Sciences); Caroline Stromberg (phytoliths; University of Washington)	
Radiolarians	David Lazarus (Museum für Naturkunde)	
Silicoflagellates	Marina Potapova (Academy of Natural Sciences); Diane Winter (Academy of Natural Sciences)	
Websites used for taxonomic authority & opinion information		
Site	Taxonomic Group(s)	
Algaebase	diatoms, nannoplankton, silicoflagellates	https://www.algaebase.org
DINOFLAJ3	dinoflagellates	http://dinoflaj.smu.ca/dinoflaj3/index.php/Main_Page
Mikrotax/Nannotax	planktic foraminifera, nannoplankton, radiolaria	http://www.mikrotax.org
Paleobiology Database	all fossil taxa; some extant taxa	https://paleobiodb.org/
Palynodata database	dinoflagellates, acritarchs	https://paleobotany.ru/palynodata

Bolboformids	Rick McCourt (Academy of Natural Sciences); Marina Potapova (Academy of Natural Sciences)	
Tropicos	botanical taxa	https://tropicos.org/home
World Register of Marine Species (WoRMS)	all extant marine taxa; some fossil taxa	http://www.marinespecies.org

Table 2. Major microfossil groups found within SOD samples, taxonomic experts consulted, and websites used for taxonomic authority and opinion information.

1.2 eODP databases

1.2.1 The Paleobiology Database

Created in 1998, the PBDB is an open data and software infrastructure centered around globally-distributed, geographically and taxonomically explicit fossil occurrence data on all organisms through all time periods. Included in the data system are “bibliographic references, taxonomic names, taxonomic opinions on synonymies and classifications, primary collection data, taxonomic occurrences, and re-identifications of occurrences” (Uhen et al., 2013). As of June 2022, the PBDB contained over 81,000 references and 458,000 taxonomic names, with over 881,000 opinions on the classification of those names; over 1,560,000 occurrences from more than 225,000 collections were also available. These data were entered by over 674 contributors from many institutions around the world. Anyone can access all of the data via the PBDB websites and REST-ful application programming interface (API). The PBDB API originated in October 2014, and since then has received hundreds of millions of requests from many different types of clients distributed around the globe. Documentation for the API is publicly available on the paleobiodb.org website (Peters and McClennen, 2016).

1.2.2 Macrostrat

Macrostrat was created circa 2005 to aggregate chronostratigraphically-stacked rock units and their properties in order to enable quantitative analyses of regional- and continent-scale patterns in the rock record (Peters, 2005, 2006, 2008; Peters et al., 2013; Fraass et al., 2015; Peters and Husson, 2017; Husson and Peters, 2017; Peters et al., 2018; Peters et al., 2022). Although the database has primarily been used to store and analyze generalized rock columns representative of relatively large geographic regions (see references above), the fundamental structure of the database is scale agnostic, making it possible to store detailed measured sections and age models for them within the same data

framework as lithostratigraphic-scale generalized columns. As of June 2022, the database contained 2,163 such regional rock columns, with 40,960 rock units distributed in North America, the Caribbean, New Zealand, South America and the deep sea realm. Continuous-time age models, generated initially algorithmically on the basis of imprecise but generally accurate constraints provided by stratigraphic superposition and correlations of units to chronostratigraphic bins that are in turn correlated with the current international timescale (Cohen et al., 2013, mod. 2022), are a key feature of Macrostrat. The SOD portion of the Macrostrat dataset, prior to eODP, comprised 387 columns with 7,124 units, with a temporal sedimentary package hiatus structure (*sensu* Peters, 2006, 2008) defined by calcareous nannoplankton zones (Peters et al., 2013; Fraass et al., 2015). Correlations of these zones to the international timescale, and stacking order of sedimentary units, are used to assign a preliminary age model to these records. Macrostrat has several simple user interfaces that aid in discovery and utilization of the data. The open API serves as the basis for these interfaces and is quite versatile; it is currently used in multiple research applications and mobile software tools, such as Rockd (Mobile app), Mancos (iOS), and Flyover Country (Mobile app).

2 Data harmonization

What follows is an overview of the workflow used to compile fossil, age, and lithology records from the three distinct SOD platforms and from the shipboard reports, to clean and standardize these records, add them to the intermediate database entitled ‘Baggage Stripper’, and incorporate them into Macrostrat and the PBDB (Fig. 1). The name ‘Baggage Stripper’ is an acknowledgement that SOD data carry a variety of structural baggage from over 50 years of data collection, and this portion of the eODP project is attempting to remove that baggage to make the data more interoperable. The ‘BS’ acronym is entirely accidental. The Python scripts and Jupyter notebook used to process eODP data from NOAA, LIMS, and Janus are available at: <https://github.com/eODP/data-processing> (see also Kwan et al., 2022), Python scripts to insert the eODP data into the BS database are available at: <https://github.com/eODP/api>, and the workflow of adding taxonomic data and associated fields to the PBDB is available at: <https://github.com/eODP/files-for-Sessa-2022> and in the supplement. Shipboard data were accessed from the data sources listed in Table 1 by developers at Whirl-i-Gig (<http://www.whirl-i-gig.com/>), who created the processing scripts and the BS database, performed initial cleaning with input as needed from the authors, and who then supplied the compiled datasets to the authors to clean and standardize.

There are numerous SOD datasets that are not available in any online repository, such as lithologic descriptions from the Janus era. This has necessitated transcribing the lithologic data manually and through optical character recognition (OCR) from the core description forms. Age-depth relationships are also not available in a standardized, digitized format that adheres to FAIR principles for most SOD programs and therefore also had to be manually transcribed from

the shipboard reports. In addition, sometimes fossil data were not incorporated into any database and were instead stored as a large table pdf within the shipboard reports. Acquiring these data required OCR software to process the pdfs, sometimes passing through Microsoft Excel to restructure the text as a table, then manual checking and formatting of the resulting fossil tables.

2.1 SOD header harmonization

Whirl-i-Gig developers first produced a unified data structure spreadsheet whereby the columns consisted of all headers used in SOD spreadsheets. There are ~250 headers, arranged in three eras (Leg/Exp. 1-96, 101-210 [130-210 paleontology only], 317-Present) and four categories: 19 common headers (e.g., Expedition, Site, Top [cm], Top Depth [m]), 27 lithology headers (e.g., Lithology Prefix, Lithology Principal Name, Lithology Suffix, Color, Minerals, Bioturbation Intensity, Bioturbation Type), 32 micropaleontological headers (e.g., Taxon Name, Taxa Comments, Micropal Scientists, Group Abundance), and 38 chronostratigraphic headers (e.g., Sample Age, Sample Zone, Source). Fraass and LeVay harmonized headers from the three eras. LIMS lithologic data was harmonized by Peters directly into Macrostrat with aid from LeVay and Fraass. This direct ingestion of LIMS data into Macrostrat was necessary because LIMS had the widest variety of headers; for example, “Comment”, “Comments”, “COMMENTS”, “comments”, “Comment (general)”, “General comment”, “Sample comment”, “BF comment”, and “Nannofossil comment” are all the same type of information functionally, but were classified as unique headers that were then manually harmonized. The hard rock files were imported into the BS database without harmonization (i.e., no effort was made to standardize terms like “2ND crystal roundness” and “2ND lithic roundness” across legs/expeditions) and are therefore not included in the above counts. Additionally, some headers could be further harmonized, but this would require additional data transformations. For example, in the chronostratigraphic data, ages had been stored as zonation schemes, datums, numerical age values, and with minimum, maximum, and average values, or sometimes as simple single values. All of this variability was retained in the BS. The explicit goal of the BS database is to harmonize the data structure as thoroughly as possible without modifying the underlying data, unless it was found to be clearly in error (e.g., misspelling of taxonomic names).

2.2 Taxonomic data workflow

Whirl-i-Gig developers compiled lists of all unique taxonomic names for each of the three data sources listed in Table 1, starting with LIMS. While most of the LIMS data were processed in bulk, some data files were misformatted and required individual processing. Numerous fossil datasets that were not incorporated into any database were transcribed manually (Table 1) and incorporated into this first data batch. The validity of all generic and higher names was checked by Sessa, with assistance as needed from LeVay, Fraass, and the researchers listed in Table 2 and by utilizing the websites listed in Table 2. Prior to the import of these taxonomic lists into the PBDB, we

first added 'taxonomic backbones' - taxonomic hierarchies of the taxa within the compiled dataset for each group listed in Table 2 - to the PBDB (see <https://github.com/eODP/files-for-Sessa-2022> and the supplement for PBDB taxon ID numbers and resolved and original taxonomic names). While SOD data are generally resolved to the species level, and species are the desired unit for research, there are considerably more species than genera. Also, species are automatically linked to a taxonomic backbone because they are always associated with genera. Thus, genera were the most efficient target for this first import. Some species names were validated on an ad hoc basis during this initial stage. As directed by research goals, species within key groups will be subjected to this workflow once all SOD generic and higher names have been entered into the PBDB taxonomic backbones.

The steps taken during the cleaning and standardizing of taxonomic entries included: correcting misspellings; standardizing to 'indet.' for all names above the generic level (many of these entries were just the higher name or included 'sp.');

standardizing informal names to formal (ex., "Miliolids" becomes "Miliolidae indet."); and moving authority, preservation, and morphologic and other descriptors into other comment fields (ex., "*Ethmodiscus* sp. fragments" - "fragments" is moved to the 'Comment' field; "*Rouxia* sp. spatulate long heteropolar (MIS)" - "spatulate long heteropolar (MIS)" is moved to the Comment field). Statistics on the resulting cleaned and standardized taxonomic dataset are provided in Table 3. For some groups, such as planktic foraminifera and ostracods, the PBDB already contained a fairly comprehensive backbone; for other groups, such as benthic foraminifera and diatoms, the backbone needed to be built nearly from scratch - compare the number of references, taxonomic opinions, and authorities entered into the PBDB for each group in Table 4.

Group	# of entries	# of distinct taxonomic entri
Bolboformids	2	2
Calcareous Nannoplankton	937	810
Chrysophyte cysts	1	1
Diatoms	717	652
Dinoflagellates	57	53
Ebridians	8	6
Benthic Foraminifera	1689	1508
Planktic Foraminifera	1019	844
Ostracods	16	15
Pollen/Palynology	142	132
Radiolarians	642	598
Silicoflagellates	33	23
Other (ex., Tintinnids; Tunicates)	15	11
TOTAL	5,278	4655
Unique higher taxonomic names (any taxon above genus)	85	

Group	# of entries	# of distinct taxonomic entries
unique genera	1060	
unique below genus	4542	
unique below species	138	

Table 3. Taxonomic entries in the first eODP-compiled dataset. “# of entries” is the total count of all entries within a taxonomic group, whereas “# of distinct taxonomic entries” is the number of valid taxonomic entries (e.g., ‘*Chaetoceros* spp.’ and ‘*Chaetoceros* spp. and similar spores’ are two entries and one distinct taxonomic entry); ‘unique below genus’ is all genus-species pairs except ‘sp.’ and ‘spp.’ Bolded taxa are the most diverse groups in the dataset, with benthic foraminifera representing 32% of all entries, followed by planktic foraminifera at 18% and calcareous nannoplankton at 17%.

Fossil Group	# References	# taxonomic opinions	# taxonomic names
Bolboformids	18	21	14
Calcareous Nannoplankton	68	49	72
Chrysophyte cysts	2	2	2
Diatoms	116	233	206
Dinoflagellates	33	27	14
Ebridians	7	7	7
Benthic Foraminifera	184	305	143
Planktic Foraminifera	14	15	7
Ostracods	1	1	1
Pollen/Palynology	44	453	74
Radiolarians	31	51	18
Silicoflagellates	18	21	13
Other (ex., Tintinnids; Tunicates)	1	2	1
Total	537	1187	572

Table 4. Number of references entered into the Paleobiology Database to create taxonomic backbones for all generic and higher names in the first eODP-compiled dataset.

The PBDB provides several advantages for housing these taxonomic data because substantial taxonomic tools have been incorporated into it over the years. For example, the PBDB tracks multiple taxonomic opinions. One example, as shown in Fig. 2, is the taxonomic nomenclature of the planktic foraminifera genus *Globorotalia*. The taxonomy of *Globorotalia* is complex, as over time this genus has been ascribed subgenera, which then were sometimes formally or informally elevated to genera. All of these revisions can and will be incorporated and stored within the PBDB. The PBDB also contains tools to disambiguate and keep separate taxonomic homonyms, which are taxonomic names that are

spelled identically but belong to two or more separate taxa, e.g., ‘*Emiliania*’ is both the name of a calcareous nannoplankton genus and a now invalid genus name of a bivalve (i.e., Sánchez, 2010). At this initial stage, our focus has been to generate the taxonomic backbones, rather than updating the various taxonomies to the current state-of-the-art, though some revisions were entered into the PBDB on an ad hoc basis.

peleobiodb.org/classic/checkTaxonInfo?taxon_no=1521&is_real_user=1

Main Menu About Download Search Actions Search the database

Basic info	Taxonomic history	Classification	Included Taxa
Morphology	Ecology and taphonomy	External Literature Search	Age range and collections

Globorotalia

Globorotaliidae

Taxonomy

Globorotalia was named by Bandy (1972) [Sepkoski's age data: T Mi-m R]. It is extant.

It was assigned to Globorotaliinae by Loeblich and Tappan (1984); to Globigerinidae by Gibson (1983) and Wienrich (1997); to Foraminiferida by Sepkoski (2002); and to Globorotaliidae by Snyder et al. (1983) and **Lam and Leckie (2020)**.

Species

G. (Hirsutella), G. (Menardella), G. (Truncorotalia), G. (Turborotalia), G. conica, G. crassata, G. hirsuta hirsuta, G. merotumida, G. minima, G. opima, G. panda, G. plesiotumida, G. truncatulinoides, G. tumida, G. ungulata

Species lacking formal opinion data

G. angulata, G. chapmani, G. munda, G. postcretacea, G. pseudomenardii, G. pusilla, G. trinidadensis, G. zealandica

[Edit taxonomic data for Globorotalia](#) - [Edit taxonomic opinions about Globorotalia](#) -
[Add/edit ecological/taphonomic data](#) -

differ when used for an individual taxon or the quantity of a particular group (e.g., planktic foraminifera). Further, the counting methods used to generate abundance vary by taxon and sometimes by expedition based on how the shipboard scientists processed samples and generated abundance data, including per field of view, per slide or tray traverse, per 300 individuals, number of fossils compared to the number of sediment particles, compared to the number of foraminifera (benthic and planktic), or the number of individuals within that particular group (e.g., percentage of a particular benthic foraminifera taxon with that entire assemblage group). It was also important to check assumptions about the abundance codes themselves; for example, in rare cases ‘F’ was used for ‘Frequent’ and not ‘Few’. To standardize these codes while ensuring that the original shipboard determinations were maintained, Peters generated a list of all codes used in each expedition by taxonomic group and Fraass collated sample processing, counting methodologies, and abundance definitions (both group and individual taxon). We have standardized these values for ease of use (ex., harmonizing ‘A’ and ‘a’ to ‘A’, ‘R?’ and ‘?R’ to ‘R?’, ‘rw’ and ‘*’ to ‘*’ for reworked, because ‘*’ is the standard symbol to denote reworking in the shipboard reports). In several instances, shipboard scientists would use an undefined code (e.g., using ‘C’ when the scheme goes directly from ‘Abundant’ to ‘Few’). In those cases, the original undefined code is retained in the abundance field of the particular taxon in the species list, and an interpreted definition of the undefined code is recorded in the comment field, as determined by Fraass. In a few instances, transitional abundances were listed (e.g., ‘C-A’) but undefined, but as both the individual values were defined, the midpoint between the two abundances was interpreted for the abundance field, with rationale provided in the comment field. All of these comments are recorded in a unified fashion. As with all eODP data, the BS database contains the original entries, the harmonized values and comments. The harmonized abundances and their corresponding units and comments were then imported into the PBDB without further modification.

“Group abundance” is a measure of the overall abundance of a particular taxonomic group in a sample and was standardized in much the same way as the abundances of the individual species. Abundance codes were harmonized across different groups and expeditions such that, for example, an ‘A’ always means ‘Abundant’. The BS contains both the original and harmonized codes and comments, and the harmonized data were incorporated into the PBDB.

The same qualitative preservational codes generally are used across taxa and expeditions/legs and therefore the standardization of preservation was comparatively simple: ‘E’ or ‘VG’ for excellent/very good, ‘G’ for good, ‘M’ for moderate/medium, ‘P’ for poor, and ‘VP’ for very poor. The BS contains unedited preservation codes; these did not require standardization. For import into the PBDB, in instances where preservation was coded as spanning categories (ex., ‘G-M’, ‘G-VG’), only the first letter was used, based on the presumption that the first letter was the most commonly seen preservation. In cases where the preservation was not contiguous (ex., ‘VG-P’, ‘G-VP’), the preservation was

recorded as ‘V’ (variable). This harmonization allowed eODP to use the existing ‘Fragmentation’ field within the PBDB’s ‘Preservation’ data table with minimal loss of data, as ‘Fragmentation’ is not a free-form field. Note that because the fragmentation of microfossil specimens is typically considered when deciding the preservation value of a sample, it is not always recorded as a distinct value within the shipboard data, and therefore not all samples have values for the Fragmentation field.

The processing of the first batch of taxonomic files began at the end of 2019, and generic and higher taxonomic names were validated by the spring of 2020, when the entry of taxonomic authorities and opinions into the PBDB for the taxonomic backbones began. The first backbone entered into the PBDB was for the calcareous nannoplankton because LeVay is an expert in this group. Over the course of a year, three undergraduate and three graduate students from Sessa’s and Fraass’s institutions entered the taxonomic data listed in Table 4. The combined efforts of these six students was equivalent to a year of full-time work. About 540 references containing 572 taxonomic names and 1,187 taxonomic opinions were entered into the PBDB.

Once taxonomic entries were cleaned and standardized, checks were run against the PBDB taxonomic backbone by Whirl-i-Gig developers using the PBDB API services to ensure that all generic and higher names were indeed within the PBDB and would be classified into their respective taxonomic hierarchies. Following these checks, taxonomic data were brought into the BS database and will then be imported from there into the PBDB (Fig. 1).

2.4 Lithology harmonization

Concurrent with the taxonomic efforts, manual entry of lithologic core descriptions not housed in online databases (ODP Leg 129 to IODP Expedition 312; Table 1) began in fall 2020. Over a year and half, four undergraduate and graduate students at Texas A&M University manually entered sediment lithologic descriptions for 12,078 individual units from 618 holes at 229 sites and 43 expeditions/legs. Descriptions are typically entered at the core level (~9.5 m resolution) and include the shipboard age assignment. The enterers worked directly off of the shipboard core summary sheets and used core depths stored in the LIMS database. After about 6 months of this workflow, one of the students developed an OCR reading program and created core summary .csv files for each hole, including age and depth. This process reduced some of the steps associated with manual entry.

Macrostrat stores hierarchical vocabularies relevant to the description of rocks; no standardization of terminology is enforced, meaning that Macrostrat accepts that there are multiple different ways to describe rocks and sediments and the focus is instead on hierarchy and nomenclature that is in use in the scientific literature (e.g., Macrostrat understands that “basalt” is a “mafic”, “volcanic”, and “igneous” rock). In order to incorporate SOD lithologic logs into Macrostrat, this lithology (and corresponding lithology attributes) vocabulary was used, allow-

ing original shipboard descriptions to persist while at the same time providing a hierarchical level of classification that allows for flexible description and retrieval. That is, it is now possible to retrieve all “carbonate”-bearing units in the SOD data, regardless of the specific lithologies assigned to the lithologic units (e.g., units described as “micrite” and “lime mudstone”, two alternative carbonate classification nomenclatural schemes, would both be retrieved in queries for “carbonate” or “sedimentary” rocks).

In practice, matching Macrostrat’s curated vocabulary of lithologies and their descriptors to the LIMS SOD data assembled by eODP required significant effort, primarily because of the heterogeneity within the LIMS data. For example, there are typically primary and minor lithology fields within the LIMS data, each of which optionally contain “prefix” and “suffix” descriptions. For example, the primary lithology in LIMS might be described as “ooze [MMK88]” with a prefix of “Clayey radiolarian” and a suffix of “with nannofossil and diatoms”. There are a total of 3,281 unique prefix-principal lithology-suffix combinations in the original LIMS dataset. Matching these terms to the Macrostrat curated vocabularies was done within the database itself. The original descriptions associated with the LIMS data remain connected to these revised and standardized descriptions, should they be required for any reason and in part because some standardization remains (e.g., for cases where spelling errors or other anomalies appear in the LIMS dataset these modifiers may not yet be included in Macrostrat, though all principal and minor lithologies have been standardized within Macrostrat’s vocabulary).

2.5 Chronostratigraphy workflow and PBDB connectivity

Similar to how the taxonomy of microfossils must be imported into the PBDB prior to further refinements of the system, Macrostrat requires a minimum level of chronostratigraphic detail (Peters et al. 2018). Because only limited age-depth relationships exist within any SOD database, eODP both manually entered age-depth information from the shipboard reports (Table 1) and collaborated with NSB to obtain stratigraphic information. In rare cases, postcruise information was used if, for example, the cruise sailed without any chronostratigraphers and therefore the shipboard reports contained no age information. The depth, core, and a variety of possible chronostratigraphic bins (e.g., calcareous nannofossil zone NN5, Eocene, and/or Chattian) were manually transcribed into web entry forms developed for Macrostrat. Additionally, NSB provided age models to the eODP team. These age-depth relationships were not incorporated in a one-to-one fashion given Macrostrat’s unit boundary-focused age model, but the ages were used to roughly calibrate the algorithm used to generate an initial age model. It is our intention that this is a halfway step, and that further development of the stratigraphy aspects of eODP will involve replication of NSB age depth relationships, with accompanying citation back to NSB. Further developments include the capability for age models to be retrieved by the PBDB and to then be served with microfossil collections to users.

2.6 Enhancing Macrostrat age models

Within Macrostrat currently, a portion of stratigraphy (a unit or subdivision of a unit) can only belong to one chronostratigraphic unit (e.g., a single biostratigraphic zone). This limitation results from the current ‘continuous time age-model’ (Peters et al., 2018), which does not allow units within a column to overlap in time. For example, a unit below cannot belong to calcareous nannofossil biozone NN2 while the unit above belongs to planktic foraminifer biozone N5, because those zones partially overlap. The solution is to place both units in a broader time bin, such as the early Miocene, preserving their stratigraphic order and resulting in a more stable but less precise stratigraphy. This is an acceptable resolution for analysis of basin-scale patterns of sedimentation over the Mesozoic and Cenozoic, like the Peters et al. (2013) study, which used Macrostrat and only calcareous nannoplankton zonations for determining marine age models. It is problematic for finer-scale analyses because zones from these larger bins may encompass several million years of geologic time. eODP plans to accommodate more complex age models within the Macrostrat schema, ideally moving towards well-defined algorithmic approaches (e.g., McKay et al., 2021). The current eODP age-depth data are confident at the Epoch scale and reasonable at the Stage level. Subdivision of eODP records below the Stage level is not advised without additional chronostratigraphic work by the end user.

3 Description of the compiled dataset and examples of results

In total, over 78,000 lithological units from 1,048 chronostratigraphically-resolved ocean drilling holes from 390 sites containing over 26,000 fossil-bearing samples with more than 5,200 taxonomic entries from 13 major biological groups form the first compiled eODP dataset. These data can be assessed via the Macrostrat API by including ‘project_id=3’ in the query.

The benthic foraminifera comprise 32% of these taxonomic entries, followed by planktic foraminifera (18%), calcareous nannoplankton (17%), diatoms (14%), radiolarians (13%), pollen and palynology (3%) and dinoflagellates (1%); all other groups listed in Table 3 comprise less than 1% of all taxonomic entries. Samples range from the late Jurassic to the Recent (Figure 3), with what is very likely a ‘Pull-of-the-Recent’ bias (Raup, 1979) that begins in the mid-Cretaceous and is particularly evident from the mid-Neogene onwards. This bias results from a more complete sampling of younger sediments relative to those in the deep past and is characteristic of global compilations of unstandardized data through geologic time (ex., Peters, 2005; Alroy, 2010b; Lowery et al., 2020). Another reason for this feature that is specific to SOD data is that younger sediments must be cored through to reach older ones.

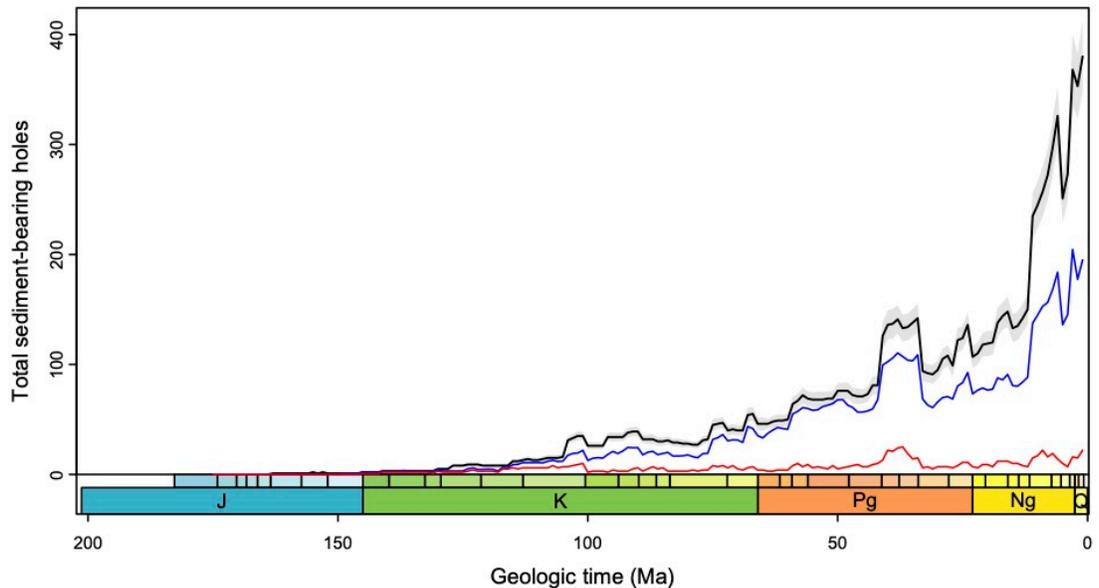


Figure 3. Number of holes bearing sediment of a particular age plotted against time. Gray envelope represents one standard deviation bootstrap error. Orange curve shows siliceous sediment, blue curve adds to the siliceous curve all carbonate-bearing sediment. The remaining lithologies encompass primarily siliciclastics (e.g., mud, siltstone).

Collating these data also allows for the generation of reproducible maps of seafloor sediments from past intervals (Figure 4). In particular, the mapping of biogenic sediments (sediments generated by organisms, i.e., diatoms producing siliceous ooze) can provide exceptional insights into how biogeochemical cycles, the preservation of sediments, and ocean-atmosphere interactions have evolved through time and across space. Most, although not all, SOD sediments are from deep ocean environments and in these settings calcareous and siliceous sediments are biogenic and not abiotically precipitated. To generate the maps in Figure 4, all eODP data within Macrostrat were downloaded using the query: https://macrostrat.org/api/units?project_id=3&response=long&status_code=in%20process&format=csv, including age, lithology, column id (a collection of units from a single location; i.e., a stratigraphic column), unit id (a portion of sediments or rock within a column, distinct from the surrounding material), thickness, and modern latitude and longitude coordinates. Sediments were defined using keywords from the lithology field as either ‘calcareous’ (keywords: ‘calcareous’, ‘foraminiferal’, ‘nannofossil’, ‘carbonate’, ‘carbonaceous’, ‘shell bed’, ‘foraminifer’, ‘chalk’, and ‘calcareous’; excluding ‘clay’, ‘sandstone’, and ‘sand’) or ‘siliceous’ (keywords: ‘siliceous’, ‘radiolarian’, ‘diatom’, ‘chert’,

‘radiolarite’, ‘biosiliceous’, ‘coquina’, and ‘diatomaceous’) . This resulted in dataframes for calcareous and siliceous sediments and excluded those sediments like sandstone, clay, etc. There are 16458 calcareous points and 3410 siliceous points, which were then confined to the epochs plotted in Figure 4, resulting in 6037 calcareous points and 726 siliceous points. Many cores contain several different biogenic lithologies for a given time interval. For example, an east Pacific core (col id 4803) contains unit id 58539 (classified as a calcareous ooze, ranges from 9.95-9.97 Ma) and the unit id 57885 (classified as a siliceous ooze, ranges from 9.97-9.98 Ma) during the Miocene; both units are plotted in Figure 4. Sediments are represented by different colored points (solid navy markers represent calcareous sediments; orange-colored open circles represent siliceous sediments). These maps were constructed in pyGplates v.036 (Müller et al., 2018) using the Seton et al. (2012) plate rotation model. Notably, the points are not rotated to paleo-position. We are employing the modern base map, but have coded the mapping function to run with the Seton plate rotation model to facilitate future developments.

These types of maps will be improved in the future with the addition of more data, more refined chronologies, paleolongitude/paleolatitude rotation via GPlates, and additional considerations, such as paleowater depth and the position of the calcite compensation depth (CCD). Despite these limitations, these maps display trends that have been found in more synthetic analyses (e.g., Lyle, 2003, Wade et al., 2020). For example, an abundance of siliceous sediments in the equatorial Pacific during the late Eocene (Fig. 4A) is also clearly apparent in the Wade et al. (2020) map from this time. There are areas and intervals with substantially higher sampling (e.g., the equatorial Pacific Ocean during the Miocene, Fig. 4C) that also are apparent in the datasets of Lyle (2003) and Wade et al. (2020). The eODP dataset records a change in the proportion of calcareous versus siliceous sediments during the Cenozoic; during the Eocene ~88% of the sediments are calcareous (1835 calcareous points to 256 siliceous points; Fig. 4A), while during the Oligocene it is ~94% (795 calcareous points to 256 siliceous points; Fig. 4B), and ~89% in the Miocene (3407 calcareous points to 417 siliceous points; Fig. 4C). This matches the expectation from the literature, with the ~1 km deepening of the CCD across the Eocene Oligocene boundary resulting in a greater abundance of calcareous sediments in the Oligocene relative to adjacent time intervals (Coxall et al., 2005; Pälike et al., 2012). eODP represents a step-change forward for the scientific ocean drilling community, one where investigations of these sorts can be readily done without painstaking, long hours generating new datasets. It is the hope of the eODP project that by following FAIR principles, these sorts of investigations can be facilitated much more readily.

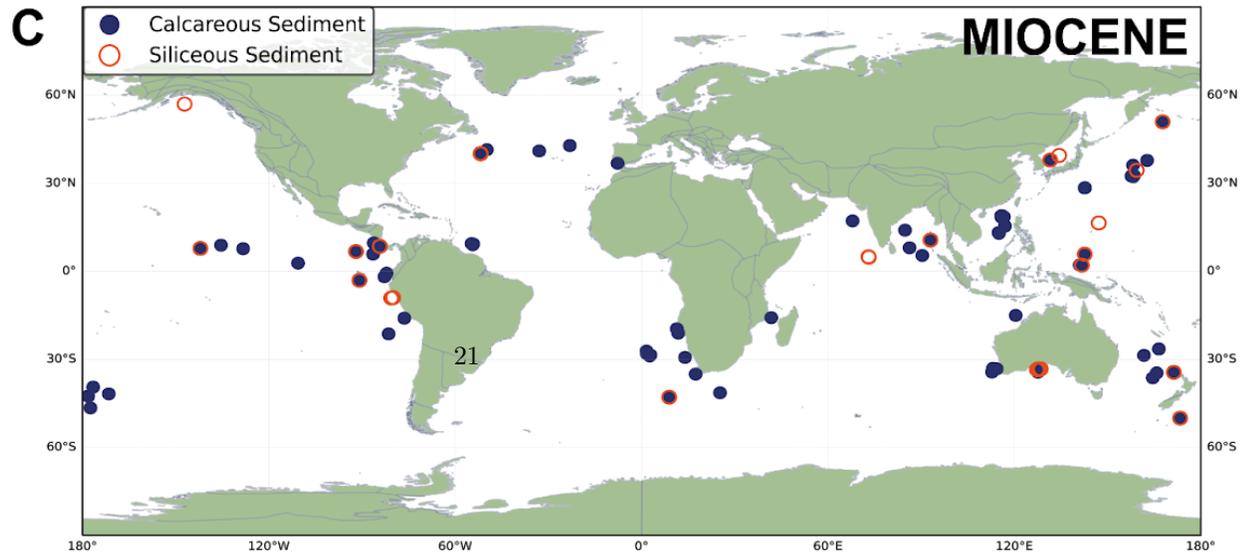
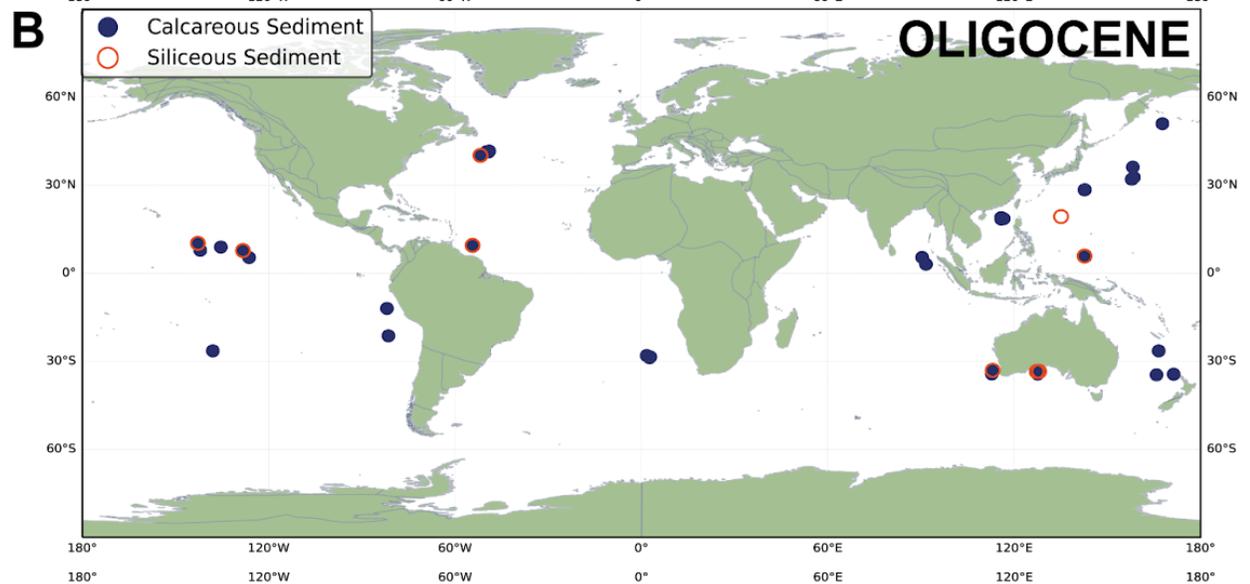
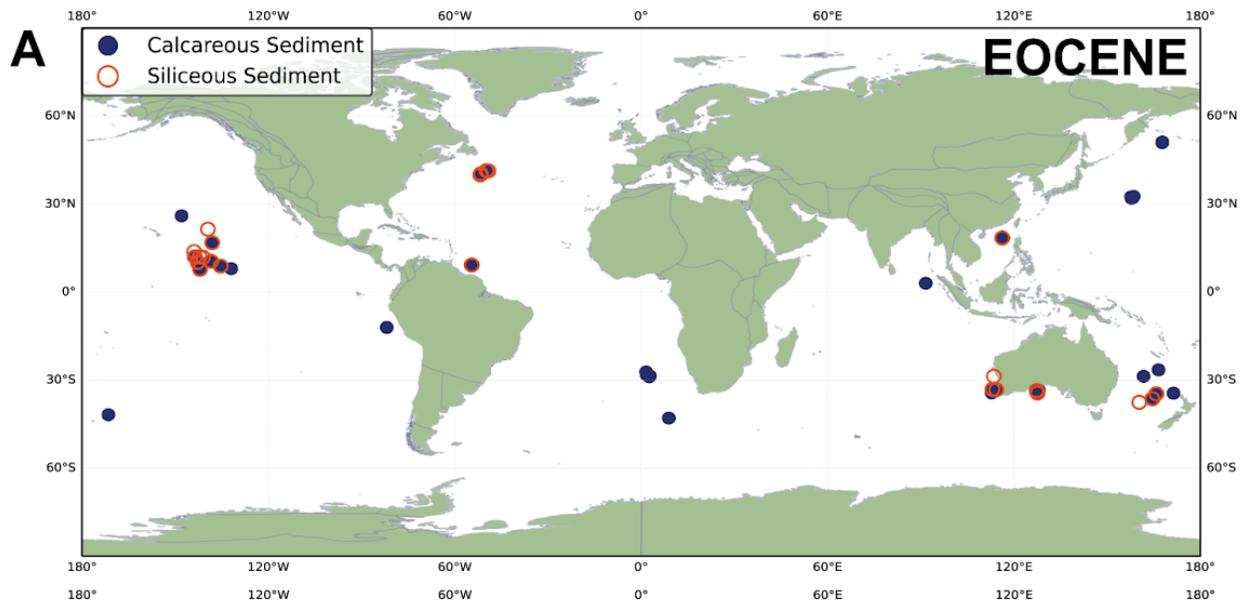


Figure 4: Maps of calcareous (blue, closed circle) and siliceous sediments (orange, open circle) presented at modern latitude/longitude. Maps were made in pyGPlates v.036 (Müller et al., 2018) using the global plate motions model and reconstructed coastlines from Seton et al. (2012). **A)** Biogenic sediment distributions from 56 to 33 Ma (Eocene); **B)** Biogenic sediment distributions from 33 to 23 Ma (Oligocene); **C)** Biogenic sediment distributions from 23 to 5 Ma (Miocene).

4 eODP community engagement:

The intention of eODP from the outset was to not only focus on data curation, but also to activate the SOD community to work on this material in a holistic way, co-designing tools in order to do so. In December 2018, we convened an EarthRates-funded workshop “Bringing Micropaleontology to the Paleobiology Database” (Workshop Report here: <https://github.com/eODP/eODP-2018-EarthRates-Workshop>), where the ~20 participants brainstormed on how to do this. The group also generated a datatype hierarchy for microfossil occurrence data, with priority-levels (required, desired, optional) for various tasks; included within the above Github repository.

The eODP project has been introduced at several major conferences and seminars including the American Geophysical Union Annual Meeting in 2019 (LeVay et al., 2019), the European Geophysical Union conference in 2020 (Fraass et al., 2020a), the EarthCube Annual Meeting in 2020 and 2022 (LeVay et al., 2020; 2022), the Geological Society of America Annual Meeting in 2020 (Fraass et al., 2020b), PaleoPERCS (Fraass, 2021), and the Geological Association of Canada - Mineralogical Association of Canada Conference in 2022 (Fraass et al., 2022a). In 2022, eODP hosted two virtual workshops on SOD and Earth Science databases, called Coding the Column, which engaged ~75 scientists in total. Several abstracts have been submitted for the upcoming 2022 Geological Society of America Annual Meeting (Fraass et al., 2022b; Jamson et al., 2022b; Sessa et al., 2022a) and American Geophysical Union Annual Meeting (Jamson et al., 2022c; Sessa et al., 2022b). The eODP project has funding for several in-person workshops, both stand alone and in conjunction with annual conferences, and it is our hope and plan to resume these activities, in-person, during late 2022 and 2023.

5 Future directions and conclusions

This is the first paper in an anticipated series; subsequent works will describe additional steps (e.g., improving genus- and species-level taxonomies, more complex and complete age-depth relationships) and address research questions that are only possible with eODP datasets. Offshore environments which are stable and continuous on million-year timescales and contain both the best-resolved fossil record and high-resolution paleoclimate records have the potential to allow understanding of the coupled ocean-climate-biosphere system at a deeper level than previously possible. It is the hope of the eODP team that these questions can be tackled both by the eODP team as well as a large community of other

scientists employing this dataset.

The status of the eODP project and conclusions can be summarized thusly:

- Using existing databases, instead of building from scratch, has several advantages. Both the PBDB and Macrostrat have existing connections to one another, either meet the needs of the SOD research community or can be adapted to them, have demonstrated user bases, and sustainability plans. Importantly, they both have existing Application Programming Interfaces (APIs) that are geared toward useful search parameters (age, lithology, taxonomic hierarchy, etc.), have existing code bases cultivated by the community, and follow FAIR principles.
- The PBDB does not currently reflect state-of-the-art taxonomy for many microfossil taxa. A planned step is to import the preexisting IODP Synonymy tables (circa 2010) developed by the Science and Technology Panel (STP) of the Integrated Ocean Drilling Program into the PBDB, alongside continued data entry work from the University of Victoria team.
- The management of age-depth relationships is complicated by the extensive requirements for generating marine stratigraphic records of high quality. eODP plans to continue developing tools as we discuss with the SOD community the best ways in which to curate these data.
- Facilitating workshops within the pandemic-era can be challenging, but despite the hurdles of virtual meetings, the SOD community remains eager to be a part of developing SOD data-resources, as seen in the participation and reception to the first EarthRates-funded workshop. We anticipate even greater success as in-person activity restarts and eODP is able to hold workshops.
- Modern SOD data has been available for more than half a century; however, it is not easily findable and interoperable, making it extremely difficult to use. The eODP project has, and continues to, devote significant effort to cleaning and harmonizing open data. This time investment highlights the difference between open data and FAIR data. If the community sees benefit in SOD data being readily usable, working towards standardizing data collection would be prudent. eODP represents, hopefully, a step toward a new era for scientific ocean drilling, with legacy data used for broader and deeper questions than before.

Acknowledgments

We are grateful for the taxonomic and age data entry efforts of: Andrew McCoy, Wunn Noon Naw, Phoebe O'Brien, Kelly Rozanitis, Chelsea Smith, and Alexis Srogota. Sidney Dangtran, Sydney Gutierrez-Gomez, and Kaylee Umberhocker are responsible for making "dark" lithologic descriptions discoverable. Morgan Underwood extracted OCR from core description pdf's and assisted the Whirl-i-Gig lead development team. The taxonomic experts in Table 2 are sincerely thanked for their assistance. The development efforts by Seth Kaufman and

Wai-Yin Kwan made this publication possible. Mark D. Uhen provided valuable discussion and guidance in the early phases of this project. Ilya Shunko, Graphic Design Consultant at Science Gateways Community Institute (SGCI; Lawrence et al., 2015), is sincerely thanked for creating the eODP logo. The eODP project acknowledges funding from the US National Science Foundation for its development (awards ICER 1927866 to L.J.L.; ICER 1928323 to S.E.P. and ICER 1928362 to A.J.F. and J.A.S) and for continued support of the International Ocean Discovery Program (IODP) program (award OCE 1326927), which provides the salary support for L.J.L. We are extremely thankful to David Lazarus and Johan Renaudie for age models and fruitful collaborative discussions. We acknowledge the longstanding efforts of many in the SOD community, too numerous to mention, in advocating for SOD data to be united and on easily accessible platforms. With respect to guiding eODP, we would like to thank Brian Huber, David Lazarus, and Ellen Thomas. This is PBDB publication ###.

Open Research

The standardized scientific ocean drilling lithology and micropaleontology datasets that compose the eODP Baggage Stripper database through the eODP GitHub [<https://github.com/eODP>]. All of the raw data can be found at [https://github.com/eODP/data-processing/tree/master/raw_data] and the processed, standardized files can be found here: [<https://github.com/eODP/data-processing/tree/master/output>]. The ‘data-processing’ repository in this GitHub contains the scripts used to standardize .csv files. All of eODP’s GitHub repositories are public.

The microfossil taxonomic entries were manually entered into the Paleobiology Database [paleobiodb.org]. The Paleobiology Database is open access and does not require a registration to view data or taxonomy. In addition to search functions through the main webpage, the Paleobiology Database can be accessed via API [<https://paleobiodb.org/#/resources>]. The Paleobiology Database is currently supported by NSF EAR 1948831.

All of the lithology data discussed and associated age constraints are stored in the Macrostrat database [macrostrat.org]. These datasets can be accessed via the Macrostrat API [<https://macrostrat.org/api>] and does not require registration. The eODP-specific datasets are flagged as a part of “Project 3” in Macrostrat e.g., [https://macrostrat.org/api/units?project_id=3&response=long&status_code=in%20process&format=csv]. Macrostrat is currently supported by NSF EAR-1150082 and ICER-1440312.

References

- Alroy, J. (2010). Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. *Palaeontology*, 53(6), 1211-1235.
- Alroy, J. (2010). The Shifting Balance of Diversity Among Major Marine Animal Groups. *Science*, 329(5996), 1191-1194. <http://www.sciencemag.org/content/329/5996/1191.abstract>
- Alroy, J., Aberhan, M., Bottjer, D. J., Foote, M., Fürsich, F. T., Harries, P. J., et al. (2008). Phanerozoic Trends in the Global Diversity of Marine Invertebrates.

Science, 321(5885), 97-100. <http://www.sciencemag.org/content/321/5885/97.abstract>

Alroy, J., Marshall, C. R., Bambach, R. K., Bezusko, K., Foote, M., Fursich, F. T., et al. (2001). Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences*, 98(11), 6261-6266. <Go to ISI>://000168883700059

Bowen, G. J., Clyde, W. C., Koch, P. L., Ting, S., Alroy, J., Tsubamoto, T., et al. (2002). Mammalian Dispersal at the Paleocene/Eocene Boundary. *Science*, 295(5562), 2062-2065. <https://doi.org/10.1126/science.1068700>

Bown, P. R. (2005). Calcareous nannoplankton evolution: a tale of two oceans. *Micropaleontology*, 51(4), 299-308. <https://doi.org/10.2113/gsmicropal.51.4.299>

Bown, P. R., Lees, J. A., & Young, J. R. (2004). Calcareous nannoplankton evolution and diversity through time. In H. R. Thierstein & J. R. Young (Eds.), *Coccolithophores: From Molecular Processes to Global Impact* (pp. 481-508). Berlin, Heidelberg: Springer Berlin Heidelberg.

Cohen, K., Finney, S., Gibbard, P., & Fan, J. (2013). The ICS International Chronostratigraphic Chart. *Episodes*, 36, 199-204.

Dunkley Jones, T., Lunt, D. J., Schmidt, D. N., Ridgwell, A., Sluijs, A., Valdes, P. J., & Maslin, M. (2013). Climate model and proxy data constraints on ocean warming across the Paleocene–Eocene Thermal Maximum. *Earth-Science Reviews*, 125, 123-145. <https://www.sciencedirect.com/science/article/pii/S0012825213001207>

Ezard, T. H. G., & Purvis, A. (2016). Environmental changes define ecological limits to species richness and reveal the mode of macroevolutionary competition. *Ecology Letters*, 19(8), 899-906. <https://doi.org/10.1111/ele.12626>

Fenton, I. S., Pearson, P. N., Dunkley Jones, T., Farnsworth, A., Lunt, D. J., Markwick, P., & Purvis, A. (2016). The impact of Cenozoic cooling on assemblage diversity in planktonic foraminifera. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1691), 20150224. <https://doi.org/10.1098/rstb.2015.0224>

Fenton, I. S., Woodhouse, A., Aze, T., Lazarus, D., Renaudie, J., Dunhill, A. M., et al. (2021). Triton, a new species-level database of Cenozoic planktonic foraminiferal occurrences. *Scientific Data*, 8(1), 160. <https://doi.org/10.1038/s41597-021-00942-7>

Fraass, A., LeVay, L., Sessa, J. A., & Peters, S. (2020, 2020/05/1). *Extending Ocean Drilling Pursuits [eODP]: Making Scientific Ocean Drilling Data Accessible Through Searchable Databases*. Paper presented at the EarthCube Annual Meeting, San Diego, CA.

Fraass, A., LeVay, L., Sessa, J. A., & Peters, S. (2020). Adapting existing database structures to work with scientific ocean drilling data. *Geological Society of America Fall Meeting 2020, Online*.

- Fraass, A. J. (2021). Towards a more holistic understanding of microfossil records and evolution. *PaleoPERCS, International online seminar series*. https://www.youtube.com/watch?v=BT3SGkSsd40&ab_channel=PaleoPERCS
- Fraass, A. J., Kelly, D. C., & Peters, S. E. (2015). Macroevolutionary History of the Planktic Foraminifera. *Annual Review of Earth and Planetary Sciences*, 43(1), 139-166. <http://dx.doi.org/10.1146/annurev-earth-060614-105059>
- Fraass, A. J., LeVay, L., Sessa, J. A., Peters, S., Kaufman, S., Kwan, W.-Y., & Jamson, K. (2022). eODP: adapting existing database structures to work with scientific ocean drilling data. *Geological Association of Canada-Mineralogical Association of Canada Joint Annual Meeting 2022, Online*.
- Fraass, A. J., Levay, L., Sessa, J. A., Peters, S., Kaufman, S., Kwan, W.-Y., et al. (2020). *eODP: adapting existing database structures to work with scientific ocean drilling data*. (Vol. 52).
- Fraass, A. J., LeVay, L. J., Sessa, J. A., Peters, S., & Jamson, K. J. (2022). The extending Ocean Drilling Pursuits (eODP) project: Stratigraphic through biotic trends. *2022 Geological Society of America Annual Meeting; Abstracts. Denver, Colorado*, 54.
- Greene, C., & Thirumalai, K. (2019). It's Time to Shift Emphasis Away from Code Sharing. *Eos Transactions American Geophysical Union*, 100, 16-17.
- Husson, J. M., & Peters, S. E. (2017). Atmospheric oxygenation driven by unsteady growth of the continental sedimentary reservoir. *Earth and Planetary Science Letters*, 460, 68-75. <https://www.sciencedirect.com/science/article/pii/S0012821X16307129>
- Ivany, L. C., Pietsch, C., Handley, J. C., Lockwood, R., Allmon, W. D., & Sessa, J. A. (2018). Little lasting impact of the Paleocene-Eocene Thermal Maximum on shallow marine molluscan faunas. *Science Advances*, 4(9). 10.1126/sciadv.aat5528.
- Jamson, K. M., Moon, B. C., & Fraass, A. J. (2022a). Diversity dynamics of microfossils from the Cretaceous to the Neogene show mixed responses to events. *Palaeontology*, 65(4), e12615. <https://onlinelibrary.wiley.com/doi/abs/10.1111/pala.12615>
- Jamson, K. M., Fraass, A. J., Sessa, J. A., LeVay, L. J., & Peters, S. E. (2022b). The extending Ocean Drilling Pursuits (eODP) project: Spatial distribution of biogenic sediments from the Cretaceous to the Recent. *GSA Annual Meeting; Abstracts, v. 54*.
- Jamson, K. M., Sessa, J. A., Fraass, A. J., LeVay, L. J., & Peters, S. E. (2022c). The extending Ocean Drilling Pursuits (eODP) project: Spatial distribution of biogenic sediments through the Cenozoic. *Fall meeting AGU; Chicago, Illinois*.
- 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(10), 1570-1596. <https://doi.org/10.1029/2019PA003632>.
- Kiessling, W., & Kocsis, Á. T. (2016). Adding fossil occupancy trajectories to the assessment of modern extinction risk. *Biology Letters*, 12(10), 20150813. <https://doi.org/10.1098/rsbl.2015.0813>
- Kwan, W.-Y., Kaufman, S., LeVay, L. J., Fraas, A. J., Peters, S. E., & Sessa, J. A. (2022). Creating a reproducible data standardization workflow using Jupyter Notebooks. . *EarthCube Annual Meeting. Notebook Proceedings Contribution 134*.
- Lawrence, K. A., Zentner, M., Wilkins-Diehr, N., Wernert, J. A., Pierce, M., Marru, S., & Michael, S. (2015). Science gateways today and tomorrow: positive perspectives of nearly 5000 members of the research community. *Concurrency and Computation: Practice and Experience*, 27(16), 4252-4268. <https://doi.org/10.1002/cpe.3526>. <https://doi.org/10.1002/cpe.3526>
- Lazarus, D. (1994). Neptune: A marine micropaleontology database. *Mathematical Geology*, 26(7), 817-832. <https://doi.org/10.1007/BF02083119>
- LeVay, L., Fraass, A., Sessa, J., & Peters, S. (2020). Extending ocean drilling pursuits (eODP): Making scientific ocean drilling data accessible through searchable databases. *2020 EarthCube Annual Meeting*. <https://doi.org/10.1002/essoar.10503496.1>
- LeVay, L. J., Fraas, A. J., Peters, S. E., Sessa, J. A., Kaufman, S., & Kwan, W.-Y. (2022). Geologic data standardization for database entry: preparing diverse datasets for hosting and accessibility. *EarthCube Annual Meeting. La Jolla, CA*.
- LeVay, L. J., Fraass, A. J., Sessa, J. A., & Peters, S. E. (2019). *Extending Ocean Drilling Pursuits [eODP]: Making Scientific Ocean Drilling Data Accessible Through Searchable Databases*.
- Lowery, C. M., Bown, P. R., Fraass, A. J., & Hull, P. M. (2020). Ecological Response of Plankton to Environmental Change: Thresholds for Extinction. *Annual Review of Earth and Planetary Sciences*, 48(1), 403-429. <https://doi.org/10.1146/annurev-earth-081619-052818>
- Lyle, M. (2003). Neogene carbonate burial in the Pacific Ocean. *Paleoceanography*, 18(3). <https://doi.org/10.1029/2002PA000777>. <https://doi.org/10.1029/2002PA000777>
- McKay, N. P., Emile-Geay, J., & Khider, D. (2021). geoChronR – an R package to model, analyze, and visualize age-uncertain data. *Geochronology*, 3(1), 149-169. <https://gchron.copernicus.org/articles/3/149/2021/>
- Müller, R. D., Cannon, J., Qin, X., Watson, R. J., Gurnis, M., Williams, S., et al. (2018). GPlates: Building a Virtual Earth Through Deep Time. *Geochemistry, Geophysics, Geosystems*, 19(7), 2243-2261. <https://doi.org/10.1029/2018GC007584>.

- Müller, R. D., Mather, B., Dutkiewicz, A., Keller, T., Merdith, A., Gonzalez, C. M., et al. (2022). Evolution of Earth's tectonic carbon conveyor belt. *Nature*, *605*(7911), 629-639. <https://doi.org/10.1038/s41586-022-04420-x>
- OConnell, S. (2019). Holes in the Bottom of the Sea: History, Revolutions, and Future Opportunities. *GSA Today*, *29*.
- Pälike, H., Lyle, M. W., Nishi, H., Raffi, I., Ridgwell, A., Gamage, K., et al. (2012). A Cenozoic record of the equatorial Pacific carbonate compensation depth. *Nature*, *488*(7413), 609-614. <https://doi.org/10.1038/nature11360>
- Peters, S. E. (2005). Geologic constraints on the macroevolutionary history of marine animals. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(35), 12326-12331. <Go to ISI>://000231675900010
- Peters, Shanan E. (2006). Macrostratigraphy of North America. *The Journal of Geology*, *114*(4), 391-412. <https://doi.org/10.1086/504176>
- Peters, S. E. (2008). Environmental determinants of extinction selectivity in the fossil record. *Nature*, *454*(7204), 626-629. <https://doi.org/10.1038/nature07032>
- Peters, S. E., & Husson, J. M. (2017). Sediment cycling on continental and oceanic crust. *Geology*, *45*(4), 323. 10.1130/G38861.1. <http://geology.gsapubs.org/content/45/4/323.abstract>
- Peters, S. E., Husson, J. M., & Czaplewski, J. (2018). Macrostrat: A Platform for Geological Data Integration and Deep-Time Earth Crust Research. *Geochemistry, Geophysics, Geosystems*, *19*(4), 1393-1409. <https://doi.org/10.1029/2018GC007467>.
- Peters, S. E., Kelly, D. C., & Fraass, A. J. (2013). Oceanographic controls on the diversity and extinction of planktonic foraminifera. *Nature*, *493*(7432), 398-401. 10.1038/nature11815. <http://dx.doi.org/10.1038/nature11815>
- Peters, S. E., & McClennen, M. (2016). The Paleobiology Database application programming interface. *Paleobiology*, *42*(1), 1-7. <https://www.cambridge.org/core/article/paleobiology-database-application-programming-interface/4D20F5CAFA1B0AC7033975418668D82B>
- Peters, S. E., Quinn, D. P., Husson, J. M., & Gaines, R. R. (2022). Macrostratigraphy: Insights into Cyclic and Secular Evolution of the Earth-Life System. *Annual Review of Earth and Planetary Sciences*, *50*(1), 419-449. <https://doi.org/10.1146/annurev-earth-032320-081427>
- Raup, D. M. (1976). Species diversity in the Phanerozoic: an interpretation. *Paleobiology*, *2*, 289-297.
- Raup, D. M. (1979). Biases in the fossil record of species and genera. *Bulletin of Carnegie Museum of Natural History*(13), 85-91.
- Renaudie, J., Lazarus, D., & Diver, P. (2020). NSB (Neptune Sandbox Berlin): An expanded and improved database of marine planktonic microfossil data and deep-sea stratigraphy. *Palaeontologia Electronica*, *23*, a11.

- Sánchez, T. M. (2010). Emiliodontia, New Name for Emiliania Sánchez, 1999, Not Emiliania Hay and Mohlen, 1967. *Journal of Paleontology*, 84(4), 781-781. <https://doi.org/10.1666/10-023.1>
- Sepkoski, J. J., Jr. (1981). A factor analytic description of the Phanerozoic marine fossil record. *Paleobiology*, 7(1), 36-53.
- Sepkoski, J. J., Jr. (1993). Ten Years in the Library: New Data Confirm Paleontological Patterns. *Paleobiology*, 19(1), 43-51.
- Sessa, J. A., Fraass, A. J., LeVay, L. J., Peters, S. E., & Jamson, K. M. (2022a). The extending Ocean Drilling Pursuits (eODP) project: Synthesizing marine stratigraphic, chronostratigraphic, and micropaleontologic data. *GSA Annual Meeting; Abstracts*, 54.
- Sessa, J. A., Fraass, A. J., LeVay, L. J., Peters, S. E., & Jamson, K. M. (2022b). The extending Ocean Drilling Pursuits (eODP) project: Synthesizing marine stratigraphic, chronostratigraphic, and micropaleontologic data. *Fall meeting AGU; Chicago*.
- Seton, M., Müller, R. D., Zahirovic, S., Gaina, C., Torsvik, T., Shephard, G., et al. (2012). Global continental and ocean basin reconstructions since 200Ma. *Earth-Science Reviews*, 113(3), 212-270. <https://www.sciencedirect.com/science/article/pii/S0012825212000311>
- Spencer-Cervato, C., Thierstein, H. R., Lazarus, D. B., & Beckmann, J.-P. (1994). How synchronous are neogene marine plankton events? *Paleoceanography*, 9(5), 739-763. <https://doi.org/10.1029/94PA01456>
- Trubovitz, S., Lazarus, D., Renaudie, J., & Noble, P. J. (2020). Marine plankton show threshold extinction response to Neogene climate change. *Nature Communications*, 11(1), 5069. <https://doi.org/10.1038/s41467-020-18879-7>
- Uhen, M. D., & Anthony D. Barnosky, B. B., Jessica Bloisd, Matthew T. Carrano, Marc A. Carrascob, Gregory M. Erickson, Jussi T. Eronengh, Mikael Fortelius, Russell W. Grahmi, Eric C. Grimmj, Maureen A. O'Leary, Austin Mastfl, William H. Pielm, P. David Pollyn & Laura K. Säiläg (2013). From card catalogs to computers: databases in vertebrate paleontology. *Journal of Vertebrate Paleontology*, 33, 13-28.
- Villier, L., & Korn, D. (2004). Morphological Disparity of Ammonoids and the Mark of Permian Mass Extinctions. *Science*, 306(5694), 264-266. <https://doi.org/10.1126/science.1102127>
- Wade, B. S., O'Neill, J. F., Phujareanchaiwon, C., Ali, I., Lyle, M., & Witkowski, J. (2020). Evolution of deep-sea sediments across the Paleocene-Eocene and Eocene-Oligocene boundaries. *Earth-Science Reviews*, 211, 103403. <https://www.sciencedirect.com/science/article/pii/S0012825220304499>
- 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(1), 160018. <https://doi.org/10.1038/sdata.2016.18>

Zachos, J., Pagani, M., Sloan, L., Thomas, E., & Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science*, 292(5517), 686-693.

Zachos, J. C., Dickens, G. R., & Zeebe, R. E. (2008). An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature*, 451(17), 279-283.

Taxonomic Workflow supplement (also available at <https://github.com/eODP/files-for-Sessa-2022>):

Within the BS, we created a ‘Non-taxonomic modifier’ column for taxonomically-relevant morphological and size information that is not a subspecies or standard modifier but is information that is more important than going into the ‘Comment’ field. For example, nannoplankton species will have different size bins that are relevant for identification and specific morphologic features like the number of rays on a Discoaster. For planktic forams, pink and white G. ruber; dextral and sinistral; intergrades of species like “Paragloborotalia continuosa/mayeri”. Benthic forams, Nodosaria spp. ‘elongate forms’. Radiolarians, intergrades like Botryostrobus auritus-australis group.

Some informal names were kept, like ‘Phytolith’ and ‘spore’ - as these have valid meanings and coarsening up to a formal taxonomic name would become less useful to meaningless (ex., phytolith would be coarsened up to “Plantae” and spore to “Life”). These informal names were placed into quotes.

Lists that were mislabeled (ex., nannofossil taxa in ‘diatom’ files) were correctly assigned.

The placement of taxonomic modifiers were standardized to always precede the name; ‘?’ was sometimes attached to the end of a name. Similarly, ‘sp.’ needed to be added to many generic questionable names (ex., ‘Arenobulimina ?’ becomes ‘? Arenobulimina sp.’). The single entry dextral and sinistral foram ratios into two entries, one for each of the sinistral and dextral counts.

Homonyms were disambiguated within the PBDB; ex., *Trinacria* is both a diatom and a bivalve; *Helminthopsis* is both an ichnofossil and diatom, *Multispinula* is both a brachiopod and a dinoflagellate, *Helicolithus* is both an ichnofossil and a calcareous nannoplankton. Note that the PBDB has a specific function for keeping homonyms separate.

Unique modifiers

?

aff.

cf.

f. (abbreviation for forma)

morph

s.s.

s.l.

var.

Instructions for PBDB entry

Start becoming familiar with the PBDB entry features by going to:

<https://paleobiodb.org/#/resources> and scroll to:

"Data Entry Tutorials" - I think the best order to watch these in is:

"Paleobiology Database Webinar 2: Entering Data" then

"Enter new reference" then

"Enter new taxon" and

"Enter new taxonomic opinions"

The videos about entering a reference, new taxon, and new taxonomic opinion are the most relevant; the "Entering Data" video covers all the PBDB features, so it contains much more than needed for just taxonomic entry, but is a good place to start for a general overview of the PBDB.

The first video in this series (webinar 1) is about downloading data, which is helpful in pulling taxonomy for various groups and to create stats on number of refs, names, and opinions entered <https://www.youtube.com/watch?v=c26nKFjbH38>.

Process for entering; example here is the calcareous nannoplankton; process was followed for all the taxa we e

Calcareous Nannofossils (mostly Coccolithophorids, but other groups, too) by Leah

Step 1:

We want to start out by completing the higher level taxonomy in the PBDB. For this I am going to refer to algaebase.org:

<https://www.algaebase.org/browse/taxonomy/?id=4359>

Kingdom through Class will be the same for all species. To see the full reference for the name, click on the Authority.

The PBDB needs to have the Kingdom (Chromista) and Class (Coccolithophyceae) updated.

Step 2:

Algaebase contains some of the orders. Use this site for what is present to the order level. Not all of the orders have an authority listed, but we will cover this below. Ignore the incertae sedis for now, we will enter those last.

Step 3:

For all other orders, orders with an authority, and lower level taxonomy, we will move to the Nannotax database:

Cenozoic and Extant: <http://www.mikrotax.org/Nannotax3/index.php?dir=Coccolithophores>

Mesozoic: <http://www.mikrotax.org/Nannotax3/index.php?dir=Mesozoic>

The linked pages list the Orders. Add any order that is either not included in the Algaebase or does not have an Authority listed in Algaebase.

To find the citation, you will need to click on the order (example: Isochrysidales). This will now list 3 families. Below the list of families you will see there is a citation (Pascher 1910).

To get to the full reference for Pascher 1910, go to “Tools” listed along the menu bar and select “References”. You can search for Pascher and get the full citation.

You can now do the same steps for the Family level.

Step 4:

After completing the families, we should download the taxonomy from the PBDB and make sure that attributions are correct (i.e. family is linked to order and not to phylum).

Step 5: Move onto genera