NSB (Neptune Sandbox Berlin):
An expanded and improved database of
marine planktonic microfossil data and deep-sea stratigraphy

Johan Renaudie, David B. Lazarus, and Patrick Diver

ABSTRACT

Thirty years ago, the Neptune Database was created to synthesize microfossil occurrences from the deep-sea drilling record. It has been used in numerous studies by both biologists and paleontologists of the evolution and distribution in space and time of marine microplankton. After decades of discontinuous development in various institutions, a significant overhaul of the system was made during the last decade, leading to the database being expanded and made available again online under the name Neptune Sandbox Berlin (NSB). In particular, the addition of an extensive stratigraphic layer has resulted in the database now being used by a wider geoscience community than its previous incarnations. We summarize here the complete history of the Neptune database development and provide a complete description of the new system, NSB, its associated website and age-model making software (NSB_ADP_wx).

Johan Renaudie. Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany. johan.renaudie@mfn.berlin
David B. Lazarus. Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany. david.lazarus@mfn.berlin
Patrick Diver. DivDat Consulting, Wesley, Arkansas, USA. pldiver@gmail.com

INTRODUCTION

The NSB database (for Neptune Sandbox Berlin) is the current implementation of the previous Neptune marine microfossil database, and as the name suggests, is developed and hosted by the Museum für Naturkunde in Berlin. It holds ca 1 million records of species occurrences, taxonomy, and geologic age information. The database has been used nearly continuously over a period of more than 25 years by a worldwide community of micropaleontologists, paleoceanographers, theoretical ecologists, and others; and has been cited in more than 110 publications. Despite the impor-

tance of the system, it has until now been documented only in much earlier, in significant ways incomplete, and in any case now seriously outdated papers (Lazarus, 1994; Spencer-Cervato, 1999). A more complete description of both the history of the project and the modern system, as provided in this paper, are thus long overdue. Specifically we describe the history of how this database developed and how the historical context affected the design of the system, how it is currently built and functions, and how it may grow into the future, along with estimates of data quality and other issues important to potential users.

HISTORY OF DEVELOPMENT

Origin, Initial Development, Goals and Design Concept (Zürich, 1989-1995)

Paleontology in the 1980s was in lively turmoil as major themes in evolutionary biology were under active paleontologic debate. Sepkoskis compilation of Phanerozoic marine families (Sepkoski, 1976 and subsequent papers) had triggered debates, which continue to this day, on the basic patterns of past biodiversity and the mechanisms that cause them. It was also clear that Sepkoski's simple first and last occurrence information for high level taxa was not sufficient to answer many questions, and thus much more detailed data on the occurrences of organisms over time would be needed. The paleontologic community eventually responded to this with the development, starting in 1998, of the Paleobiology Database (PBDB: http://fossilworks.org/?page=FAQ; Alroy et al., 2001). In principle, however, the rich record of (mostly Cenozoic aged) planktonic protist species diversity that had been, by the late 1980s, documented by nearly twodecades of deep-sea sediment coring could provide even better data than the fossil record of macro organisms. However, the occurrence data was scattered through hundreds of publications and although in part captured digitally by the Deep Sea Drilling Program (at the time, the formal name of the international deep sea drilling effort) was not easily accessible by working scientists for large-scale synthesis. Further, taxonomic concepts and cataloging of the species had lagged behind the descriptive work on hundreds, if not thousands, of new species that these cores had recovered, so that workers struggled with the extensive (paper only) literature simply to standardize concepts for the relatively small fraction of observed species that had been selected as biostratigraphic markers. Lastly, due to long standing

divisions between micropaleontology and the rest of paleontology, and possibly due to the largely dismissive reception of then recent micropaleontologic work on the theory of punctuated equilibrium, most micropaleontologists were not interested in pursuing paleobiologic questions, preferring instead to concentrate on (relatively easily funded) applied geologic research in biostratigraphy and the newly developing field of paleoceanography, neither of which needed such large syntheses of available microfossil data.

Despite these challenges, in the late 1980s a group of micropaleontologists under the leadership of one of the junior authors of this paper (Lazarus) began to develop a new database of published marine microfossil occurrence data based on the materials recovered by the deep-sea drilling programs. The initial goals (Lazarus, 1994) were to provide the ability to carry out broad syntheses of marine microfossil data to better understand broader patterns of evolution in the context of paleoceanographic change, as part of the newly developing subject of earth system science research. Other goals included assisting the declining number of taxonomic specialists by providing more efficient ways to find and synthesize taxonomic literature to improve marine microfossil taxonomy (Lazarus, 1994). The project was based within the micropaleontology group of H. Thierstein at the ETH Zürich, who was official manager of the Swiss National Science Funding agency (SNF) grant and active participant in the project's development.

The central design elements of the original NSB system were derived from the data, and methods of analysis already well established in marine micropaleontologic research. Occurrence data was assumed to be for lists of species in a series of sediment samples taken from (drilled) deep-sea sediment sections. These primary observations needed only to be mapped to occurrences by geologic age via the use of an age model giving geologic age vs depth in section, and the taxonomy updated as needed to uniform species concepts. Only basic information about the sections and sample locations in these needed to be recorded, as all other section-related data (lithology, geochemistry, etc.) were available in other deep-sea drilling program supported data archiving systems. Initially, a large fraction of the data was already available digitally from the drilling programs, primarily in Fortran card image format files (later published on CD-ROM: Deep Sea Drilling Project, 1989). In 1993 data access was improved via the creation of ODP's Janus online relational database

system (Mithal and Becker, 2006). Thus a relatively simple data structure would be sufficient to provide data on the occurrences of species in time and space.

Inherent in the design was the focus on species as the primary taxonomic unit of data rather than a hierarchy of higher units such as genera or families, which were the primary units of Sepkoski's database, as well as the succeeding PBDB. The focus on species reflects the desire in deep-sea sediment research to resolve the geologic age as finely as possible in support of high temporal resolution studies of oceanographic change, and is in turn derived from earlier work on developing high resolution, species level stratigraphy of geologic sections for oil exploration (Bolli et al., 1985). Species were also the only taxonomic level which could readily be identified, as the biologic-functional and genetic information needed for higher level systematics of these mostly protist taxa was lacking (and is only recently starting to become available).

It should be noted here that Neptune shares a geologic section oriented structure with some other systems such as the Geobiodiversity Database (GBDB; Fan et al., 2013), but differs sharply in the individual sample ('collection') based structure used in PBDB. Neptune and PBDB also differ in the initial assumption that Neptune would be one of a complementary suite of data resources and would rely on other databases, e.g., the drilling program Janus database for lithologic information. Many other databases, in particular PBDB, need to hold much more detailed geologic data linked to the samples, as no complementary databases similar to those of the drilling programs are available. Another difference is the absence in deep-sea sediments of the complex nomenclature of classical lithology (formations, beds, etc.): samples in deep-sea sediment research are identified simply by the name of the section and a variously formatted sample label giving the stratigraphic depth location within the section.

In detail many other requirements arose that complicated the actual database design. For example, the practice in early (DSDP era) drilling was to label the samples provided to the scientists only with the (unique to deep-sea drilling programs) sample label hierarchy of core, section, and depth interval within section (in cm). The actual depth in the hole (e.g., in meters below sea-floor: mbsf) was not provided to the paleontologists, and thus normally did not appear in their published reports. In order to link these published data on species

occurrences to depth in section, and thus to the age model, the depth in section had to be calculated, using supplementary data tables for cores in Neptune that were obtained from the drilling programs. Neptune's initial implementation thus had a few tables more than initially planned, but less than one dozen, which for relational databases was indeed a simple design.

Technical constraints of 1980s computing also significantly affected early design decisions. One important decision was the need to keep the total size of the database as small as possible as even the workstation class computers at that time had RAM capacities in the low single digit Mb range (typically just 4 Mb), and large hard drives were only ca 100 Mb in size. This in particular affected the mapping of published occurrence data into the system. The normal practice in marine micropaleontology was, and still is, to report occurrence data in 2D matrices (typically a spreadsheet file) of sample names by taxa names, with a mixture of codes to represent occurrences. Blank cells are assumed to indicate absence of the species, even if in practice, they can also occasionally be due to the failure to record the presence of a species due to simple forgetfulness by the observing scientist. This however only would be likely to (occasionally) occur for species of no known biostratigraphic value, whose presence in samples would be recorded simply to indicate general assemblage composition, or for some other secondary purpose. Only rarely would a special code, such as '-' be used to indicate that the species was explicitly searched for but not found. Given the focus in most published studies on reporting biostratigraphic markers, and that these almost by definition do not extend over the majority of the studied section, the large majority of data cells in the published matrices were blank. The database took advantage of these characteristics of the data to include in the database only non blank cells, which reduced the size of the main occurrences data table by an order of magnitude (from 11 to just 1 Mb: Lazarus, 1994), thus permitting the primary data tables to reside entirely in RAM, and a consequently dramatic improvement in computation speed.

Other decisions included keeping the biostratigraphic, magnetostratigraphic and other event data used to create age models external to the database, as the database did not have any need to access these data-only the age model derived from them was used internally to calculate ages. Nor, unusually, was there a table to provide the publication source of the data. This was done as at

the time virtually all deep-sea drilled section micro-paleontology data was published in the DSDP Scientific Reports volumes, and there was a nearly perfect one paper per microfossil group per Leg pattern, so that the database fields 'fossil group' and 'Leg' values served to uniquely identify the source of the data in the system.

Neptune was intially programmed in what at the time was a very powerful, relational (although lacking a dynamic data definition capability), and uniquely graphical and thus for non programmers easy to use system: 4th Dimension software ('4D'; Ribardière, 1985). 'Mainframe' derived SQL systems (e.g. IBM products, Oracle) were still in the early phases of development, very expensive and difficult for non-programmers to use, while other (often not SQL-based) desktop databases of the time (e.g., dBase, Paradox) were relatively under-powered and/or tied to the increasingly obsolete DOS operating system. 4D was only available for Apple's Macintosh computers, which was conveniently the standard system used by the research group at the ETH. 4D included a full procedural programming language with a Pascal-like syntax, which also fit well with the ETH environment - Pascal was widely used in the 1980s, had been developed by N. Wirth at the ETH and was the programming language all ETH students were required to learn. Neptune was initially a single user application as a (local network only) multiple user (client-server) version of 4D was only released in 1993.

A species level only focus for taxonomic information meant that in principle a single data table could be used to hold taxonomic names. A further important design decision was to preserve the original taxonomic names given in the source publications, with name standardization being done via links between names. Preserving the original names rather than entering only standardized names added complexity to the database, but only by doing this is it possible to 'roll back' changes in the taxonomic meaning of the name that may occur over time as taxonomy improves. For example, if two taxa, initially published as X and Y are later thought to be synonymous and the Y is changed to X, it is not possible to restore just the X values that originally were Y back to Y if an even later study reverses the synonymization. Such problems are in fact quite realistic given knowledge in the last decades on pseudocyrptic spieces in protist plankton that reverse long standing trends towards 'lumping' finely resolved morphospecies together. Standardization included not only formal synonyms

but also a wide variety of other conceptual relationships between names such as questionable identifications, subspecies, and a variety of other, often informal qualifiers that are used by working scientists when they record occurrences in actual samples (e.g., cf, sensu latu, etc.). The need to accomodate these identification complexities in real-world occurrence data distinguishes the taxonomic list content of Neptune from a formal taxonomic catalog system, where such complexities are largely ignored.

Relational databases are not very well suited to representing such hierarchical links, and most implementations of hierarchy in relational systems either include a static set of tables, each for a pre-defined taxonomic rank; or represent the links as a type of pointer to another record, either in the same table or another table. Links referencing records within a single table is the form adopted by Neptune, and is a flexible and simple way to record a variety of relationships. The type of relationship was also recorded by a separate field (taxon status) which indicated via simple letter codes if the taxon was a valid name (and thus did not have a link value), a formal synonym, a questionable identification of another name, a non species level identification (typically genus spp.), or a name whose taxonomic status was not yet determined (U, for unknown). The only disadvantage of this type of representation is that the logic needed to compile the records indicated by the links cannot be expressed in the declarational relational database SQL language but instead requires a procedural language to implement, which was not by default included in early SQL database systems. This issue was not considered as a problem when Neptune was created as SQL was not part of 4D, while a suitable procedural language was.

The initial implementation of Neptune covered the main groups of planktonic microfossils found in deep-sea sediments-planktonic foraminifera, calcareous nannofossils (coccolithophores and the extinct, enigmatic group discoasters), planktonic diatoms, and radiolaria (i.e., Polycystinea). In the 1980s very little data on other groups had been published by the deep-sea drilling projects other than benthic foraminifera, but these were seen as problematic due to the complex, poorly understood taxonomy and were thus left for future development. Standardization of the names in Neptune - ca 8,000 records in the initial implementation (Lazarus, 1994) was done by taxonomic experts for each group. K. von Salis and J.P. Beckmann, both at the ETH, edited the name lists for the calcareous

nannofossils and planktonic foraminifera, respectively. Diatoms were edited by C. Sancetta, then at the NSF, USA, while Lazarus, C. Nigrini (independent consultant), and J. P. Caulet (MNHN, Paris) edited the radiolarian names. The completeness of these efforts varied according to time available and the state of taxonomic knowledge of the group, with the lists being reasonably complete for the calcareous groups but significantly less so for the siliceous microfossil clades. It should be noted here that even in the initial implementation it was clear that in principle, synonyms should be made on a per publication basis, not globally for a name (Lazarus, 1994). The effort required for this was seen then as prohibitive (and remains so), and thus was not done, nor were the data structures in Neptune designed to support such a fine-grained synonymy. Taxa names in Neptune were represented in the occurrences table by relatively short alphanumeric codes (e.g., RSTYA0090 for the radiolarian *Stylatractus universus*), both to save storage space and to maintain compatibility with DSDP, which had established this coding system for all taxonomic data entered into their data archives. Radworld, a taxonomic catalog database of radiolarians developed by Caulet and Nigrini, also adopted this code system.

Quantitative geologic age values for all occurrence data were a central element in Neptune's design. The established standard method for this in the field of deep-sea sediment studies was the creation of an age model for each drilled section, which via a set of linked straight line segments gave a continuous mapping of depth in the section to geologic age. Only the end points of the line segments were stored in Neptune and the upload data procedure used the stored age model to assign ages to each new sample as it was loaded into the system. No internal representation of the different geochronologic reference calibration systems was made, and all numeric values were based on the then current time scale (Berggren et al., 1985). Towards the end of the initial ETH development phase a global mapping of these ages was done to the newer Berggren et al. (1995) time scale. A separate data table in Neptune stored basic metadata about the age models stored in the system.

To develop the age models, a special purpose stand-alone program was developed - Age Depth Plot (ADP) (Lazarus, 1992), written in True Basic (Kemeny et al., 1985), a language which supported GUI app development on the Macintosh platform. The program read in event and/or zonal age estimates from a user provided tab-text file, a file (derived from the drilling program databases) giving the core depths for the section, converted the age data from DSDP sample label format to depth in section, plotted the data as a depth in section (Y, or vertical axis) vs geologic age (horizontal axis) scatterplot, with, as was convention in the drilling programs, the origin in the upper left. It also provided a dynamic line-drawing function, which allowed the user to draw in a 'line of correlation' (LOC) - a set of line segments with differing lengths and slopes that visually best fit the data, including an option for hiatuses: line segments with identical depth values but different age values, which plotted as horizontal lines. ADP imposed controls on the drawing of the segments to ensure a monotonic increasing relationship between depth and geologic age. The control points defining the line were exported and uploaded into the database. For documentation the image of the plot with the fitted LOC was also saved.

Deciding which holes (sections) from the hundreds that had been drilled to upload into Neptune took considerable time. Specialists for each microfossil group ranked holes according to overall importance for their group - quality and quantity of occurrence data and the age model for the section. The joint ranking for the various groups largely determined the priority for preparing and uploading occurrence data, although an effort was also made to maximise coverage by group, geologic age and geography, and to avoid uploading two very similar sections, as judged by these criteria. Manpower constraints and the tendency for geologically younger intervals to have better section geochronology and better geographic coverage led to the decision to load data initially only for the Neogene (ca 0-23 Ma).

Most data was extracted from the DSDP CD-ROM, but for the more recently drilled ODP sections, for which no digital occurrence data was then available, data was uploaded from spreadsheet files contributed by external colleagues, or in a few cases by re-typing the data from the publication into a spreadsheet file. A 4D procedure in Neptune uploaded the data matrix for a section, checking that all taxa in the file were also in the taxonomy table, and then loading the sample names into one table, with the (non blank) occurrences for each sample being loaded into the main occurrences table. For each chosen section, group members collated the available biostratigraphic and magnetostratigraphic data and used the ADP program to develop the age model for the section. All occurrence data and age models were checked for con-

sistency by Lazarus before uploading to the database. M. Pianka-Biolzi, H. Hilbrecht, and C. Spencer-Cervato joined the group at the ETH in this labor-intensive phase of data compilation and entry. By 1994, data from over 100 sections and nearly 300,000 occurrence records had been entered into the system.

Neptune contained analytic functions, the most important of which recovered all occurrences of desired taxa including data recorded under synonymous names. A graphical application (Age Range Chart, or ARC), written in True Basic, was co-written by Lazarus and U. Störlein to plot multiple taxa occurrences by geologic age in multiple sections. This was meant to aid taxonomic and biostratigraphic studies. Details of these functions are given in Lazarus (1994), which was the first significant documentation of the new system and was intended as the initial publication in a series to follow with increasingly analytic content. The extensive compilation of biostratigraphic and magnetostratigraphic data used to make the age models, together with the models and age depth plots for all DSDP sections, was published in Lazarus et al. (1995), with a copy of the content being archived at the NGDC in the USA (now part of NCEI). A subsequent volume for ODP sections was initially planned but not carried out. Two initial analytic studies based on the stratigraphic compilation and age model library were published (Spencer-Cervato et al., 1993, 1994), which documented extensive patterns of systematic diachroniety in what were assumed by many to be essentially isochronous plankton biostratigraphic events.

Interregnum (1996-2001)

In 1995 Lazarus left the ETH for a permanent position in Berlin; ownership and future development of Neptune was retained by Thierstein at the ETH. Several other project members reduced their involvement, and software development largely came to a stop. Instead, from 1996 to 1998, Spencer-Cervato systematically analysed data in Neptune, added Paleogene data into the system and data from ODP Legs up to 135, Content expanded to over 160 sections. She published the results of her studies (Spencer-Cervato, 1999) as a monographic-length work. This publication included most of the system description given in Lazarus (1994), a detailed summary of the data content and data quality, and a summary of the initial joint publications of the group (Spencer-Cervato et al., 1993, 1994). She noted that simple compilations of occurrence data in the system were sensitive to

data outliers, in particular to geologic age errors due to uncertainties in the placement of hiatuses in the age models. Spencer-Cervato left the ETH group in 1998 and further development of the system was given to M. Knappertsbusch (Natural History Museum, Basel), a former student of Thierstein. Knappertsbusch and co-workers developed a large library of additional age models for sections not yet added to Neptune, but did not otherwise substantially expand the database. An attempt was made to put Neptune online, but the 4D software was not well suited to this purpose. Neither the ETH or Basel groups had yet used Neptune analytically to address the paleobiologic questions for which it had initially been created.

The Chronos Project (2002-2008)

Spencer-Cervato (returning her family name to Cervato and now based at the University of Iowa, Ames) had by 2002 become part of the NSF cyberinfrastructure project Chronos, whose goal was the integration of geochronostratigraphic data and the provision of a global framework of data infrastructure and geologic timescales for scientific research (Chronos Workshop Members, 2002). Although Neptune was initially only a minor element in Chronos, when Cervato took over management of the project revitalizing Neptune became a major focus of the work. Thierstein permitted the transfer to Chronos in Iowa. P. Diver, a consultant working for Chronos, ported the database content from 4D to Postgres, an modern, internet, client-server, open source SQL database system, retaining most of the data structures of the 4D version. Synonomy however was re-implemented with a separate data table, and extensive SQL was written to support basic queries. Project member D. Fils built a web front end for Internet access to Neptune content. This included a few simple queries for taxonomy and stratigraphy work. The ARC program was ported into Java, and its graphic output made directly available at the Chonos website. The ADP program was ported to Java as a stand-alone program (Bohling, 2005). This new version lacked some of the features of the True Basic version but added the ability to directly download stratigraphic event data from the Neptune database. As no structures for events or event content were present in Neptune to download, these were automatically calculated from the occurrences table as the deepest and shallowist occurrences of fossils in sections. Determination of event locations from occurrences in sections is normally done by biostratigraphic experts, who can decide whether

scattered, rare, or questionable occurrences of species near the tops and bottoms of ranges should be included or not. The event locations from Chronos ADP thus often differed significantly from those used by the ETH workers, who had derived the event locations in accordance with the initial biostratigraphic publications. Micropaleontologists associated with the project, particularly B. Huber and M. Leckie, assisted with data selection and taxonomy, with an exclusive focus on carbonate microfossils. Content increased to ca 480,000 occurrence records from this work.

Although the Chronos provided web interface was not well suited to many scientific needs, the availability for the first time of Neptune online attracted a significant number of users. Many obtained data by requesting custom downloads from Chronos staff. Others accessed this new resource from a query link at the PBDB website that had been established by PBDB head J. Alroy in 2006 at the request of the micropaleontology group of PBDB, which Lazarus at that time headed. Nearly 30 publications citing Neptune appeared between 2003 and 2009, including several in high impact journals such as Nature and PNAS (full list of known publications citing/using Neptune can be found at http://nsb-mfn-berlin.de/history.html). Many of these publications began finally to explore the paleobiologic questions, which had initially inspired the idea of creating the database in nearly 20 years earlier. Notably, early users of the system came from a broad range of fields, and included not only micropaleontologists and paleoceanographers but also invertebrate and vertebrate paleobiologists, theoretical ecologists, and researchers from other specialties.

Funding for Chronos ended in 2008. The server hosting Neptune continued to run in Ames, but without regular expert maintenance soon began to develop problems. Fils, who had moved to the Consortium for Ocean Leadership (Washington, D.C.) tried to stabilize the system and explored options for moving it to a different hosting organisation. Users who downloaded data began to report problems with data integrity (e.g., erroneous ages for nannofossil occurrences - published in Lloyd et al., 2012). Lazarus et al. (2012) (see below) could not identify such problems, but, however, analyzed different data (on radiolarians), using an older set of data obtained from the Chronos Neptune database via the PBDB link in 2008.

The IODP Taxonomic Name List Project (TNL: 2008-2010)

Twenty years after the start of the Neptune project, the taxonomic content of the names list was becoming seriously obsolete as new taxa and revisions to taxonomic nomenclature continued to be published. Nor had the original lists been complete even when initially produced. Neither Thierstein's groups nor the Chronos effort had found time and resources to address this. The opportunity came in the late 2000s as the deep-sea drilling program changed from a single ship, single lead institution effort to a multiple platform, multiple institution management system under IODP. The US IODP organisation based in College Station, Texas, had already been working on the taxonomic names for several years, where R. and I. Goll, and other local IODP staff had systematically been vetting and adding basic metadata to all taxa names in the US Janus system. Faced with the need to merge taxonomic data from multiple database systems (e.g., the system supporting the Japanese ship Chikyu, the EU consortium platforms), IODP's IT managers reached out to the micropaleontology community via a workshop (Houston, 2006) to fund a new project to harmonize the taxonomic nomenclature in their various systems. With strong advocacy by IODP managers E. Söding, J. Collier, and P. Blum a new Paleontology Coordination Group (Söding and Lazarus cochairs) was created to carry out this task. Under this umbrella, several workers (P. Bown, P. Diver, J. Dolven, S. Feist-Burkhardt, D. Harwood, K. Hooks, B. Huber, M. Iwai, D. Lazarus, J. Lees, A. Sanfilippo, N. Suzuki, S.W. Wise, and J. Young) worked from 2008-2010 to create harmonized, updated, accurate lists of names (TNLs) for the planktonic foraminifera, calcareous nannofossils, diatoms, radiolaria, and dinoflagellates. Each group of workers tackled the problem differently as the state of taxonomic knowledge and resources available differed dramatically by fossil group, but all lists were made to the same standard form, and used similar criteria to evaluate names. For the foraminifera, nannofossils, and dinoflagellates, community catalogs already existed so the work mostly was in reference to these systems (the foraminifera and nannofossil catalogs, in updated form, are currently are online at Mikrotax (www.mikrotax.org); the dinoflagellate catalog was based on Williams et al. (1998). Taxonomic catalogs for the siliceous microfossils did not exist and required compiling the primary published literature, an enormous task which led some years later to a new publication for the

radiolarians (details and results given in Lazarus et al., 2015). At the end of the TNL effort, IODP decided to no longer develop the list management work by a regular updating system, or even to incorporate the new lists in their databases. By then a new international management struture of largely independent national platform efforts had been chosen as the future of the drilling programs, and there was no longer the same need to merge database content. The TNL lists were soon used to overhaul and update the taxonomy system of a new implementation of Neptune: NSB. The living radiolarian species names in the TNL were also used to populate the Polycystinea section of the living marine biodiversity database WoRMS (Lazarus et al., 2011).

Repatriation of Neptune to Berlin as NSB (2009-2011)

L.-H. Liow and N. Stenseth of CEES, Oslo, early users of Neptune (Liow and Stenseth, 2007; Liow et al., 2010) were concerned about the fate of this new resource after Chronos had ended and hosted a small workshop in Oslo in early 2009 with Lazarus and Young. A project was conceived to implement a much simpler, scientist-maintainable offline version, initially for internal only use by members of the three research groups (CEES, Berlin, and London). The system would be released to the public later, when group members agreed. The Chronos group agreed to release Chronos Neptune for the project, with the name of the new system being changed to Neptune Sandbox Berlin (NSB) to distinguish it from the 'primary' Chronos instance of Neptune. In early 2010 after a long process of contract specification between the administrative staff of Oslo and Berlin, CEES provided a modest grant to Lazarus and Diver to extract Neptune from the complex software stack developed by Chronos for the broader goals of the Chronos project. By late 2010 an initial version in Postgres was running at the MfN and was accessible via standard SQL query tools. No web interface was available, or possible with the limited funding provided, but basic data upload functionality was provided by an external python app. Only four papers were directly generated from this initial 'members only' offline implementation: Hannisdal et al. (2012), Lazarus et al. (2012), Renaudie and Lazarus (2013), and Lazarus et al. (2014). While the other papers were studies of diversity dynamics, the Lazarus et al. (2012) paper specifically addressed the main concern raised earlier by Spencer-Cervato (1999): that occurrence data in

Neptune included a significant number of outliers, i.e., species occurrences at geologic ages that were not in accordance with the bulk of the occurrence data for that species. Lazarus et al. (2012) confirmed that such outliers existed, noted that the reasons were complex and included not only age uncertainties associated with hiatuses but also reworking of fossils from older sediments, plus other reasons. More importantly, a simple method of 'Pacman' trimming a calibrated fraction of the oldest and/or youngest occurrences in species data was introduced that, if not fully correcting the data, dramatically reduced the magnitude of outliers in the dataset, without the need to individually check all data values. Data thus treated was suited to many types of analyses, including the paleobiologic ones that were already starting to be addressed by Chronos Neptune users.

Earthtime and Development of Current Web Accessible NSB (2013-2016)

Based on a fall 2013 talk given by Lazarus at the annual German deep-sea drilling conference that presented preliminary results for Lazarus et al. (2014) and described NSB, H. Pälike, who chaired the biogeochronologic Earthtime EU project (http://earthtime-eu.eu/earthtime/), saw an opportunity to link the databasing goals of Earthtime EU with NSB. In 2014 a new 1.5 year project was approved by the Earthtime EU steering committee that supported J. Renaudie (already working as a postdoc with Lazarus in Berlin) to substantially extend NSB to include biogeochronologic data and to create a website for online access. This was the first substantial funding for NSB since Neptune had been repatriated to Berlin. NSB went online in the summer of 2014, incorporating the TNL taxonomy, and with the ability to apply Pacman trimming to downloaded occurrence data. In 2015 the system was expanded multiple times with new occurrence data (670K records by June, 2015), search functions, mapping the internal ages to use the newer Gradstein et al. (2012) age scale as default, and numerous other smaller improvements. Nearly 150 age models were newly entered or revised during this phase of development. Paleolatitude and longitude values were created for all samples in the system in 2016 by Renaudie.

Age models were developed using a combination of the Chronos version of ADP, simple x-y plots using R, or just copying the values from publications where no modification was deemed neccessary (e.g., sites with robust astrochronostratigraphy based age models). All these methods

had limitations, and in the summer of 2015 Lazarus and Diver began a new implementation of the ADP program in python, which merged the full functionality of both the original program of Lazarus (1992) and the later version by Bohling (2005).
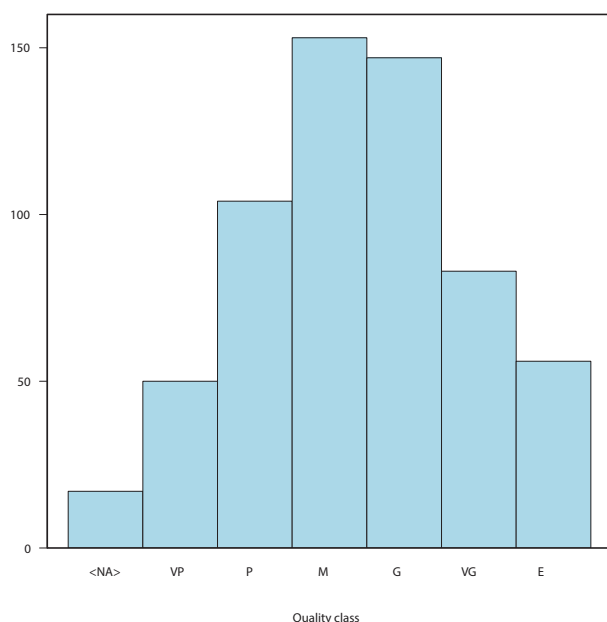
As a result of this phase of work, the code base was changed radically from the CEES supported port of the Chronos database. NSB now was a minimal SQL tables database core, with most functionality based in the website and newly written in python. Several tables inherited from the Chronos project as well as much of the more complex SQL code supporting them were retired, as the external python scripts were much easier to maintain and also ran much faster than the server-side SQL had. Synonym management for example was returned to the single taxon name table with pointers as this fit much better how taxonomic specialists viewed synonymy and was easier to maintain, with the procedural (though now in python) code system used in the original 4D version. Importantly, due to these changes, Renaudie had become effectively the main developer and maintainer of NSB, with Lazarus largely providing background management and data quality support.

One other major change occurred in this time interval. J. Young, with taxonomic help from several colleagues, had created by 2008 a new website - Nannotax - to host the community calcareous nannofossil catalog. In 2014 he decided to expand the catalog content to include actual summaries of occurrence data from NSB for each species. As neither database had a services/protocol layer to issue or respond to queries, a simple SQL access and query mechanism was established for the purpose. As described below, providing data to Nannotax (subsequently renamed to Mikrotax) substantially increased the visibility of NSB to micropaleontologists, and, via their use of Nannotax/Mikrotax, dramatically increased the total numbers of NSB users. A similar link via SQL query was created to the GBDB in 2015 to allow users to graphically view age models, although the impact on total numbers of users of NSB has not been, in comparison to Nannotax, very substantial.

Recent Developments

While many new sections and their age models had been added to Neptune/NSB over the years, pre-existing age models had not been systematically reviewed or improved for most sections in NSB since the initiation of Neptune in the 1990s. The geochronologic framework of NSB was definitely in need of review and updating. Lazarus in

2017-2018 created a complete inventory of the age models used in NSB, creating quality estimates for each model and reference folders with the primary data and history of model development, including those created but until now not used by Knappertsbusch et al. The source stratigraphic files and age model plots were centrally stored on a cloud server. At this time it was discovered that during the Chronos phase of Neptune development numerous erroneous age models had been uploaded into the system, although only for holes newly added by Chronos, and only those with just carbonate microfossils. These erroneous age models typically had only a single short line segment relating age to depth over but a small fraction of the total section depth, and typically were from one to, in extreme cases, several million years displaced from the available geochronologic data for the depth interval in the hole. These age model errors were the likely source of most of the erroneous ages for nannofossil occurrences reported by earlier workers. Given the limited number of sections involved and the very short depth intervals spanned by the erroneous models, fortunately only a small fraction of the total data in NSB had been affected. However, as samples without geologic ages are not used in any of the search outputs, a significant fraction of the sample-linked occurrence data loaded into the system was going unused, i.e., all samples from these sections outside the depth range of the erroneous model. Renaudie had taken over the python ADP project from Diver and Lazarus and had completed the new python ADP program in 2017. This was used by Lazarus to revise ca 80 age models in the system, including some older ETH era models and all Cenozoic sections with erroneous models from the Chronos project. Intervals in sections where age model uncertainties existed, or with other evidence of problems, e.g., reports of reworking were flagged so that they would not by default be included in search outputs. J. Young volunteered time in late 2017 to revise an additional nearly 40 Chronos-era age models, many also erroneous, for Cretaceous sections. Although age model development remains a work in progress, the quality of the age estimates in NSB has significantly improved (see Figure 1) and has been much better documented for more transparency. Equally important, the extension of age models to samples in many Chronos era sections that had largely lacked them substantially increased the number of usable occurrences to over 700K records, despite the removal by flagging of many other occurrence

FIGURE 1. Age model quality histogram.

records linked to samples with reworking or other problems.

The Chronos version of Neptune, having become increasingly difficult to even reach on the internet, has since ca 2018 effectively disappeared, with only a notice that all services have now migrated to other projects still left at the former website.

NSB TODAY

Current Database Design

The database is constructed around a central table of paleontological occurrences (neptune_sample_taxa; see Figure 2), linking a taxon with a sample, along with two qualifiers: the taxon abundance (which, if given, can be a numerical abundance, a percentage or a semi-quantitative category) and a note on the occurrence quality, if needed. More specifically, the latter records if the identification is ambiguous (marked by a question mark), if the occurrence is thought to be a reworked specimen (noted by 're'), or if the specimen found was only a fragment (noted by 'frg'). At the date of writing, 2461 occurrences out of 767600 have such a negative quality qualifier. Each sample points to a record in the table of samples (neptune_sample), and each taxon points to a record in the taxonomy table (neptune_taxonomy). The taxon referred to in the paleontological occurrences table bears the exact name that was given

to it in the original publication where the occurrence was reported.

Information about the samples is contained in table neptune_sample. So far, only DSDP, ODP, or IODP deep-sea drilling samples are contained in the database, thus constraining the form of this table: four columns give the sample leg, site, and hole (aggregated under the key hole_id), the core, the section (field section_number) and the interval top (sect_interval_top). Core catchers are arbitrarily given the section number 9 (IODP cores never contain more than eight sections). For convenience, the corresponding depth of the sample is given in field sample_depth_mbsf: in practice, it is computed from table neptune_core and the information supplied in this table's columns core, section_number, and sect_interval_top. Similarly, the sample age (sample_age_ma) is calculated based on the sample depth and the age model of the site (if present in the database) from table neptune_age_model. Finally, the sample paleocoordinates (paleo_longitude and paleo_latitude) are computed based on the sample age and site, using the table neptune_paleogeography. Normally each sample taken from deep-sea drilling sections was studied for a single group of fossil organisms, and this practice is reflected in the database: each sample has a single field for the fossil group, and metadata on this studied group, i.e., the group abundance in the sample (sample_group_abundance) and its preservation (sample_preservation). In addition the field sample_status is used to flag samples believed to be reworked, or located in a geologically disturbed interval of the section. Finally, the field dataset_id links tables neptune_sample and neptune_sod_off, the latter containing metadata relating to the original dataset from which the sample comes from. Though this field was present in the legacy Neptune database it was unfortunately in practice left blank prior to this new update (see section on "data quality control"). Each sample is, in theory, characterized by a unique combination of fields hole_id, sample_depth_mbsf, fossil_group, dataset_id. In practice an integer key (sample_id) is used to ensure each data record can be uniquely identified, even for multiple studies of the same sample interval, and/or data without a known dataset_id.

The table neptune_core as mentioned before contains for each drilled hole the core name (defined as a unique combination of fields hole_id and core), its top depth (core_top_mbsf), and its total, recovered length (core_length). This is not only sufficient to calculate the depth of a given
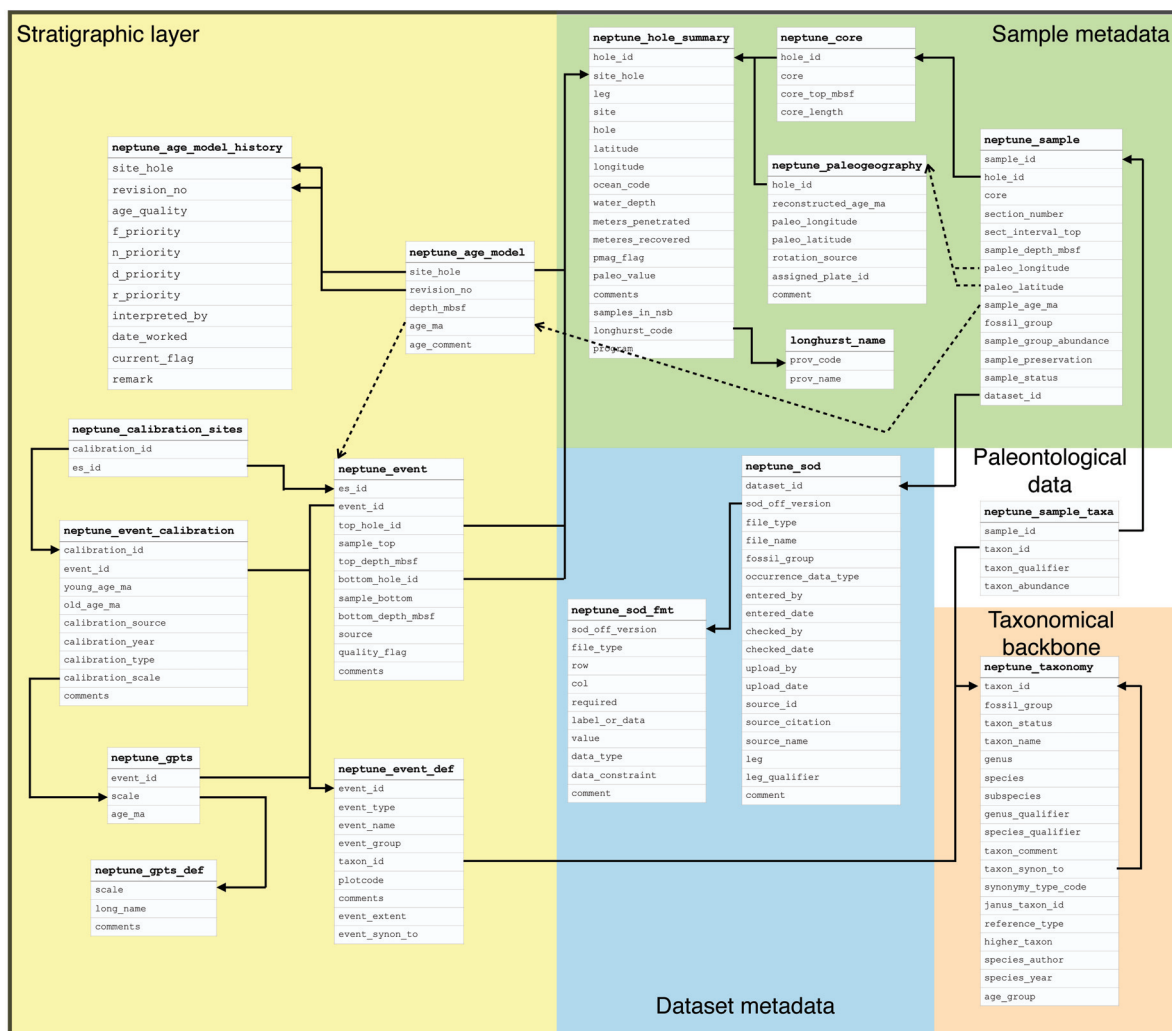
FIGURE 2. Current architecture of the NSB database.

sample given its name (i.e., as per IODP nomenclature of the form "Leg-Site-Hole-Core-Section, Interval", e.g., "120-751A-10-2, 10") but also to perform sanity checks on sample names (i.e., if the depth interval of the sample is not contained in the length of the core).

The neptune_hole_summary table has metadata on the hole containing the sample. Apart from the needed fields for IODP sample nomenclature (leg, site, and hole), it contains the hole latitude and longitude, a code giving the name of the ocean the hole come from (field ocean_code whose values can be ATL, ANT, PAC, MED, and IND depending on whether the sample comes, respectively, from the Atlantic, Southern, and Pacific oceans, the Mediterrannean Sea or the Indian Ocean), the water depth of the hole, information on the drilling process (i.e., how many meters of sedi-

ments were penetrated and recovered) as well as subjective metadata on the value of the paleomagnetic or paleontological data at the site (pmag_flag and paleo_value). The field longhurst_code (which links to table longhurst_name) gives the name of the ecologic planktonic province in which the hole currently is located, according to Longhurst (1988) biogeographic classification.

In addition to the Longhurst biogeographic data, an additional table was created during NSB's development in Berlin in the sample metadata layer: the table neptune_paleogeography. As mentioned previously, this table is used to compute the paleocoordinates of samples based on which sites they come from and the age of the sample. The table contains a reconstructed paleolatitude and paleolongitude for each site in the database, calculated at 0.50 Myr intervals. Each of the paleocoor-

dinates were calculated based on the site's current latitude and longitude using GPlates (Boyden et al., 2011), along with a given plate rotation file (named in the corresponding field of the table: currently, only the rotation file from Seton et al., 2012 has been implemented).

The legacy age model system consists of two tables: neptune_age_model and neptune_age_-model_history. The former contains the age models themselves as a list of tie-points, keyed by the name of the hole and a revision number to allow several different age models for a single site to coexist. The latter table contains metadata for the age models and more importantly shows which of the different age models for a given site is the one currently used in the database to calculate ages for samples and fossil occurrences. Other metadata include the name of the age model uploader, the day of the upload, a quality assessment for the age model, and remarks, usually on the quality of the overall model or of specific sections of the model and/or the source of the data if needed. Four additional columns (f_priority, n_priority, d_priority, and r_priority) give a priority assessment on the value of the occurrence data published from the hole, on a scale from 0 (no data of value) to 3 (highly important occurrence data). The data in these fields are a legacy from the initial ETH phase of development and in the future should be migrated to a logically better suited table with metadata about holes, e.g., neptune_hole_summary.

Finally, linked via the neptune_sample field dataset_id, the neptune_sod table contains metadata on the specific occurrence data tables (called, somewhat misleadingly by micropaleontologists 'rangecharts') used to populate NSB. Metadata include the reference of the publication, along with various metadata specific to the SOD format (see Lazarus et al., 2018).

The TNL based taxonomic table of NSB, neptune_taxonomy uses an integer key system developed for the TNL project - a unique, seven decimal integer ID (taxon_id) where the leftmost two digits give the higher level taxon group and the remaining five the individual taxon name, and in addition fields for genus, species, subspecies (if needed) ,and genus and species qualifiers (i.e., "?", "cf" etc.); a field indicating the validity of the taxon name; taxon_status, which could be one of V (valid species), S (synonym), B (for a valid subspecific group), G (open-nomenclature taxon, with genus assignment), Q (questionable identification, e.g., *Catinaster coalithus*?), U (for unknown, i.e., status not determined), I (for invalid, i.e., taxon name

wrongly assigned to the group by the author), and H (for higher-level taxon: i.e., open-nomenclature taxon, unresolved even to the genus-level). For the time being, the few silicoflagellates and ebridians present in the database were originally in the publications, and still are, entered as radiolarians, with a taxon_status of "I". Here also future improvement to define and use additional fossil group values, or simply to use the encoded values in the key field is needed.

Additional fields are present to link the integer taxon_id used in NSB to the mixed alphanumeric system used in the previous incarnation of Neptune (nsb_taxon_id) or to a different, integer system used in the Janus database (janus_taxon_id); to qualify the type of synonymy, to record the author-year combination and to give additional comments (such as the Epoch or stratigraphic Stage to which the taxon belongs to, or any other info deemed useful).

A function to resolve the synonymies, nsbtax-syn written in pgSQL, is made available on the database cluster. Similarly the web portal uses a python-based synonymy resolver and the companion R package an R-based one (see Supplementary Material). All three are recursive as some synonyms point to another synonym: this is due to the presence of objective synonyms (immutable) and subjective ones (which are more likely to change).

The new stratigraphic layer includes several tables: neptune_event_def, neptune_event_calibration, neptune_event, neptune_gpts and neptune_gpts_def. The central table is neptune_event_def (Figure 1), which contains the definitions of the stratigraphic events. All types of events are held in this table including biostratigraphic and magnetostratigraphic events, and in principle many other types such as isotope-based stratigraphic events. At the moment it holds events for planktonic foraminifera, calcareous nannofossils, diatoms, radiolaria, dinoflagellates, Bolboforma, and magnetostratigraphy. Note that Bolboforma, an extinct group of calcareous microfossil has no information for it in the taxonomy or occurrence tables, although such data could be added in the future. A single event is defined by a unique combination of event_type – i.e., whether it is a first (BOT) or last (TOP) occurrence, an ecological acme (ACME), a first (FCO) or last (LCO) common occurrence, an evolutionary transition (EVOL), a local/regional disappearance (DISP) or reapparance (REAP), or a stratigraphic zone sensu lato (ZONE) –, event_name (in the case of biostra-

tigraphic event, a taxon name) and event_extent (i.e., the geographical extent of the event). This table also holds, for those events based on taxa, the taxon_id in the taxonomy table. This allows unambiguous matching of taxonomy to event definitions and allows matching events defined with names subsequently synonymized, e.g., the top and bottom of the species currently called *Globoconella conomiozea*, where both events were defined under the older name *Globorotalia conomiozea*.

The table neptune_event_calibration contains the calibrations of the biostratigraphic events: it has a one-to-many relationship with neptune_event_def in order to allow several different calibrations to exist for a single event. The age estimates in this table are given on the same age scale (GPTS) as in the original publication in which they were calibrated (field calibration_scale). The type of calibration is indicated in field calibration_type (M for magnetostratigraphic, O for orbital tuning, F, N, or D correspond to the three microfossils groups; any other type of calibration is spelled out, i.e., Isotope, Ammonite, etc.).

The third table, neptune_event contains events in sections, whether they've been used for calibrations, for building age models, or just observed in individual sections. It is conceptually similar to the table neptune_sample_taxa. It has a key (field es_id) to allow the resolution of the many-to-many relationship between this table and table neptune_event_calibration, as one calibration (a record in the calibration table) can be based on several observed sites (multiple records in the events table) and one event in one site can be used by several calibrations (the relationship is resolved by the table neptune_calibration_sites, which links the calibrations to the events in sections that were used to produce them).

Finally the tables neptune_gpts and neptune_gpts_def contain the ages of the magnetochrons (referred to by their event_id) according to various geomagnetic polarity time scales (GPTS): Berggren et al. (1985), Berggren et al. (1995), Cande and Kent (1995), Gradstein et al. (2004), Pälike et al. (2006), and Gradstein et al. (2012). This table allows the translation, via code, of age estimates from one GPTS to another. All age estimates in the micropaleontological occurrence layer of the NSB database are externally reported using the most recent GPTS in the system, i.e., Gradstein et al. (2012). More recently published scales, e.g., Ogg et al. (2016) differ only

trivially in the Cenozoic from the Gradstein et al. (2012) values, and thus were not yet implemented.

Content Improvements

Prior to the database renovation (2014-2017), Neptune held 500,622 microfossil occurrence records from 42,034 samples, a taxonomic list of 11,767 taxa names and age models for 289 holes. As of today, NSB contains 768,057 microfossil occurrences values from 61,139 samples, with a taxonomy list of 18,915 taxa names. It also contains 659 age models for 458 holes, based on 27,072 individually observed stratigraphic events, from 2 812 defined stratigraphic events, with 2 641 calibrations.

In the legacy database, data were entered without keeping track of their sources: 30 years later, this design choice makes checking the quality of the data entry or the quality of the source material challenging. This is why substantial effort has been made to not only document the sources of the new datasets entered since 2010 but also track back the sources of the legacy data (and correct the data accordingly). At the time of writing, 48,862 out of 61,139 samples are sourced (i.e., 79.9% of them); or 598,536 out of 768, 057 species occurrences are sourced (i.e., 77.9%). This allowed us to flag and eliminate duplicate samples and also in a few cases to add back missing samples from large datasets, which in the Chronos phase were sometimes not fully loaded.

The legacy age model library quality was reassessed. Although still a work in progress, major improvements in average age model quality have been achieved. Problematic age models (or problematic sections in age models) were sought by using a variety of outlier detection algorithms on the microfossil samples or directly on the stratigraphic events. In particular a 'Pacman' profiling (Lazarus et al., 2012) was performed to identify suspect sections of age models (in particular incorrect placement of hiatuses). The creation of the table neptune_event and neptune_event_calibration also allowed computing the difference between calibrated age and projected age on the current age model for each stratigraphic event, which helped identify age models for which this mean difference was statistically higher than on the overall database.

In addition to problematic age models, age models for holes containing a large number of undated occurrences were also made or extended to include all the samples. As a result, the number

of undated occurrences fell to ca. 8.5% (i.e., 65,308 undated occurrences).

As of the time of writing (Figure 1), 108 age models out of the 463 current ones are considered poor or very poor (median sample age error ca 1 my or more, accounting for 19.3% of all dated samples - mostly older DSDP sections), while 230 are considered moderately good to good (median error ca 0.5 my, 58.9% of all dated samples) and 123 very good to excellent (median error ca .5 to .2 my - or less, for orbitally tuned sections, 21.8% of all dated samples). These are subjective impressions, mostly based on the observed scatter of the stratigraphic event data vs the fitted line of correlation in age depth plots, but as work on age models continues we expect the relative percentages of sections with good or better quality age estimates to further increase.

Database Access through the Web Portal and the NSP_ADP_wx Software

Web portal. In addition to being accessible through the postgreSQL command line interface or any one of a number of generic GUI database manager applications, the NSB database is also accessible through a web portal at the following address: http://nsb-mfn-berlin.de. This website is programmed in the Django Web Framework (Django Software Foundation, 2013) and is being served by Apache HTTP Server (Apache Software Foundation, 2005). The advantages of the Django Web Framework are threefold: 1) server-side, as Django is essentially just a python module, everything is coded in python 2.7 (Python Software Foundation, 2010), which not only allows the website to be maintained by a programmer with a background in scientific programming rather than web programming, but, via the extensive support in Python for scientific data analysis, also supports scientific computations to be performed directly by the server on the database outputs in the native language of the website; 2) Django was specifically conceived as a bridge between database and website, meaning it natively resolves many database-specific security issues (such as SQL injections); 3) client-side, though the website is dynamic (i.e., web pages will change depending on the state of the database and on the choices of the user), web pages are rendered as simple HTML pages (with the exception of a single page containing a jQuery script, see below), meaning not only that they are lightweight but also that they are compatible with virtually any browser/platform.

The website allows the user to perform a wide variety of predefined searches on the database:

- search for microfossil occurrences (see Figure 3): this is the website main functionality. The user can search the micropaleontological occurrence data by geographic (latitude, longitude, ocean basin, deep-sea drilling expedition), taxonomic (fossil group, genus, species) and/or temporal (time span in million years) criteria and filter the data by taxonomic (questionable identifications, open-nomenclature taxa) and/or stratigraphic quality (possible reworking, age model robustness). The query output can in addition to this be reported using a different GPTS than the one used in the database, and the taxonomy can be resolved using the TNL synonym mappings (see above). User defined pacman trimming of the output to remove most outliers (Lazarus et al., 2012) is also offered as an option. The result of the query is a tab-separated csv file that the user can either download right away or anytime during the following month (see Figure 4), if the user is registered.

This file contains a large amount of information including (on top of, obviously, the species name, the resolved name if synonymy resolution was selected by the user, and detailed information on the sample in which the species was found) a wide array of data and metadata such as an estimate of the paleocoordinates of the sample, based on Seton et al. (2012) plate motion model and the sample estimated age according to the current age model, the publication in which the occurrence was reported, and the current ecological province in which the sample is found according to Longhurst's (1988) classification of the ocean biome.

- search for microfossil taxonomy (see Figure 5): through this interface the taxonomic name list (TNL) can be directly explored. This is particularly useful if synonymies of microfossil occurrences were resolved using the previous interface: the user can then check which species were considered synonyms and resulted in the observed, aggregated occurrences.

- search for stratigraphic event calibration (see Figure 6): in addition to providing the various available calibrated ages for bio- or magnetostratigraphic events, it also shows in which samples were those events found, and what the estimated age of the samples is according to the age model for its site in NSB. This provides the user with a quick way to judge either the age model quality of a given site or the quality of the event calibration, the geographic diachroneity of the event, or (if an event lacks any modern calibration) the potential age range of an event. The query box for this page is provided with a jQuery script allowing text completion, in order to bypass any issues that arise from looking for a bio-

FIGURE 3. Query page for microfossil occurrences.

FIGURE 4. Result page for microfossil occurrences (the specific query here was for radiolarians from samples between 20 and 25 Ma).
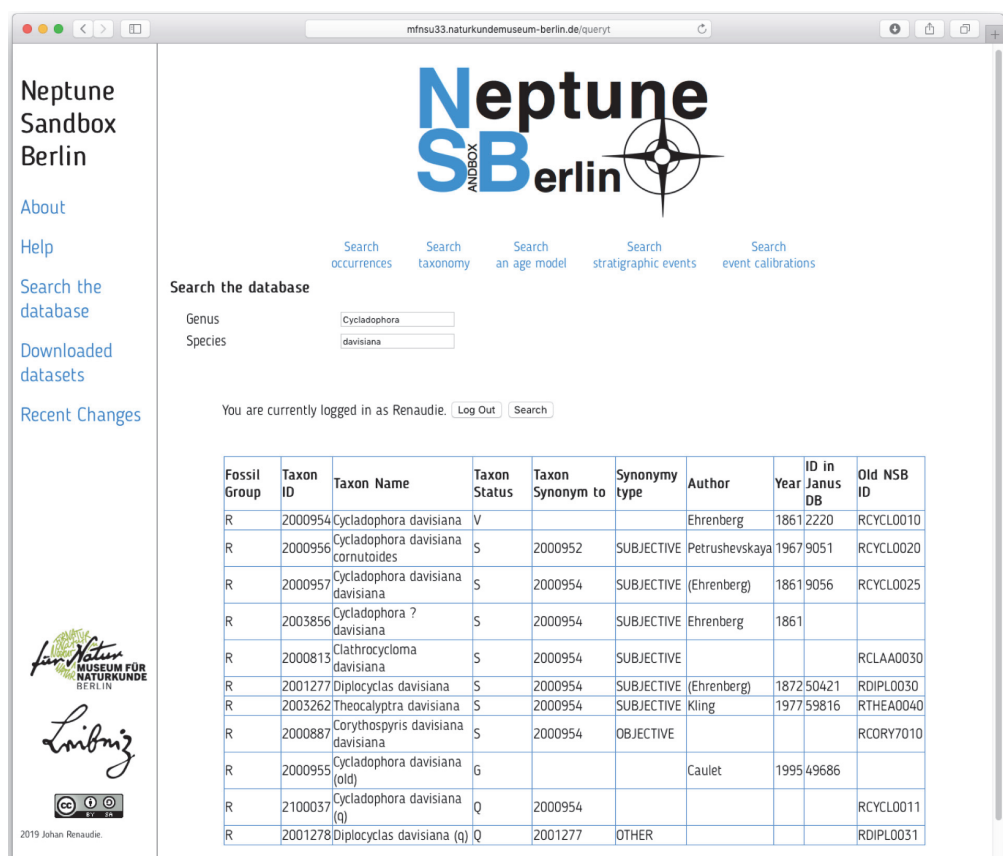
FIGURE 5. Result page for microfossil taxonomy.

stratigraphic event for which the base species is known by different (subjective or objective) synonym names.

- search for stratigraphic events in specific sites (see Figure 7): for each event stored in NSB for a given hole, the upper and lower limits of the event are given, as well as the source of the datum recognition, the age according to the latest calibration (interpolated in the GPTS scale chosen by the user if needed, so that all ages are given on the same scale), the source of that calibration and a comment (usually if the event was specifically used by the calibration source).

- search the age model library (see Figure 8): each age model is presented as a simple table of age-depth tiepoints. All age models for a given hole are returned by the search, as well as metadata such as whether the age model is the one currently used in the database (the current model is also visually highlighted on the page), the name of the user that uploaded the age model as well as the date of upload, a quality assessment on the age model and remarks (either warning notes on a specific portion of the age model or sources for astronomical, isotopic calibrations, or any kind of additional stratigraphic information used for the age model but not found in any tables of the database).

NSB_ADP_wx. The software NSB_ADP_wx (current version 0.7) was written in python 2.7 (Python Software Foundation, 2010), plus the python modules matplotlib (to plot the data and to interact with the plotting area; Hunter, 2007), pandas (to manipulate downloaded data tables), SQLalchemy (to interact with the database; McKinney, 2010; Bayer, 2014), and wxPython (to build the Graphical User Interface; Dunn, 2014). It is available (see http://github.com/plannapus/nsb_adp_wx/releases) as a script or as a double-clickable software for Mac OSX 10.13+ and Windows 10. Its complete user guide is given here as Supplementary Material.

The starting window is always available and displays information (including errors and warnings) during the session. The user can chose to work based on biostratigraphic data and age models included in NSB, from their own files or a mixture of both. In any case, the source of the data will be displayed at any time in the main window (see Figure 9) where the age-depth plot will be displayed and will be also present on any saved pictures. If the user chooses to work from NSB data, they will get to choose from all sites for which data is present in the database (see Figure 10) and

FIGURE 6. Result page for stratigraphic event calibrations.

which age model revision to use if there is any (see Figure 11), at which point they will see metadata on the selected age model revision.

The age-depth plot window displays the data and age-depth relationship (line of correlation) (see Figure 10), and the user interacts with the graph using a combination of the mouse and keyboard keys to issue specific commands. The main window graphic elements and commands are:

- the stratigraphic event data for the hole. By default the names of the data are not displayed, but can be shown by simply clicking on, or near, the datum. Each stratigraphic data category (i.e., fossil group or magnetostratigtraphic) is displayed using a different symbol and colour (if the user chose to display the age-depth plot in colour). Those symbols and colours are customizable by the user by editing the file plot_config.txt found in the folder REF. This file being a tab-delimited text file is easily modifiable. Any number of additional stratigraphic event groups can be easily added as a new line and will be recognized by the software once this file is edited.

- a line-of-correlation, as a green line. If none was selected, the starting age model will be a line linking the minimum and maximum ages and depths of the biostratigraphic data. Each tie-point of the line of correlation is draggable (though it can not be dragged in a way that would create a stratigraphic inversion, e.g., higher in the section than the previous one). A tie-point can be inserted at the current mouse position using the 'i' letter key, and a tie-point hovered by the mouse can be deleted by

FIGURE 7. Result page for stratigraphic events on a single site.

using the 'd' key. If two tiepoints are at the same depth, a hiatus will be created and highlighted in red. If two tiepoints ar so similar in depth that the slope between them implies a sedimentation rate lower than 0.5m per Myr, a warning will be displayed inviting the user to make it an hiatus, and the slope will be highlighted in red. This is to prevent sample depths from an interval between tie points inadvertantly being assigned intermediate ages if the interval is most likely missing in a hiatus, but allows such intervals of 'condensed' sedimentation to be created if there is (external to the program) evidence for it. At any point the current line-of-correlation is accessible and savable in numerical form with menu elements.

- the timescale (at the top), the GPTS (at the bottom), and the cores and sections limits (at the right).

In addition to this, the user can, using menu elements, project the biostratigraphic data on the line-of-correlation (to estimate the age of poorly defined events for instance), save, and load new data or a new line-of-correlation and compare two lines-of-correlation (see Figure 12).

The format of the input (when not downloading straight from NSB) and output files, although slightly cumbersome, respect the format used in the previous implementations of this Age-Depth plot program for the sake of backward-compability

(see User Guide in the Supplementary Material for more information). Pre-made files containing the sites' core depth information (used to translate sample names into depth for instance, and to draw the core stack on the right side of the plot, when not downloaded straight from the database) for most of the DSDP and ODP legs are provided with the software (in subfolder "CORE").

DISCUSSION

On the NSB Model

The first thing that should be noted is the uniqueness of the fossil record it compiles. Not only is the microfossil record the richest paleontological record (Lazarus 2011), but the samples recorded in the database deriving from deep-sea drilling campaigns (DSDP, ODP and IODP) have, thanks to the unique multidisciplinarity of these scientific projects, a phenomenal amount of information available for each of them, and lastly also offer a reasonably complete plankton biogeographic and stratigraphic coverage of the deep-sea sedimentary record.

It should be mentioned that this record is not without its flaws, and those flaws should be kept in mind when working with this kind of data: because of its vastness, coupled with the fact that most of

18

FIGURE 8. Result page for age model query.

the published data were not produced to document biodiversity (usually biostratigraphy was, instead, the goal), and, often, the lack of manpower preventing the completion of more complete surveys, micropaleontologists working on deep-sea drilling sediments have been recording data frequently biased toward the biostratigraphic or ecological markers needed for their study, while often discarding a significant proportion of the species present in their samples (see Lazarus, 2011). Also micropaleontology, as a science, has been constrained by technology and evolving methodologies. Sample preparation practices have changed throughout the last 50 years in a way that makes older data different enough from younger ones to be noteworthy: for instance, the use of 63 to 80µm sieves to study radiolarian faunas used to be prevalent while now a 45µm sieve is the standard, which has the effect of reducing the amount of smaller species reported in older studies.

The NSB database has a certain number of unique positive features: in particular, an expert,
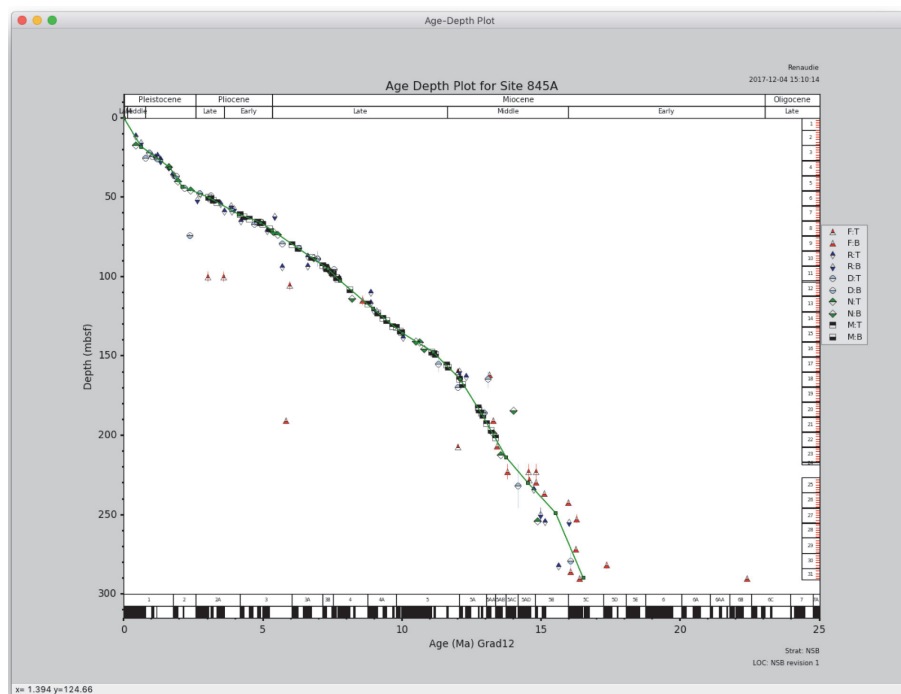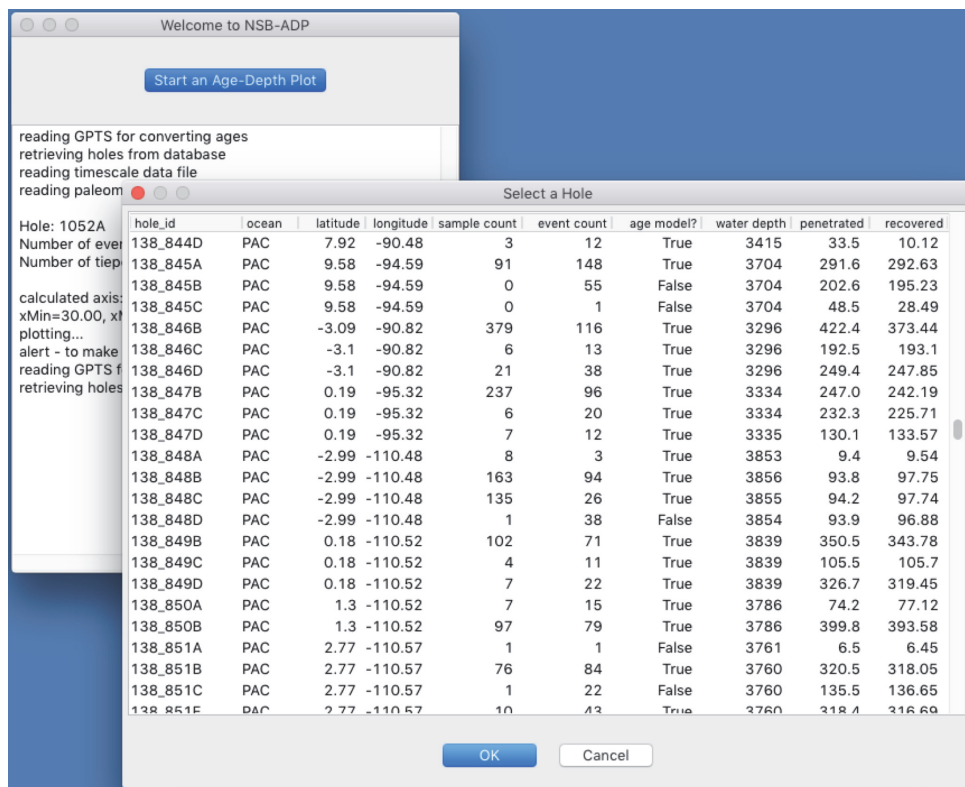
FIGURE 9. Age-Depth plot window.

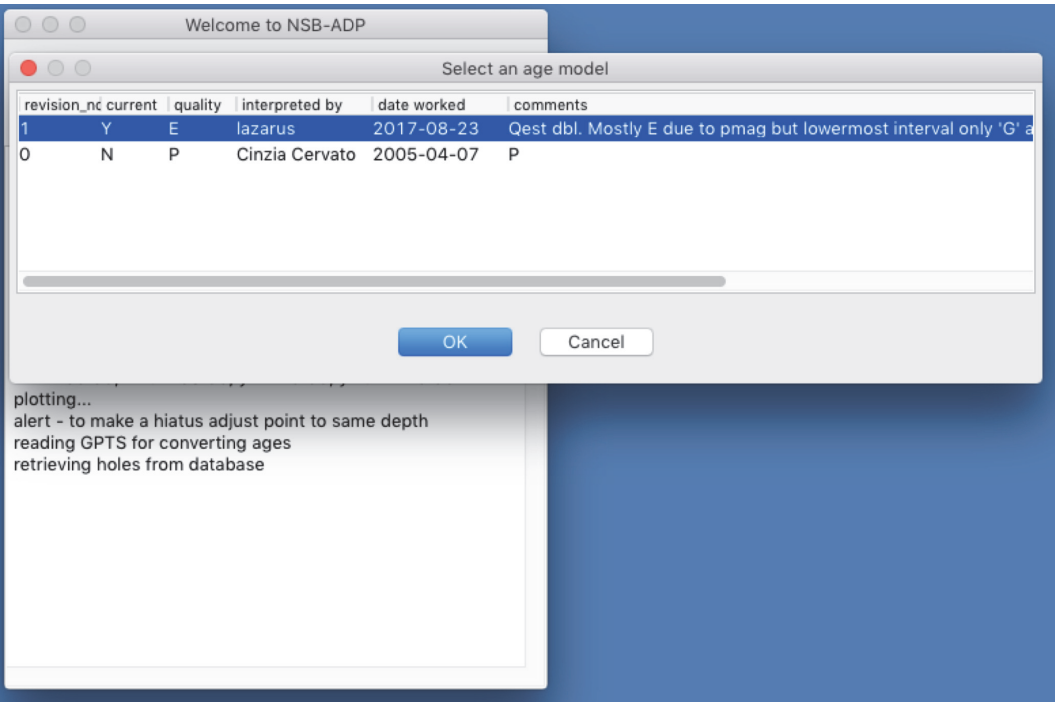**FIGURE 10.** Hole selection in NSB_ADP_wx.

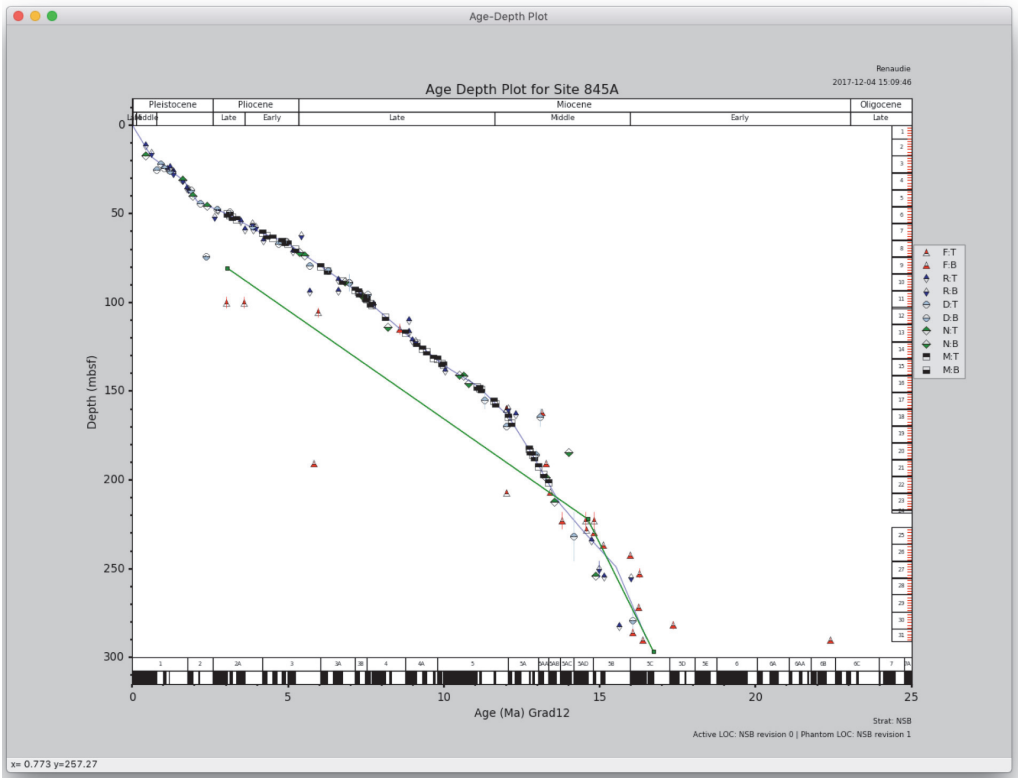**FIGURE 11.** Age model selection in NSB_ADP_wx.

FIGURE 12. LOC comparison.

quorum-based, vetting system for its core taxonomic and stratigraphic framework, and a dynamic, numerical, "higher"-resolution age assignment system.

While the data in NSB record with high fidelity the data reported in the original dataset, thus documenting the taxonomic opinions of the original author, species can also be resolved using the TNL taxonomic backbone resulting from a quorum of experts of each taxonomic group. Similarly the stratigraphic library, while based on data from the original published stratigraphic studies, contains not only newer age models produced specifically for NSB by experts, but also vetted (through their age quality assignment) by experts. Only a handful of age models are reproduced from the literature without modificationonly when they correspond to the state of the art (e.g., astrochronologically-tuned age models).
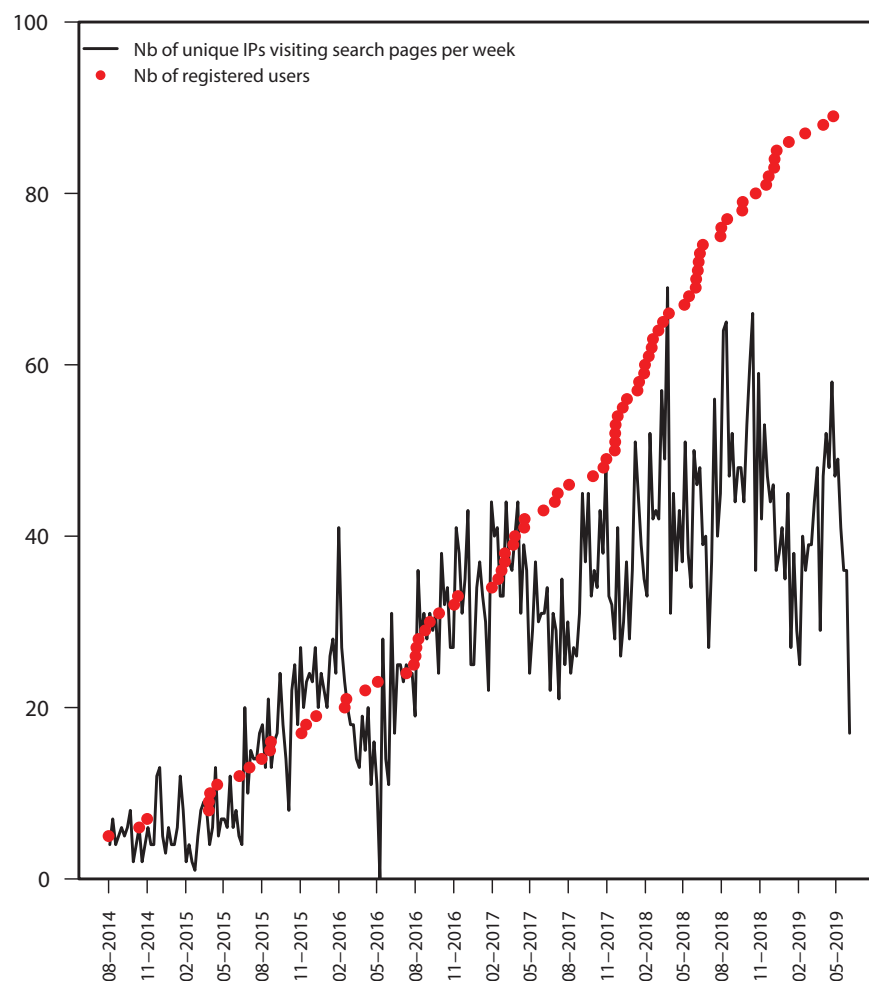
This approach to sample age assignment is also, to our knowledge, unique to NSB as a geoscientific database: sample ages are neither based automatically on the original age assignment proposed at the time in the original dataset, nor are they simply categorized in broad categories (Stage, Epoch, etc.) but are based on a numerical age-depth model, which changes everytime our knowledge of the section's stratigraphy, or of the stratigraphic data (placement and calibration) improves.

On the History of Neptune

Several aspects of the history of this database project are worth discussing. The most obvious is the remarkable complexity and variability of the path over which development took place. Between the initial concept in the late 1980s in Zürich and the current system in Berlin, development involved at least three main groups in three countries, contributions from dozens of researchers, and funding from a wide variety of sources, with substantial intervals in between in which no funding was available at all to maintain or develop the system. Indeed, since the end of the Earthtime EU grant in 2015, NSB in Berlin has been developed as a voluntary effort without any specific funding for it. This pattern of erratic funding and the vital role played by voluntary effort is in fact common to many other science databases (Baker, 2012; Attwood et al., 2015; Gabella et al., 2018), and also to other paleobiology databases. PBDB for example, although it is a much younger system, has shifted institutional hosts several times, including from the USA to Australia and back, changed lead developers along the way, and has had to repeatedly hunt for renewed funding. All the while, a massive, voluntary effort by the paleontology community kept PBDB growing and contributing to fundamental research. Although these paleontology databases are not so intensely used (or as expensive) as to be comparable to a major physical facility such as a nuclear accelerator, even more comparable (in the intensity of use) parts of science infrastructure such as a small telescope, or small marine station can normally expect a much steadier, more predictable level of funding with which to plan and carry out development. Given an approximate half life of 10 years for science database projects (Atwood et al., 2015), an oft-told aphorism about the digital world is that 'digital information lasts forever, or five years, whichever comes first' (Rothenberg, 1995) can be paraphrased as 'support for digital database development lasts forever, or 10 years, whichever comes first'. Developing the cyberinfrastructure needed for twenty-first century science is still a challenge for the world's science funding systems.

A second point, and partially reflecting the point above, is the great importance played by voluntary contributions by numerous individuals to the success of the project. A great many individuals, not all named above, have contributed time for taxonomic list editing, programming, age modeling, and other tasks; others have allocated discretionary funds to keep the project afloat in lean times. We are deeply greatful to all who have helped. A related point is the importance of (voluntary) community involvement. In the past, Neptune has followed a somewhat different path than, e.g., PBDB, which is supported by hundreds of active contributors. In the initial phase of development in the early 1990s online access was not technically feasible, limiting the ability to connect to a broad community. The database only went online in the early 2000s via Chronos, and was not then effectively connected to a community of supporters. This latter reflects in part the demise, soon after Neptune's online appearance, of the Chronos project. It also reflects an unusual aspect of Neptune - the majority of the early direct users (ecologists, paleoceanographers, macro-organism paleobiologists) were not themselves marine micropaleontologists, and thus were not well equipped to help maintain the system, as they have not enough knowledge of age modeling of deep-sea sections or the taxonomy of the fossils in the system. Now that NSB is better connected to the community of marine micropaleontologists, both by its own website and in particular via widely used linked resources such

FIGURE 13. Weekly visitors and registered users.

as Mikrotax, it may be hoped that community involvement will improve, particularly in specialists capable of helping to maintain and improve the data content. Still, the unusual degree of cross-disciplinary interest in NSB remains both a strength and a weakness.

Lastly, in databases as in so many other human activities, a project is more likely to be successful when it is led by an individual, or group of individuals, who are strongly committed to the project and provide the leadership and drive to keep the project moving forward. In Neptune's 'interregnum' this was lacking, and nearly led to the extinction of the resource. Ideally, such resources should not only be hosted by institutions, they should be seen by the host institution as a central element of their activity profile, and thus provided with institutional resources for long-term management and support. This seems to be a key characteristic for other science databases that have persisted for longer periods of time (Attwood et al., 2015).
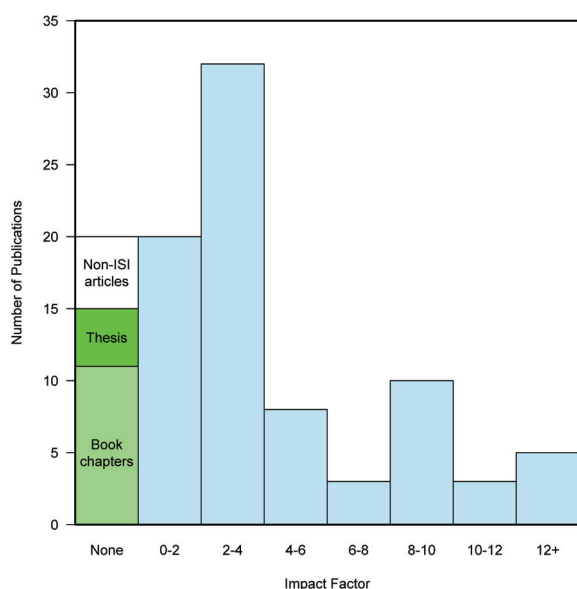
CONCLUSIONS

Past Usage of the Database

Since the creation of the web portal (http://nsb-mfn-berlin.de), almost 90 users have registered (see Figure 13). In 2018, search pages averaged ca. 40 unique IPs a week visiting them, for a total of 1 258 unique IP visiting over the year. Additionally, NSB delivers occurrence data directly to the community catalog system Mikrotax (http://mikrotax.org), which can be seen on each species and genera pages (Young et al., 2019). Over 2018, Mikrotax was visited by 24 161 unique visitors.

Since the creation of the Neptune database in 1990 to the beginning of 2019, 112 scientific publications have been written about or used the database (including 11 book chapters, 6 PhD theses, and 81 articles in ISI-listed journals; see Figure 14), with themes varying from paleobiology (e.g., Rabosky and Sorrhanus, 2009; Hannisdal et al. 2017), paleoceanography (e.g., Cermeño et al

FIGURE 14. Histogram of publications using or about NSB, spread by Clarivate Analytics' Impact Factor (JIF, 2017).

2015), marine sedimentology (e.g., Muttoni and Kent 2007), theoretical biology (e.g., Lewitus and Morlon 2017), and systematics (e.g., Fordham et al. 2018). A complete list of publications using NSB or about NSB can be found at http://nsb-mfn-berlin.de/history.html.

Future Developments

For NSB to continue to develop in the future it will need to be anchored into a longer term support framework so that it is less dependent on just a few individuals, whose career demands cannot be depended on to provide stable support indefinitely. As of this writing initial steps to create this framework have been made. An international scientific advisory board has been constituted to oversee and advise on the long-term development of the database. Planning is now underway to establish NSB as a long-term institutionally supported resource at its host institution, the Museum für Naturkunde in Berlin, and to define NSB as part of larger national and international science database networking initiatives such as Germany's NFDI and the IUGS managed DDE. Although integration into institutional activities is in itself no guarantee of adequate funding, it makes the long-term perspective clear that is essential for obtaining future funding from external granting agencies, and is at least likely to ensure an internal, institutional baseline level of support to maintain the system.

Longer term development of NSB as a globally available resource in an institutional framework also implies a significant degree of additional technical development. At a minimum, NSB needs to become a data provider using established technologies for data exchange in global scientific data networks. Currently the accepted standard for this is via provision of an API to network clients. While this is also still in the early planning stages, there are already in existence APIs that, if not complete enough to give access to all data types in NSB at least provide an essential core level of functionality. These include the ADCDEFG protocol (Petersen et al., 2018) developed primarily for exchange of museum collection data, and the EarthLife Consortium API, which is already in use by other large paleontology research databases such as the PBDB and Neotoma (http://earthlife-consortium.org).

Future growth in NSB will also depend on more effectively engaging the external user community, primarily via the provision of improved data analytics, either directly from NSB, or indirectly via networked data feeds to cooperating databases. Examples of such cooperation with Mikrotax and GBDB need to be extended. Improved tools for user data input are also needed for occurrence datasets and age models. The existing data links to Mikrotax need to be made bidirectional so that, as Mikrotax matures into a comprehensive taxonomic catalog system for all microfossil groups, it can be the primary data provider for the taxonomic name system used in NSB.

Lastly, the data content of NSB needs to be expanded to additional microfossil groups, data types, and geologic time intervals. Some of this data content will hopefully come from a more engaged user community, other data may be captured as part of specific, targeted externally funded research projects. The focus on high quality data from sections with numerical age models, particularly sections with complementary paleoenvironmental data means that NSB will always be a more selective, time, space, and taxa limited system than more general databases such as PBDB and GBDB. The upside is that NSB will continue in the future to provide the best available data on oceanic species diversity dynamics and chronology of the deep-sea sediment record.

Code and Data Availability

The compiled software NSB_ADP_wx is available on the following GitHub repository: http://github.com/plannapus/nsb_adp_wx/releases.   All

data discussed here are available on the NSB website discussed herein (http://nsb-mfn-berlin.de). Though we suggest users register for an account as it allows them to keep track of their previous queries, any unregistered user can use the guest account "guest" (with password "arm_aber_sexy"). Instructions on how to reach the database directly using a database manager software such as pgAdmin3 or Navicat can be found on the website portal.

A companion R package can be found on the following GitHub repository: http://github.com/plannapus/NSBcompanion/releases. Its user guide and the user guide for NSB_ADP_wx are given as supplementary material.

Author Contributions

JR and DBL wrote the manuscript (respectively, the current state of NSB and its history). JR created the stratigraphic layer of the database and the website. JR and DBL maintain the database. JR, DBL, and PD made the discussed modifications to the database. PD and JR wrote NSB_ADP_wx, while DBL designed it.

REFERENCES

Alroy, J., Marshall, C., Bambach, R., Bezusko, K., Foote, M., Fürsich, F., Hansen, T. A., Holland, S., Ivany, L., Jablonski, D., Jacobs, D.K., Jones, D.C., Kosnik, M.A., Lidgard, S., Low, S., Miller, A.I., Novack-Gottshall, P.M., Olszewski, T.D., Patzkowsky, M.E., Raup, D.M., Roy, K., Sepkoski Jr, J.J., Sommers, M.G., Wagner, P.J., and Webber, A. 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences,* 98(11):6261-6266. https://doi.org/10.1073/pnas.111144698

Apache Software Foundation. 2005. Apache HTTP Server, version 2.2. http://www.apache.org/

Attwood, T.K., Agit, B., and Ellis, L.B.M. 2015. Longevity of biological databases. *EMBnet.journal,* 21:e803. https://doi.org/10.14806/ej.21.0.803

Baker, M. 2012. Databases fight funding cuts. *Nature,* 489:19. https://doi.org/10.1038/489019a

Bayer, M. 2014. Sqlalchemy, p. 291-314. In Brown, A. and Wilson, G. (eds.), *The Architecture of Open Source Applications*, volume II. Mountain View, AOSA.

Berggren, W.A., Hilgen, F., Langereis, C., Kent, D.V., Obradovich, J., Raffi, I., Raymo, M.E., and Shackleton, N. 1995. Late Neogene chronology: New perspectives in high-resolution stratigraphy. *Geological Society of America Bulletin,* 107:1272-1287. https://doi.org/10.1130/0016-7606(1995)107<1272:LNCNPI>2.3.CO;2

Berggren, W.A., Kent, D.V., Flynn, J.J., and Van Couvering, J.A. 1985. Cenozoic geochronology. *Geological Society of America Bulletin*, 96:1407-1418. https://doi.org/10.1130/0016-7606(1985)96<1407:CG>2.0.CO;2

Bohling, G.C. 2005. Chronos age-depth plot: A java application for stratigraphic data analysis. *Geosphere*, 1:78-84. https://doi.org/10.1130/GES00009.1

Bolli, H.M., Saunders, J.B., and Perch-Nielsen K. 1985. *Plankton Stratigraphy*. Cambridge University Press, Cambridge.

Boyden, J.A., Müller, R.D., Gurnis, M., Torsvik, T.H., Clark, J.A., Turner, M., Ivey-Law, H., Watson, R.J., and Cannon, J.S. 2011. Next-generation plate-tectonic reconstructions using GPlates, p. 95-113. In Keller, G.R. and Baru, C. (eds.), *Geoinformatics: Cyberinfrastructure for the Solid Earth Sciences*. Cambridge University Press, Cambridge.

Cande, S.C. and Kent, D.V. 1995. Revised calibration of the geomagnetic polarity timescale for the late Cretaceous and Cenozoic. *Journal of Geophysical Research: Solid Earth*, 100:6093–6095. https://doi.org/10.1029/94JB03098

Cermeño, P., Falkowski, P.G., Romero, O.E., Schaller, M.F., and Vallina, S.M. 2015. Continental erosion and the Cenozoic rise of marine diatoms. *Proceedings of the National Academy of Sciences,* 112:4239-4244. https://doi.org/10.1073/pnas.1412883112

Chronos Workshop Members. 2002. *Integration of Chronostratigraphic Databases for the 21st Century.* Unpublished report, Amherst, MA.

Django Software Foundation. 2013. Django Web Framework, version 1.5.2. https://www.djangoproject.com/

Dunn, R. 2014. wxPython, version 3.0. https://wxpython.org/

Fan, J., Chen, Q., Hou, X., Miller, A.I., Melchin, M.J., Shen, S., Wu, S., Goldman, D., Mitchell, C.E., Yang, Q., Zhang, Y., Zhan, R., Wang, J., Leng, Q., Zhang, H., and Zhang, L. 2013. Geobiodiversity database: a comprehensive section-based integration of stratigraphic and paleontological data. *Newsletters on Stratigraphy,* 46:111-136. https://doi.org/10.1127/0078-0421/2013/0033

Fils, D., Cervato, C., Reed, J., Diver, P., Tang, X., Bohling, G., and Greer, D. 2009. CHRONOS architecture: Experiences with an open-source services-oriented architecture for geoinformatics. *Computers & Geosciences,* 35:774-782. https://doi.org/10.1016/j.cageo.2008.02.035

Fordham, B.G., Aze, T., Haller, C., Zehady, A.K., Pearson, P.N., Ogg, J. G., and Wade, B. S. 2018. Future-proofing the Cenozoic macroperforate planktonic foraminifera phylogeny of Aze & others (2011). *PLoS ONE,* 13:1-38. https://doi.org/10.1371/journal.pone.0204625

Gabella, C., Durinx, C., and Apple, R. 2018. Funding knowledgebases: towards a sustainable funding model for the UniProt use case. *F1000Research,* 6:2051. https://doi.org/10.12688/f1000research.12989.2

Gradstein, F.M., Ogg, J.G., Schmitz, M.D., and Ogg, G.M. 2012. *The Geologic Time Scale.* Elsevier, Amsterdam.

Gradstein, F.M., Ogg, J.G., and Smith, A.G. 2004. *A Geologic Time Scale 2004.* Cambridge University Press, Cambridge.

Hannisdal, B., Haaga, K.A., Reitan, T., Diego, D., and Liow, L.H. 2017. Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record. *Proceedings of the Royal Society of London B: Biological Sciences,* 284(1858). https://doi.org/10.1098/rspb.2017.0722

Hannisdal, B., Henderiks, J., and Liow, L.H. 2012. Long-term evolutionary and ecological responses of calcifying phytoplankton to changes in atmospheric CO_2. *Global Change Biology,* 18:3504-3516. https://doi.org/10.1111/gcb.12007

Hunter, J.D. 2007. Matplotlib: a 2D graphics environment. *Computing in Science and Engineering,* 9:90-95. https://doi.org/10.1109/MCSE.2007.55

Kemeny, J. G., Kurtz, T. E., and Elliott, B. 1985. *True BASIC: the Structured Language System for the Future.* Addison-Wesley, Reading.

Lazarus, D. 1992. Age Depth Plot and Age Maker: age modelling of stratigraphic sections on the Macintosh series of computers. *Geobyte,* 2:7-13.

Lazarus, D. 1994. Neptune: a marine micropaleontology database. *Mathematical Geology,* 26:817-832. https://doi.org/10.1007/BF02083119

Lazarus, D., Barron, J., Renaudie, J., Diver, P., and Türke, A. 2014. Cenozoic planktonic marine diatom diversity and correlation to climate change. *PloS ONE,* 9:e84857. https://doi.org/10.1371/journal.pone.0084857

Lazarus, D., Dolven, J.K., and Sanfilippo, A. 2011. Polycystinea. WoRMS: www.marinespecies.org.

Lazarus, D., Renaudie, J., Lenz, D., Diver, P., and Klump, J. 2018. Raritas: a program for counting high diversity categorical data with highly unequal abundances. *PeerJ,* 6:e5453. https://doi.org/10.7717/peerj.5453

Lazarus, D., Spencer-Cervato, C., Pika-Biolzi, M., Beckmann, J.P., von Salis, K., Hilbrecht, H., and Thierstein, H. 1995. Revised Chronology of Neogene DSDP holes from the World Ocean. *Ocean Drilling Program Technical Report,* 24: 301pp.

Lazarus, D., Suzuki, N., Caulet, J.-P., Nigrini, C., Goll, I., Goll, R., Dolven, J.K., Diver, P., and Sanfilippo, A. 2015. An evaluated list of Cenozic-Recent radiolarian species names (Polycystinea), based on those used in the DSDP, ODP and IODP deep-sea drilling programs. *Zootaxa,* 3999:301-333. https://doi.org/10.11646/zootaxa.3999.3.1

Lazarus, D., Weinkauf, M., and Diver, P. 2012. Pacman profiling: A simple procedure to identify stratigraphic outliers in high-density deep-sea microfossil data. *Paleobiology,* 38:144-161. https://doi.org/10.1666/10067.1

Lewitus, E. and Morlon, H. 2017. Detecting environment-dependent diversification from phylogenies: a simulation study and some empirical illustrations. *Systematic Biology,* 67(4):576-593. https://doi.org/10.1093/sysbio/syx095

Liow, L.H., Skaug, H.J., Ergon, T., and Schweder, T. 2010. Global occurrence trajectories of microfossils: Environmental volatility and the rise and fall of individual species. *Paleobiology,* 36:224-252. https://doi.org/10.1666/08080.1

Liow, L.H. and Stenseth, N.C. 2007. The rise and fall of species: Implications for macroevolutionary and macroecological studies. *Proceedings of the Royal Society of London B: Biological Sciences*, 274:2745-2752. https://doi.org/10.1098/rspb.2007.1006

Lloyd, G.T., Pearson, P.N., Young, J.R., and Smith, A.B. 2012. Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea. *Paleobiology*, 38:569-584. https://doi.org/10.1666/11041.1

Longhurst, A. 1998. *Ecological Geography of the Sea*. Academic Press, Cambridge.

McKinney, W. 2010. Data structures for statistical computing in python, p. 51-56. In van der Walt, S. and Millman, J. (eds.), *Proceedings of the 9th Python in Science Conference*.

Mithal, R. and Becker, D. G. 2006. The Janus database: providing worldwide access to ODP and IODP data. *Geological Society, London, Special Publications*, 267(1):253-259. https://doi.org/10.1144/GSL.SP.2006.267.01.19

Muttoni, G. and Kent, D.V. 2007. Widespread formation of cherts during the early Eocene climate optimum. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 253:348-362. https://doi.org/10.1016/j.palaeo.2007.06.008

Ogg, J.G., Ogg, G., and Gradstein, F.M. 2016. *A Concise Geologic Time Scale 2016*. Elsevier.

Pälike, H., Norris, R.D., Herrle, J.O., Wilson, P.A., Coxall, H.K., Lear, C.H., Shackleton, N.J., Tripati, A.K., and Wade, B.S. 2006. The heartbeat of the Oligocene climate system. *Science,* 314:1894-1898. https://doi.org/10.1126/science.1133822

Petersen, M., Glöckler, F., Kiessling, W., Döring, M., Fichtmüller, D., Laphakorn, L., Baltruschat, B., and Hoffmann, J., 2018. History and development of ABCDEFG: a data standard for geosciences. *Fossil Record*, 21:47-53. https://doi.org/10.5194/fr-21-47-2018

PostgreSQL Global Development Group. 2009. PostgreSQL, version 8.4. http://www.postgreSQL.org/

Python Software Foundation. 2010. Python Language Reference, version 2.7. http://www.python.org/

Renaudie, J. and Lazarus, D.B. 2013. On the accuracy of paleodiversity reconstructions: A case study in Antarctic Neogene radiolarians. *Paleobiology,* 39:491-509. https://doi.org/10.1666/12016

Ribardière, L. 1985. 4th Dimension. Design reference manual. ACI Inc., Paris.

Rothenberg, J. 1995. Ensuring the longevity of digital documents. *Scientific American*, 272:42-47. https://doi.org/10.1038/scientificamerican0195-42

Sepkoski, J. J. 1976. Species diversity in the Phanerozoic: species-area effects. *Paleobiology*, 2(4):298-303. https://doi.org/10.1017/S0094837300004930

Seton, M., Müller, R., Zahirovic, S., Gaina, C., Torsvik, T., Shephard, G., Talsma, A., Gurnis, M., Turner, M., Maus, S., and Chandler, M. 2012. Global continental and ocean basin reconstructions since 200Ma. *Earth-Science Reviews*, 113:212-270. https://doi.org/10.1016/j.earscirev.2012.03.002

Spencer-Cervato, C. 1999. The Cenozoic deep sea microfossil record: explorations of the DSDP/ODP sample set using the Neptune database. *Palaeontologia Electronica,* 2(2):a13. https://doi.org/10.26879/99013 palaeo-electronica.org/content/2-2-neptune

Young, J.R., Bown, P.R., Wade, B.S., Pedder, B.E., Huber, B.T., and Lazarus, D.B. 2019. Mikrotax: developing a genuinely effective platform for palaeontological geoinformatics. *Acta Geologica Sinica,* 93:70-72. https://doi.org/10.1111/1755-6724.14249

SUPPLEMENTARY MATERIAL

Supplementary material contain the user manual for the software NSB_ADP_wx (described in the manuscript), along with a companion R package, NSBcompanion, mimicking the main functionalities of the website. Files are zipped and available at https://palaeo-electronica.org/content/2020/2966-the-nsb-database.