One sentence summary: Paleobiological data systems are highly diverse, widely used, immensely valuable, and are a relatively low-cost solution to sustain hundreds of years and billions of dollars of scientific discovery and building a climate-smart future.

Fossils for Future: the billion-dollar case for paleontology's digital infrastructure

Dowding, EM[1]*; Dunne, EM[1]*; Collins, KS[2]. Cryer, K[1]; De Baets, K[3]; Dimitrijević, D[1]; Edie, SM[4]; Finnegan, S[5,6]; Kiessling, W[1]; Lintulaakso, K[7]; Liow, LH[8]; Little, H[4,] Na, Lin[9]; Peters, SE[10]; Renaudie, J[11]; Saupe, EE[12]; Seuss, B[1]., Sessa, JA[13]; Smith, JA[14]; Uhen, MD[15]; Williams, JW[16], Kocsis, ÁT[1]*

1. GeoZentrum Nordbayern, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
2. Natural History Museum, London, England
3. Institute of Evolutionary Biology, University of Warsaw, Warsaw, Poland
4. Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington DC, United States of America
5. Department of Integrative Biology & Museum of Paleontology, University of California, Berkeley,California, United States of America
6. Smithsonian Tropical Research Institute, Panama
7. Natural Sciences Unit, Finnish Museum of Natural History, Helsinki, Finland
8. Natural History Museum and Centre for Planetary Habitability, Department of Geosciences, University of Oslo, Oslo, Norway
9. Nanjing Institute of Geology and Palaeontology, Nanjing, China
10. Department of Geoscience, University of Wisconsin-Madison, Madison, WI, US

11. Museum für Naturkunde, Leibniz-Institut für Evolutions-und Biodiversitätsforschung, Berlin, Germany
12. Department of Earth Sciences, University of Oxford, Oxford, United Kingdom
13. Academy of Natural Sciences of Drexel University, Philadelphia, United States of America
14. Department of Earth and Environmental Sciences, University of Minnesota Duluth, Duluth, Minnesota, U.S.A.
15. Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia, United States of America
16. Department of Geography, University of Wisconsin-Madison, Madison, WI, US

* Corresponding authors

Abstract

The digital revolution has transformed paleontology through the development of open-access, community-driven databases that underpin some of the most impactful research in biodiversity, climate change, and extinction dynamics. These systems safeguard high-effort, volunteered data and have revealed major macroevolutionary patterns, including mass extinctions. However, of 118 paleontological and Earth science databases reviewed, 95% had lifespans under 15 years, putting decades of investment at risk. As paleontological data infrastructures enter a third generation—marked by modular design, improved data provenance, and cross-platform integration—there is growing potential to support multi-scalar, interdisciplinary research across Earth and Life sciences. We advocate for strategies to enhance database longevity, including sustained funding models, stronger institutional support, and modular backend architectures that better link international community databases to each other and to fossil specimens.

1. Introduction

The study of the history of life on Earth is inherently multidisciplinary and conducted at scales from local to global. This scientific inquiry draws from geology, biology, chemistry, archaeology, and mathematics, amongst others, to reconstruct ancient ecosystems, investigate the drivers of biodiversity, and forecast how life will respond to today's changing environments (Dietl & Flessa, 2011; Dillon et al., 2022; Kiessling et al., 2023). The fossil record is essential for understanding biodiversity and Earth system processes operating at timescales beyond the 20th- and 21st-century window of instrumental observations and provides examples of past Earth system states with instructive analogies to the societally novel climates that are now emerging (Burke et al., 2018; Pandolfi et al., 2020). From their very beginning, paleontological databases (See Glossary) played pivotal roles in enabling the field to scale up from site-level studies to global-scale research, laying the groundwork for influential research such as identifying mass extinctions and their roles in macroevolution (Raup & Sepkoski, 1982) and the earliest evidence of climate-driven species range shifts and ecosystem transformations (Davis, 1976; Bernabo and Webb, 1977). The subsequent migration of paleontological databases to online platforms and data systems–*encompassing the database, its system for community governance and data curation, and any associated software services*– increased their accessibility and amplified their impact by enabling broader collaboration and reproducibility (Williams et al. 2018; Uhen et al. 2013).

Today, openly accessible, community-run data systems function as collective repositories for scientific data and knowledge, providing the means for quantitative analyses of the history of life on Earth (Figure 1; Guo, 2017). These databases are invaluable for reconstructing ancient ecosystems (e.g., Cribb & Darroch, 2024), tracing evolutionary pathways (e.g., Alroy et al, 2008), studying climate- and human-driven eco-evolutionary dynamics at continental to global scales (Pandolfi et al. 2020, Mottl et al. 2021, Lang et al. 2023, Gordon et al. 2024), and predicting future biological and geological changes - or assessing the limits to

predictability in an increasingly novel world (e.g., Fitzpatrick et al. 2018; Stern & Gerya, 2023). By integrating these paleontological databases with other open data systems, Earth system scientists can tackle increasingly complex, multifaceted questions that are top priorities in global change research (National Academies of Sciences, 2018; Wang et al., 2021; Kiessling et al., 2023).

Representing developers, leaders, curators, and users of 15 community-run paleontological databases (see Supplementary; Table 2), we examine the current data landscape to assess the volume, variety, and value of data held in community-curated, open access databases, the challenges faced by these databases, and the opportunities for sustainable growth and scientific discovery. Focussing on Earth Science and paleontological databases and systems, we examine diversity dynamics within our shared data landscape to build a roadmap toward sustainable funding, and provide recommendations for continued researcher, maintainer, developer, and funder investment.

2. Paleontological data and databases: An Overview and History

2.1. Key Concepts

Paleontology aims to reconstruct the history of life across the broadest possible range of spatiotemporal scales and throughout the geological record (Figure 1). Here paleontology encompasses closely related fields including but not limited to paleobiology, biostratigraphy, and paleoecology. As our understanding of geological processes evolves, new scientific questions emerge, and our interpretation of the fossil record is updated. This, in turn, affects our understanding of the processes that we infer from it and drives new primary-data collection campaigns (e.g. fieldwork) and the reinterpretation and reanalysis of existing data. Examples include taxonomic reidentification of old fossils following new finds (Godfrey & Collareta, 2022), the re-dating of core samples and refinement of the geological time scale using newer and improved methods and data (e.g. Niebuhr & Wilmsen, 2023),

re-interpretation of the environmental/depositional context (e.g. Stiles et al, 2022), the incorporation of paleobiogeographic patterns into tectonic models (e.g. Torsvik et al., 2025), and the development of new analytical methods to quantify ecosystem rates of change (Mottl et al. 2021).

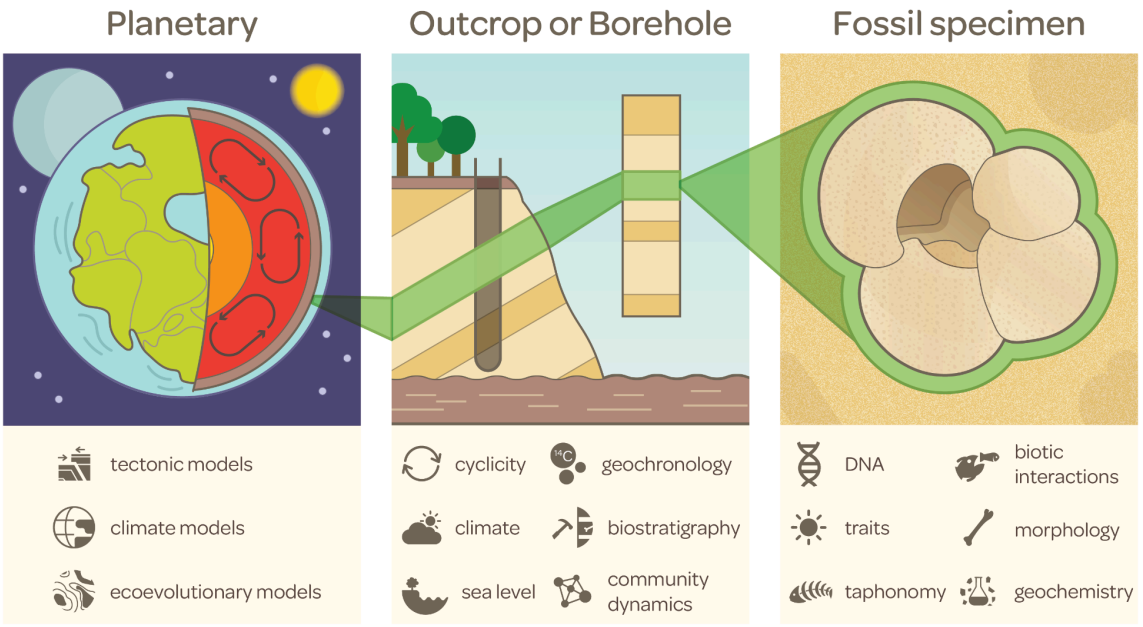

Figure 1. *Paleontological information in an Earth system context. From left to right, planetary or global-level information can be used to understand tectonic processes, climate and landscape evolution, and eco-evolutionary processes across timescales ranging to billions of years. Outcrop- or borehole-level data represent local- to regional-scale time series that can be used to reconstruct climate, geochronology (age), sea level fluctuations, and community dynamics. Finally, specimen-level data is the foundational unit in paleobiology for analyses of, e.g. taxonomy, biotic interactions, geochemistry, functional ecology, and taphonomic processes.*

Paleontologists work with two primary forms of data: 'fundamental data' and 'processed data'. 'Fundamental data' are direct observations and sampling of the sedimentary record and fossil specimens within these sediments. Examples of *fundamental data* include geospatial locations, physical samples, multimedia recording, counts, and geochemical analysis. When these *fundamental data* are subject to further interpretation, such as through taxonomic study and analyses of morphology, preservation, and biotic associations, they are translated into *processed data*. For example within database structures, age controls (e.g.

radiocarbon dates) are *fundamental data* and age-depth models (used to estimate the age of different depths within a sediment core or stratigraphic profile) are *processed data* and are frequently revised. Although fundamental and processed data exist on a continuum, whenever possible, paleontological databases should maintain the strongest links to the evidence and fundamental data. This foundation functions to provide evidence provenance and ensure against corollary risks when databases (e.g. Neptune Sandbox; NSB) are used as data sources for other, secondary databases (e.g. BioDeepTime; Smith et al, 2023a).

2.2. Database development history

2.2.1. The past: First-generation databases

First-generation compilations of paleontological data focussed on the collation of processed data, such as the inferred temporal (i.e. stratigraphic) distribution of fossil taxa using harmonised taxonomic lists across sites, which are the minimum requirement for assessing the history of biodiversity (e.g. Phillips, 1860; Sepkoski, 1982 and the shifting distribution of taxa across space and time (Bernabo and Webb, 1981, Huntley and Birks, 1983). These were initially collated as physical repositories (e.g. the John Williams Index of Paleopalynology; Riding et al. 2012) or as offline digital entities (e.g. Sepkoski's Compendium, Sepkoski, 2002, and the first version of the Neptune database; Lazarus, 1994). These first-generation databases were often built either by individual scientists over their careers or within small research teams.

2.2.2. The present: Second-generation data systems

As the field advanced, paleontologists gained further understanding of the various factors that distort the structure of the fossil record (e.g., Raup, 1972), new research questions emerged (e.g., reconstruction of past biomes and terrestrial carbon sequestration; Prentice et al. 1993), and paleontologists developed new skills and methods in large-scale data analytics and quantitative methods to address emergent questions. This, in turn, led to new

efforts to reanalyze existing databases. For example, in deep-time biodiversity analytics, the field progressed from recording inferred first- and last-appearance dates of taxa as endpoints of their existence (Sepkoski 2002) to the recording of occurrences from the entire stratigraphic record of the taxa (e.g., Alroy et al, 2001, 2008). The second-generation data systems (e.g.,Neotoma; Williams et al 2018) incorporated storage of multiple kinds of fundamental data, including the geographic coordinates of fossil sites, taxon abundance and traits, depths in boreholes, and lithological characteristics, amongst others.

In parallel, the leadership and development of these databases increasingly shifted from a few individual experts to community-governed data systems. For example, in Quaternary palynology, individual efforts to build databases and map continental-scale plant distributions for North America and Europe (Bernabo and Webb, 1981, Huntley and Birks 1983) expanded to continental-scale databases around the world, each with their own data leaders and stewards (Grimm et al. 2024). Multiple paleontological data systems incorporated well-developed community governance systems, such as leadership councils and experts charged with stewarding and curating specific kinds of data (Williams et al. 2018, Uhen et al., 2023).

The data structures of current, second-generation databases vary significantly, reflecting their founding aims and user communities. As examples, the Paleobiology Database (PBDB) was originally developed to investigate Phanerozoic diversity change (Alroy, 2008), NOW (New and Old Worlds database of fossil mammals) focused on Cenozoic mammal macroevolution (Jernvall and Fortelius 2002), Neotoma was designed to study species range shifts during the Quaternary glacial-interglacial cycles across multiple taxonomic groups (Grimm et al., 2024), and the Geobiodiversity Database (GBDB) was specifically designed to handle high-resolution stratigraphic data by linking fossil occurrences to detailed geological sections (Fan et al., 2014). Others, like BioDeepTime (Smith et al., 2023a), aggregated and standardized time series data from other databases, while integration-focused platforms

such as Deep-Time Digital Earth facilitated the synthesis of massive, diverse datasets and data types across paleontology and geoscience (Wang et al., 2021).

All these databases continue to grow in scope and incorporate new kinds of data. As new questions emerged and with the diversification and increased accessibility of data (Ross-Hellauer et al., 2022; Smith et al., 2023b), the range of scientific applications of these second-generation databases far surpass their original scope and yield input for thousands of scientific studies (Supplementary: ‘Database Use’). For example, PBDB occurrence data have been used for climatic modelling (Marcilly et al, 2021), landscape evolution (Fernandes et al, 2019), and paleogeographic models (Cao et al, 2017, Torsvik et al., 2025). Similarly, NOW data have been used to study macroevolutionary expansion (Žliobaitė, 2024) and Neotoma for reconstructing past climates (Chevalier et al. 2020), constraining past land cover dynamics and the terrestrial carbon cycle (Blarquez et al. 2015), and documenting cross-continental species invasions (Alverson et al. 2021). The scientific utility and applications of these databases thus continue to grow and diversify, as do the databases themselves.

2.3. The near future: From databases to third-generation data systems

Paleontology is poised for its next transformative phase, in which second-generation databases will be better integrated with each other and with other components of the paleontological, Earth and life sciences data infrastructures (Fig. 3), to address more integrative, cross-disciplinary, and multi-scalar questions.The transition to the third generation systems is already begun with cross database integration a focus in backend-development, with improved efficiency offered by modular design waiting to be capitalised upon (Deng and Li, 2013; Kaufman et al 2018; Deng et al 2024). Examples of research topics that can be advanced through improved integration of paleontological data

within and across scientific disciplines are described in the *Big Questions Projec*t (Smith et al. 2025). Representative questions include "*How do external environmental drivers (e.g., plate tectonics, global temperature, sea level) influence the structure of biological systems at different spatiotemporal scales?"*, " *How does the prevailing climate state experienced by species and communities influence their response to perturbation?"* and "*To what extent are the phases of events (e.g., collapse, recovery) during extinctions consistent across different biotic crises?"* Addressing these integrative questions requires scalable, connected data that captures, for example, phenotypic variation among individuals in a population, in conjunction with high stratigraphic resolution, paleoenvironmental, and specimen-level information. These scientific needs demand further advances in how paleontological data are reported, structured, integrated, managed, and sustained. Cross-institutional aggregation of museum collection specimen information into iDigBio (Nelson & Paul 2019) and GBIF (Telenius, 2011) are excellent examples that are made available by biodiversity data standards, such as the Darwin Core (Wieczorek et al. 2012) and ABCDEFG (Petersen et al. 2018), featuring a growing scope of associated semi-structured metadata (Hardisty et al. 2022).

The development of integrative platforms, such as Deep-Time Digital Earth (Wang et al. 2021) and the continued growth of existing databases to support new data types such as ancient environmental DNA (Williams et al. 2023), are striking movements towards third-generation databases. Careers of an entire generation of scientists are now based on access to open, interoperable data (Koch et al. 2018; Li et al 2023). At the same time, new concerns have arisen about whether these databases encode and perpetuate past and present inequities and how best to reduce these inequities to truly fulfill the deeper mission of these databases to ensure democratized data access for all (Rolin 2015; Monarrez et al. 2022, Raja et al. 2022).

Paleontological databases are also changing to help advance the new data standards that have emerged in the open sciences. FAIR data principles (Findability, Accessibility,

Interoperability, and Reusability; Jacobsen et al., 2020) have rapidly become a cornerstone of open-data policy, guiding how data should be shared, structured, accessed, and reused. The TRUST principles (Transparency, Responsibility, User focus, Sustainability, and Technology; Lin et al., 2020) provide best-practice guidance for how digital repositories can establish long-lasting relations of trust with their user communities. Complementing the FAIR principles, the CARE Principles for Indigenous Data Governance (Collective benefit, Authority to control, Responsibility, and Ethics; Carroll et al. 2020, Jennings et al. 2025) encourage open data movements to prioritise co-design with Indigenous Peoples and collective benefits. With increasing community awareness and utilization, improved data standards are beginning to be applied to paleontological data. FAIR and TRUST principles are currently being implemented by the paleo-community, and CARE principles are starting to gain traction (Dunne et al., 2025). Despite growing awareness and discussion of these principles, their implementation remains uneven among research institutions, publication protocols, or funding application frameworks and more work is needed to align the principles and their implementation. Finally, community governance by scientific experts of paleontological data is particularly critical for both TRUST and reproducibility, because of the many steps involved to make precise and accurate inferences about past biodiversity dynamics from fossil data (Boldgiv et al. 2025). By supporting those who create and curate palaeontological data to govern its use, more effective, context-sensitive strategies can be developed to address issues of access, provenance, and data equity (e.g., Lendermer et al, 2020; Wang et al, 2021; Sterner et al, 2023; Hurst et al, 2025).

3. Landscape Survey: The Current State of Paleodata

3.1. Data collection and analysis

An online meta-analysis of available paleo- and Earth science databases was conducted using search terms in multiple languages (Supplementary Table 4). Between November

2024 and March 2025, academic journals, data repositories (e.g. Zenodo, Dryad), reference lists, and aggregators (Google Scholar, Web of Science) were searched for records of relevant *Community-run* and *Open Access* databases. Community-run databases were required to not be attached to a state governing body including state-funded museums or geological survey, and are considered *Open Access* by the requirement of having the data free for general use. The period of activity was identified by the first publication of the database in the peer-reviewed scientific literature and its endpoint was identified through the last update to the web service, data repository, and/or latest published article. Of the 171 paleontological and Earth systems databases that were identified, 118 were tagged as Open Access and community-run. Using these 118 databases, we analysed their extinction rates, origination rates, and diversity dynamics using 'divDyn' (Kocsis et al., 2019; Figure 2). The per capita extinction and origination rates were analyzed using a rolling mean of year-to-year database activity, whilst the sampled-in-bin diversity used an extended decadal time series to account for boundary conditions.

We also assessed the replacement value of the data stored in three databases (Paleobiology Database (PBDB), GeoBiodiversity Database (GBDB), Neotoma Paleoecology Database (Neotoma). These three were selected based on their longevity and the access provided by the database maintainers to their 2024–2025 records (see Supplementary). The replacement value of the individual sample records (data measured on a fossil specimen or other physical sample) and collection records (sites; a group of fossil occurrences that are geographically and stratigraphically connected) were calculated in USD. Following the valuation procedure of Thomer et al. (2025), conservative sample and collection values were determined by calculating the cost of data replacement, including sample preparation and analysis, fieldwork, labor ($150 per sample record; $3000 per collection record). These values represent an estimated average across occurrence, core, and time-series data amongst others. It should be noted that some records may be

cheaper or far more expensive than these estimates depending on collection procedures and analysis. The valuation also assumes that the sites that host the published data are still accessible, i.e. not destroyed by human land-use or natural processes such as earthquakes. The data from inaccessible locations are therefore irreplaceable and priceless.

Research effort, storage, maintenance, curation, and expertise were not calculated, resulting in conservative values that do not cover the entire cost to replace the extant data nor do they cover the significant costs in labour, server hosting, and infrastructure development that go into setting up and sustaining databases. Further the results do not cover the article processing costs to publish a scientific paper (mean $2300 USD; Pinfield et al. 2016) nor the value of papers published (estimated at over $5000 USD per item; Rousseau et al. 2021).

Table 1. *The value of the samples and collections (sites) stored within three active paleo-databases in USD. Conservative value estimates are taken from the valuation framework of Thomer et al. (2025) and do not include collection and curation labour, storage, development, maintenance, and institutional overhead, which are collectively more than double the presented estimates. The original valuation of Neotoma in Thomer et al. (2025) has been expanded to include the Paleobiology Database (PBDB) and Geobiodiversity database (GBDB). Samples refer to individual records, for example species occurrence in the PBDB. A collection refers to a grouping of samples for example in a geographical site such as an outcrop in GBDB, or field location in Neotoma.*

DB	Samples (n)	Collections (n)	Sample Value ($)	Collections Value ($)	Total ($USD)
PBDB	1,653,699	240,405	248,054,850	721,215,000	
GBDB	580,049	217,969	87,007,350	653,907,000	
Neotoma	12,281,094	25,168	1,842,164,100	75,504,000	
			2,177,226,300	1,450,626,000	**$3,627,852,300**

3.2. Historical trends

Based on our web search, database origination rates peaked in the 1970s and 1990s, with a tertiary peak in the 2010s (Figure 2). Nearly 50% of our corpus of databases (*n* =118) became inactive within just five years, and fewer than 15% survived a full decade. Only a rare 5% remained active for over 15 years (Figure 2). This five-year timing coincides with the standard funding program of many large research grants, e.g. through the European Research Council or the National Science Foundation of the United States of America. This means that, after 5 years, up to 65% of value-added data effort, representing years of data aggregation, data harmonization and cleaning, technical development, and scientific labour, is left unmaintained and sometimes inaccessible.

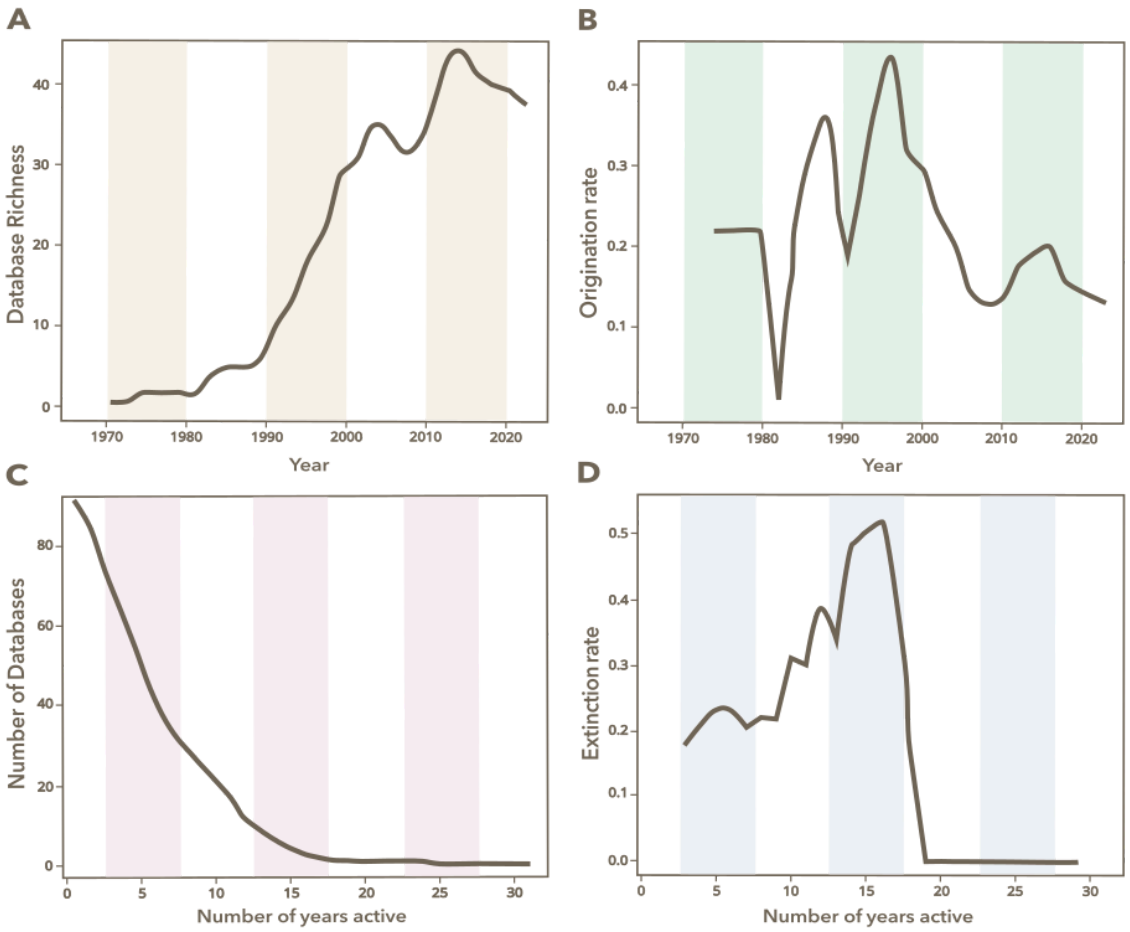

Figure 2*. Diversity dynamics of 118 community-developed paleontological databases (DB) from the 1970s to 2024. A. The range-through richness of databases by year,* ***B.*** *The origination rate of DBs through time, indicating areas of peak activity for novel DB development between 1995 and 2005.* ***C.*** *Diversity of DBs as a function of years active (i.e. database survivorship) showing the loss of >80% of*

DB diversity by 10 years of activity. ***D.*** *The rolling mean per capita extinction rate of DBs as a function of years active since inception with peaks at 5, 15, and 25 years of activity.*

Though some database development efforts are intended for short-term use and do not assume database longevity, the loss of these databases is not just a scientific concern, it also represents a significant economic waste (Figure 2; Table 1). The cost of allowing valuable data infrastructure to degrade is not theoretical but quantifiable and substantial. The best-case scenario for at-risk databases is integration with larger data systems, an example being the current integration of 34 constituent databases into Neotoma. Neotoma was launched in 2007 and maintains constituent databases that date back to the 1980s. The data protected and expanded by Neotoma was recently estimated to cost over $1.5 billion USD to replace (Thomer et al., 2025). This cost estimate is conservative, because it only represented data-replacement costs and did not account for estimates of time saved through ready data access and the cost of replacing software services. Despite this high valuation and proven utility to the community (e.g. Mottl et al. 2021, Adeleye et al. 2023, Li et al. 2023; Gordon et al. 2024), even long-lasting success stories such as Neotoma are at risk due to reliance on grant-based funding.

4. Towards the Third Generation of paleontological databases

We present here a series of actionable recommendations to address the existing structural and community challenges within the paleontological and Earth science data landscape (Table 2). To address data fragmentation and redundancy in databasing effort, the immediate priority is to maximize the value of existing services while laying the groundwork for long-term solutions.

4.1. Modular, interoperable data system

The scientific community and governing bodies (e.g. funders) must move away from the current trend of creating databases that are not interoperable. Instead, they must move toward those designed for integration, to break the cycle of effort and loss. While broader challenges around data infrastructure are often shaped by political and institutional forces beyond the control of individual researchers, the scientific community can take meaningful action through improved data practices (Deng et al, 2020; Ramachandran et al., 2021; Ross-Hellauer et al., 2022; Dunne et al., 2025). Examples like Neotoma, NOW, and the PBDB, which have remained active for over 15 years and continue to serve global communities across disciplines, demonstrate the efficacy and resilience of collaborative stewardship. However, databases and other resources like the *Biodiversity Heritage Library (https://www.biodiversitylibrary.org/),* are vulnerable to 'extinction', such as through cyberattacks, but more commonly due to funding termination. For example, recent attacks on the Museum für Naturkunde Berlin cyber infrastructure resulted in the loss of community access to Neptune Sandbox Berlin database (NSB). The NSB was intermittently funded (16 funded years since 1990) and maintained by an individual expert (see Supplementary, *Curator Review*), held hundreds of thousands of marine plankton microfossil species from hundreds of deep-sea ocean drilling sections, an invaluable resource that contributed to other databases including BioDeepTime (Smith et al. 2023a) Microtax (Huber et al. 2016), the GBDB (Fan et al. 2014), Triton (Fenton et al. 2021) amongst others. Through this attack, not only was a key resource for microfossil taxonomists, evolutionary (paleo)biologists, and paleoceanographers impacted, but the data provenance of the dependent databases has become compromised. The lack of funding and dedicated technical support resulted in a lack of failsafes at the museum. Instead, through community activity, external versions of the NSB, e.g., held through Zenodo (Renaudie et al. 2023) are contributing to database recovery, further highlighting the value of community contributions to sustaining data resources.

By prioritising interoperability, modularity, and long-term integration, we can build a resilient and pluralistic community of data systems that safeguard multiple dimensions of value of scientific data and ensure its continued relevance to scientists, external stakeholders and the general public (Sterner et al, 2023; Majeed and Hwang, 2024). To this end, we recommend the transition to a decentralised modular data network (Figure 3), where core components like those responsible for taxonomy, stratigraphy, and specimen provenance are built with a flexible scope. This would function, at first, following a movement from the fragmented and uncoordinated data landscape (Figure 3A) to pooled, pluralistic frameworks (Figure 3B; Sterner et al 2023). Pluralistic approaches to data pooling maintain domain independence and flexibility, permitting field-specific misalignment (e.g. the unit differences in terminology and grouping seen between core-based micropaleontology and global macrofossil biogeography in terms of spatial and temporal binning). Modules within this system serve as interlocking elements, offering researchers the basis to develop extension structures required to answer novel scientific questions within a broader, connected data landscape (Figure 3B). For example, to develop a novel database to answer questions about fossil biotic interactions (e.g. BITE; Huntley et al. 2023), a new data structure is required, developed specifically to tie a biotic interaction and the organisms to a rock-specimen. However outside of this novel database element, core elements such as taxonomy, stratigraphy, and geography could be downloaded from *pro forma* modules from the idealised framework (Figure 3B), meaning the only new element to be constructed is the one that specifically captures biotic interaction data. This approach saves time on database construction, ensures interoperability, and safeguards against database loss. The suggested solution mimics the general tendency of corporations that move from large monolithic applications to microservices to meet demands of scalability and a fast development cycle (Ponce et al., 2019l Majed and Hwang, 2024), and is particularly suited to scientific research that is globally distributed in nature (Deng and Li 2013; Kaufman et al, 2018; Deng et al 2024).

To realize their full potential, databases must, whenever possible, maintain direct links to physical specimens and samples (e.g. through the International Generic Sample Numbers, [https://ev.igsn.org/]), users (e.g. persistent identifier through ORCID [https://orcid.org/]), usage (e.g. DATACITE for DOI mining [https://datacite.org/]), and improve linkages to other databases (see Section 2.3). Museums, research institutes, and public collections are a foundation of this system, providing crucial metadata that ties scientific conclusions to real-world evidence (see Johnson and Owens, 2023; Boldgiv et al. 2025). Increasing connection to specimen evidence provides the most powerful opportunity to ensure that digital records remain verifiable, reproducible, and scientifically robust (Schriml et al 2020). Strengthening the connections between physical specimens and their digital representations will support the long-term sustainability of databases, facilitate interdisciplinary research, and enable the next generation of large-scale paleontological analyses (DeMiguel et al., 2021; McManimon & Natala, 2021; Deng and Li, 2013). By reinforcing these links, we can ensure that databases remain powerful tools for discovery while upholding scientific transparency and rigor.

Developing Application Programming Interfaces (APIs), which enable one software program to request services or data from another without needing to know the internal workings of the other system, that adhere to Open Science standards is crucial for ensuring seamless exchange of information between data systems, regardless of their underlying technologies. Additionally, employing data harmonization tools (Grenié et al., 2023) can streamline the integration process by automatically reconciling differences in data formats, units of measurement, and terminologies. For example, the *fossilpoll* workflow (Flantua et al, 2023; hope-uib-bio.github.io/FOSSILPOL-website/en/index.html) pulls data from Neotoma, harmonizes the age-depth models, and builds harmonized taxonomic names lists. These workflows create opportunities to distribute effort and have scientists outside the database leaders/curators add value while still provenancing back to the databases. Further, such tools can leverage machine learning algorithms to identify and merge duplicate records,

standardize taxonomic names, and align stratigraphic information, potentially reducing the manual effort required for data integration. Similar to data stewardship practices within Neotoma (Grimm et al, 2018), these automated processes need to undergo expert validation to ensure accuracy and reliability. As a general rule, because of the complexity of fossil data and the implicit knowledge often embedded within paleontological datasets, we recommend that analytical and curatorial workflows employ human-in-the-loop approaches, rather than fully automated systems to avoid ‘garbage in-garbage out’ situations and the influence of programmer biases (Bircan and Özbilgin, 2025).

4.2. Financial support

This perspective by developers and maintainers showcases the vast and diverse opportunities offered by well-curated and long-lasting paleontological data systems. At the same time, we highlight the risks and challenges facing fundamental data resources in paleontology and elsewhere (Thomer et al. 2025), as well as the strategies needed to secure their future for the benefit of all science (Smith et al, 2023b, 2025). In this context, using paleontological data as an example, we propose a path forward for sustainable development, funding, and stewardship to safeguard community-built scientific data systems for future generations. Whilst we focus here on open digital resources for the democratization of science, the investment in such resources must come with better linkages to, and explicit support for, museums and physical repositories (Allmon et al., 2018; Marshall et al., 2018; McManimon & Natala, 2021; Dunne et al., 2025).

Databases have been developed and maintained through a combination of funding and unfunded volunteer/service work (Thomer et al, 2025). The persistence of these databases through all this financial precarity is a testament to their importance and the work of many scientists to keep them going. Investing in sustainable, modular data infrastructure not only enhances the longevity, accessibility, and utility of scientific data, but also protects the

immense financial and intellectual investment already made. Funding is essential for ensuring that community-curated data continue to inform cutting-edge science well into the future.

Table 2. *A roadmap to sustainable funding*

Action	Description
Embed sustainability from inception	Design databases with modular architecture and interoperability in mind. This enables future integration into broader infrastructures and reduces redundancy, lowering long-term maintenance costs
Establish core infrastructure grants	Advocate for dedicated infrastructure funding schemes, distinct from research project grants, that support long-term maintenance, technical upgrades, and data curation
Develop cross-sector partnerships	Collaborate with museums, universities, government agencies, and industry partners to co-invest in shared data resources
Quantify and communicate value	Systematically assess the scientific and economic value of databases, as done for Neotoma (~$1.5 billion USD), to demonstrate return on investment and attract strategic funding
Adopt attribution standards	Promote data citation, DOI assignment, and recognition mechanisms to incentivise community data contributions and support funding applications that highlight demonstrable use
Foster community governance	Create steering bodies or consortiums to coordinate long-term strategy, technical development, and funding pipelines across institutions and borders

Besides optimizing the use of already acquired funding, long-term sustainability hinges on moving beyond short-term, project-driven funding models (Tables 2 and 3). Advocating for

policy support at institutional, national, and international levels is required to create an enabling environment for these systems to thrive. Network-level integration provides a means to ensure continued relevance, usability, and return on investment beyond the end of a research project's funding cycle (Thomer et al, 2025).

Engaging policymakers and funding agencies in discussions about the importance of Earth science and paleontological community data networks can help secure the necessary support and resources (e.g., the USA's Geoscience Congressional Visit days). Core infrastructure funding, akin to utilities for the scientific community, should be secured through national and international bodies, ensuring that databases are treated as essential research infrastructure (e.g., the German NFDI initiative [nfdi.de] with NFDI for Earth [nfdi4earth.de] the Chinese National Natural Science Fund Key Basic Research Infrastructure program, [nsfc.gov.cn/english/site_1/funding/E1/2024/06-12/364.html], which support geo-data infrastructure). Within our proposed funding roadmap (Table 2), we recommend demonstrating the economic, societal, and scientific value of open data through public–private partnerships and cost–benefit analyses, approaches already proven effective in initiatives like Neotoma. Ultimately, we wish to see the establishment of a dedicated international non-profit organization, akin to CERN, to be responsible for the financial sustainability of the geological data landscape.

4.3. Community governance and goals

We propose a phased, community-guided transition toward a sustainable, specimen-based, and explicitly modular data infrastructure—one that is grounded in the principles of FAIR, CARE, and TRUST, and ensures proper attribution (Lin et al., 2020; Carroll et al., 2020; Jacobsen et al., 2020; Jennings et al., 2023). As artificial intelligence, large-scale web scraping, and automated data aggregation become increasingly common tools, the paleontological community must actively shape how its openly accessible data are structured, cited, and reused. A modular and well-governed framework will allow us to

respond nimbly to these technological developments while preserving the integrity and provenance of our data. Central to this vision is strong, inclusive community governance—led by the researchers, data stewards, and institutions who know the data and needs of the researchers best. By harmonising efforts and redistributing responsibilities through open consultation, we can build an equitable and future-ready infrastructure that supports both innovation and accountability in paleoscience.

***Table 3**. Recommendations for the sustainable development of community-developed data resources and the related benefits derived from their implementation. Where the benefits are* Rigor and Reliability *(1),* Ability to address new Questions *(2),* Faster and more Inclusive Dissemination of Knowledge *(3),* Broader Participation in Research *(4),* Effective use of Resources *(5),* Improved Performance Research Task*s (6) and* Open Publication for Public Benefit *(7; see supplementary Table 3 for expansion and descriptions).*

Recommendation	Details	Benefits
A Incentivise data contributions	Create systems (and a scientific culture) for increased acknowledgement, attribution, and citation for data contributions.	4, 5, 6
B Establish a framework for data integration	Develop a standardized framework for integrating diverse Earth System databases, ensuring interoperability and data quality transparency.	1, 2, 5, 6
C Secure sustainable funding	Advocate for dedicated funding streams to support the development, maintenance, and enhancement of modular data systems.	All
D Promote Open Science practices	Encourage the adoption of open science practices, including open data, Open Access publications, and collaborative research initiatives.	All
E Invest in technology and innovation	Leverage technological advancements to enhance data integration, analysis, and visualization capabilities.	1, 2, 6

F Build and foster global collaborations	International collaborations and partnerships create a comprehensive and diverse global network of paleontological data.	2, 4, 6
G Ensure ethical and legal compliance	Addressing ethical and legal considerations, including data privacy, security, and intellectual property rights, ensures responsible data management and sharing.	1, 4, 6, 7
H Advocate for policy support	Advocating for policy support at institutional, national, and international levels is required to create an enabling environment for these systems to thrive.	All

Promising steps are already underway. Initiatives like the ARC Centre of Excellence for Australian Biodiversity and Heritage (CABAH; epicaustralia.org.au) exemplify how community-led, transdisciplinary frameworks can successfully balance Indigenous knowledge systems, biodiversity and paleodiversity data, and open infrastructure. In 2023 CABAH produced 127 journal articles and welcomed over 60,000 attendees to their public programs and events. CABAH's approach is collaborative, bringing together researchers, Indigenous communities, industry, and policy partners. This momentum is furthered by ensuring that decisions around standards, attribution, and data validation are made through inclusive consultation with a broad cross-section of the community, including historically underrepresented groups and the global majority. Community buy-in for data attribution and validation will facilitate community confidence in Open Data resources.

True integration goes beyond technical aspects and requires active collaboration between scientists and technical experts from varied disciplines (Table 3, Figure 3). Establishing interdisciplinary data standards, training programs, research teams, and projects can facilitate this collaboration. Through this effort, we may be able to develop common research frameworks and questions that guide data integration efforts that can align the objectives of different disciplines (Rolin, 2015). For instance, questions about the impact of climate change on biodiversity through geological time can serve as a unifying framework for integrating paleontological, geological, and climatological data.

5. Conclusions

Paleontological data systems are critical resources for the advancement of Earth System research. By committing to the development and maintenance of decentralised, interconnected, modular data systems, we are capable of addressing pressing questions about our planet and creating a more interconnected scientific community. This effort is already well underway and following success stories like Neotoma, integrated support systems can protect and sustain our community-developed data resources. Together, these recommendations align structural reforms with scientific needs and community values. The path forward requires a collective effort, sustained funding, and a commitment to collaboration, ensuring that paleontological data remain useful resources for future generations.

Acknowledgements

The paper was written in the context of the 'Integrated Record of Ancient Life' (IRAL) working group. We thank the Paleosynthesis Project and the Volkswagen Stiftung for funding that supported this project (Az 96 796). Figures 1 and 3 designed by Miranta Kouvari and Nuria Melisa Morales Garcia from Science Graphic Design (sciencegraphicdesign.com).

Author Contributions

E.M.Do. designed the data survey, collected and analysed data, and led the writing. Á.T.K conceptualized the IRAL project, E.M.Du, and Á.T.K, secured funding, assembled the IRAL working group, and contributed significantly to framing and writing. K.C. assisted with data collection and analysis. J.W and J.A.S contributed significantly to editing and framing. All authors contributed data, for the databases and systems they maintained, and all were

involved in writing, editing, and framing the article. B.S. offered significant logistics and management support.

Supplemental information

GIT: https://github.com/dowdingem/IRAL

References

1. Adeleye, M. A., Haberle, S. G., Gallagher, R., Andrew, S. C., & Herbert, A. (2023). Changing plant functional diversity over the last 12,000 years provides perspectives for tracking future changes in vegetation communities. Nature Ecology & Evolution, 7(2), 224–235. https://doi.org/10.1038/s41559-022-01943-4
2. Allmon, W. D., Dietl, G. P., Hendricks, J. R., & Ross, R. M. (2018). Bridging the two fossil records: Paleontology's "big data" future resides in museum collections. In G. D. Rosenberg & R. M. Clary (Eds.), Museums at the Forefront of the History and Philosophy of Geology: History Made, History in the Making (pp. 35–44). Geological Society of America. https://doi.org/10.1130/2018.2535(03)
3. Alroy, J. (1992). Conjunction among taxonomic distributions and the Miocene mammalian biochronology of the Great Plains. Paleobiology, 18(3), 326–343. https://doi.org/10.1017/S0094837300010873
4. Alroy, J., Marshall, C.R., Bambach, R.K., Bezusko, K., Foote, M., Fürsich, F.T., Hansen, T.A., Holland, S.M., Ivany, L.C., Jablonski, D. & Jacobs, D.K., (2001). Effects of sampling standardization on estimates of Phanerozoic marine diversification. Proceedings of the National Academy of Sciences, 98(11), pp.6261-6266.
5. Alroy, J., Aberhan, M., Bottjer, D. J., Foote, M., Fürsich, F. T., Harries, P. J., ... & Visaggi, C. C. (2008). Phanerozoic trends in the global diversity of marine invertebrates. Science, 321(5885), 97-100.
6. Alverson, A.J., Chafin, T.K., Jones, K.A., Manoylov, K.M., Johnson, H., Julius, M.L., Nakov, T., Ruck, E.C., Theriot, E.C., Yeager, K.M., Stone, J.R., 2021. Microbial biogeography through the lens of exotic species: the recent introduction and spread of the freshwater diatom Discostella asterocostata in the United States. Biol Invasions. https://doi.org/10.1007/s10530-021-02497-5
7. Bernabo, J.C., Webb, I., T., 1977. Changing patterns in the Holocene pollen record of northeastern North America: A mapped summary. Quaternary Research 8, 64–96.
8. Bircan, T., & Özbilgin, M. F. (2025). Unmasking inequalities of the code: Disentangling the nexus of AI and inequality. Technological Forecasting and Social Change, 211, 123925.
9. Blarquez, O., Aleman, J.C., (2015). Tree biomass reconstruction shows no lag in postglacial afforestation of eastern Canada. Canadian Journal of Forest Research 46, 485–498. https://doi.org/10.1139/cjfr-2015-0201
10. Boldgiv, B., Lkhagva, A., Edwards, S., Stenseth, N.C., Bayarsaikhan, J., Altangerel, D., Usukhjargal, D., Dovchin, B., Gombobaatar, S., Batsaikhan, N., Warinner, C.,

Hart, I., Galbreath, K., Greiman, S.E., Malaney, J., Murdoch, J.D., McLean, B., DeWitte, S.N., Manzitto-Tripp E., Chin, K., Karim, T. S., Simpson, C., Stevens, N J., Dunnum, J. L., Cook, J. A. & Taylor W.T.T. (2025) Global natural history infrastructure requires international solidarity, support, and investment in local capacity, Proc. Natl. Acad. Sci. U.S.A. 122 (6) e2411232122, https://doi.org/10.1073/pnas.2411232122

    a. **This manuscript has analogous aims to the current manuscript, highlighting the value of coordinated but pluralist approaches to improving and protecting data infrastructure.**

11. Burke, K.D., Chandler, M., Haywood, A.M., Lunt, D.J., Otto-Bliesner, B.L., Williams, J.W., 2018. Pliocene and Eocene provide best analogues for near-future climates. Proceedings of the National Academy of Sciences, 115, 13288–13293. https://doi.org/10.1073/pnas.1809600115
12. Cao, W., Zahirovic, S., Flament, N., Williams, S., Golonka, J., & Müller, R. D. (2017). Improving global paleogeography since the late Paleozoic using paleobiology. Biogeosciences, 14(23), 5425–5439. https://doi.org/10.5194/bg-14-5425-2017
13. Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. Data Science Journal, 19, 43. https://doi.org/10.5334/dsj-2020-043
14. Chang, J., Cui, B., Nargesian, F., Asudeh, A., & Jagadish, H. V. (2024). Data distribution tailoring revisited: Cost-efficient integration of representative data. The VLDB Journal, 33(5), 1283–1306. https://doi.org/10.1007/s00778-024-00849-w
15. Chevalier, M., Davis, B.A.S., Heiri, O., Seppä, H., Chase, B.M., Gajewski, K., Lacourse, T., Telford, R.J., Finsinger, W., Guiot, J., Kühl, N., Maezumi, S.Y., Tipton, J.R., Carter, V.A., Brussel, T., Phelps, L.N., Dawson, A., Zanon, M., Vallé, F., Nolan, C., Mauri, A., de Vernal, A., Izumi, K., Holmström, L., Marsicek, J., Goring, S., Sommer, P.S., Chaput, M., Kupriyanov, D., 2020. Pollen-based climate reconstruction techniques for late Quaternary studies. Earth-Science Reviews. https://doi.org/10.1016/j.earscirev.2020.103384
16. Cribb, A. T., & Darroch, S. A. F. (2024). How to engineer a habitable planet: The rise of marine ecosystem engineers through the Phanerozoic. Palaeontology, 67(5), e12726. https://doi.org/10.1111/pala.12726
17. Davis, M.B., 1976. Pleistocene biogeography of temperate deciduous forests. Geoscience and Man XIII, 13–26.
18. DeMiguel, D., Brilha, J., Alegret, L., Arenillas, I., Arz, J. A., Gilabert, V., Strani, F., Valenciano, A., Villas, E., & Azanza, B. (2021). Linking geological heritage and

geoethics with a particular emphasis on palaeontological heritage: The new concept of ‘palaeontoethics.’ Geoheritage, 13(3), 69. https://doi.org/10.1007/s12371-021-00595-3

19. Deng, M., & Di, L. (2013). Building open environments to meet big data challenges in Earth sciences. In Big Data Techniques and Technologies in Geoinformatics (pp. 67–88). CRC Press.
    a. **The developers perspective on database development and infrastructure necessary for ensuring future proof, functional, financed, resources.**
20. Deng, Y., Song, S., Fan, J., Luo, M., Yao, L., Dong, S., Shi, Y., Zhang, L., Wang, Y., Xu, H., Xu, H., Zhao, Y., Pan, Z., Hou, Z., Li, X., Shen, B., Chen, X., Zhang, S., Wu, X., … Wang, E. (2024). Paleontology Knowledge Graph for Data-Driven Discovery. Journal of Earth Science, 35(3), 1024–1034. https://doi.org/10.1007/s12583-023-1943-9
21. Deng Y., Fan. J., Wang, Y. Shi, Yang. J., & Lu Z. (2020). Current Status of Paleontological Databases and Data-driven Research in Paleontology. Geological Journal of China Universities, 26(4), 361–383.
22. Dietl, G. P., & Flessa, K. W. (2011). Conservation paleobiology: Putting the dead to work. Trends in Ecology & Evolution, 26(1), 30–37. https://doi.org/10.1016/j.tree.2010.09.010
23. Dillon, E. M., Pier, J. Q., Smith, J. A., Raja, N. B., Dimitrijević, D., Austin, E. L., Cybulski, J. D., De Entrambasaguas, J., Durham, S. R., Grether, C. M., Haldar, H. S., Kocáková, K., Lin, C.-H., Mazzini, I., Mychajliw, A. M., Ollendorf, A. L., Pimiento, C., Regalado Fernández, O. R., Smith, I. E., & Dietl, G. P. (2022). What is conservation paleobiology? Tracking 20 years of research and development. Frontiers in Ecology and Evolution, 10, 1031483. https://doi.org/10.3389/fevo.2022.1031483
24. Dunne, E. M., Chattopadhyay, D., Dean, C. D., Dillon, E. M., Dowding, E. M., Godoy, P. L., Smith, J. A., & Raja, N. B. (2025). Data equity in paleobiology: Progress, challenges, and future outlook. Paleobiology, 51(1), 237–249. https://doi.org/10.1017/pab.2024.61
    a. **Summary of data equity issues in data access, data contribution, and data use and provides an expanded framework for addressing challenges.**
25. Fan, J., Hou, X., Chen, Q., Melchin, M. J., Goldman, D., Zhang, L., & Chen, Z. (2014). Geobiodiversity Database (GBDB) in stratigraphic, palaeontological and palaeogeographic research: graptolites as an example. Gff, 136(1), 70-74.

26. Fenton, I. S., Woodhouse, A., Aze, T., Lazarus, D., Renaudie, J., Dunhill, A. M., Young, J.R. & Saupe, E. E. (2021). Triton, a new species-level database of Cenozoic planktonic foraminiferal occurrences. Scientific Data, 8(1), 160.
27. Fernandes, V. M., Roberts, G. G., White, N., & Whittaker, A. C. (2019). Continental-Scale Landscape Evolution: A History of North American Topography. Journal of Geophysical Research Earth Surface, 124(11), 2689–2722. https://doi.org/10.1029/2018jf004979
28. Fitzpatrick, M.C., Blois, J.L., Williams, J.W., Nieto-Lugilde, D., Maguire, K.C., Lorenz, D.J., 2018. How will climate novelty influence ecological forecasts? Using the Quaternary to assess future reliability. Global Change Biology 24, 3575–3586. https://doi.org/10.1111/gcb.14138
29. Flantua, S. G., Mottl, O., Felde, V. A., Bhatta, K. P., Birks, H. H., Grytnes, J. A., ... & Birks, H. J. B. (2023). A guide to the processing and standardization of global palaeoecological data for large-scale syntheses using fossil pollen. Global Ecology and Biogeography, 32(8), 1377-1394.
30. Godfrey, S. J., & Collareta, A. (2022). A new ichnotaxonomic name for burrows in vertebrate coprolites from the Miocene Chesapeake Group of Maryland, U.S.A. Swiss Journal of Palaeontology, 141(1), 9. https://doi.org/10.1186/s13358-022-00250-6
31. Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G. M. L., Sagouis, A., & Winter, M. (2023). Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. Methods in Ecology and Evolution, 14(1), 12–25. https://doi.org/10.1111/2041-210X.13802
32. Grimm E. C., Blois, J.L., Giesecke, T., Graham, R.W., Smith A.J., & Williams J.W. (2018) Constituent databases and data stewards in the Neotoma Paleoecology Database: History, growth, and new directions. Past Global Changes Magazine 26(2) 64-65. https://doi.org/10.22498/pages.26.2.64
33. Grimm, E.C., Bradshaw, R.H.W., Brewer, S., Flantua, S., Giesecke, T., Goring, S., Lézine, A.M., Takahara, H., Williams, J.W., 2024. Databases and their application, in: Mock, C.J., Elias, S.A. (Eds.), Encyclopedia of Quaternary Science. Elsevier.
34. Guo, H. (2017). Big Earth data: A new frontier in Earth and information sciences. Big Earth Data, 1(1–2), 4–20. https://doi.org/10.1080/20964471.2017.1403062
35. Hardisty, A. R., Ellwood, E. R., Nelson, G., Zimkus, B., Buschbom, J., Addink, W., Rabeler, R. K., Bates, J., Bentley, A., Fortes, J. A., & others. (2022). Digital extended specimens: Enabling an extensible network of biodiversity data records as integrated digital objects on the internet. BioScience, 72(10), 978–987.

36. Huber, B. T., Petrizzo, M. R., Young, J. R., Falzoni, F., Gilardoni, S. E., Bown, P. R., & Wade, B. S. (2016). Pforams@microtax: A new online taxonomic database for planktonic foraminifera. Micropaleontology, 62(6), 429–438. https://www.jstor.org/stable/26645533
37. Huntley, B., Birks, H.J.B., 1983. An Atlas of Past and Present Pollen Maps for Europe: 0-13000 Years Ago. Cambridge University Press, Cambridge.
38. Huntley, J., Skawina, A., Dowding, E. M., Dentzien-Dias, P., De Baets, K., Kocsis, Á., Labandeira, C., Liow, L. H., Petsios, E., Smith, J., & Chattopadhyay, D. (2023). Biotic Interactions in Deep Time (bite): Developing a Specimen-Level Database to Address Fundamental Questions in Ecology and Evolution. Abstracts With Programs - Geological Society of America. https://doi.org/10.1130/abs/2023am-390127
39. Hurst, S., Moore, M. W., Simpson, A., Salisbury, S. W., Ahoy, S., Kitchener, C., & Betts, M. J. (2024). More than museums: care for natural and cultural heritage in Australia. Geoconservation Research, 7(2), 1-38. https://doi.org/10.57647/GCR-2024-SI-SY25
40. Jacobsen, A., De Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., … Schultes, E. (2020). FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence, 2(1–2), 10–29. https://doi.org/10.1162/dint_r_00024
41. Jennings, L., Anderson, T., Martinez, A., Sterling, R., Chavez, D. D., Garba, I., Hudson, M., Garrison, N. A., & Carroll, S. R. (2023). Applying the 'CARE Principles for Indigenous Data Governance' to ecology and biodiversity research. Nature Ecology & Evolution, 7(10), 1547–1551. https://doi.org/10.1038/s41559-023-02161-2
42. Jennings, L., Jones, K., Taitingfong, R., Martinez, A., David-Chavez, D., Alegado, R. 'Anolani, Tofighi-Niaki, A., Maldonado, J., Thomas, B., Dye, D., Weber, J., Spellman, K.V., Ketchum, S., Duerr, R., Johnson, N., Balch, J., Carroll, S.R., (2025). Governance of Indigenous data in open earth systems science. Nature Communications 16, 572. https://doi.org/10.1038/s41467-024-53480-2
43. Jernvall J, Fortelius M.(2002). Common mammals drive the evolutionary increase of hypsodonty in the Neogene. Nature 417, 538–540 https://doi.org/10.1038/417538a
44. Johnson, K. R., Owens, I. F., & Global Collection Group. (2023). A global approach for natural history museum collections. Science, 379(6638), 1192-1194.
45. Kaufman, D.S., Abram, N., Evans, M.N., Francus, P., Goosse, H., Linderholm, H., Loutre, M.F., Martrat, B., McGregor, H.V., Neukom, R., St George, S., Turney, C., von

Gunten, L., (2018). Open-paleo-data implementation pilot: The PAGES 2k special issue. Climate of the Past 14(5),593–600. https://doi.org/10.5194/cp-14-593-2018

46. Kiessling, W., Smith, J. A., & Raja, N. B. (2023). Improving the relevance of paleontology to climate change policy. Proceedings of the National Academy of Sciences, 120(7), e2201926119. https://doi.org/10.1073/pnas.2201926119
    a. **Highlights the specific and useful contributions of palaeontology and quantitative analysis of the fossil record for modern environmental challenges.**
47. Koch, A., Glover, K.C., Zambri, B., Thomas, E.K., Benito, X., Yang, J.Z., 2018. Open-data practices and challenges among early-career paleo-researchers. PAGES Magazine 26, 54–55.
48. Kocsis, Á. T., Reddin, C. J., Alroy, J., & Kiessling, W. (2019). The r package divDyn for quantifying diversity dynamics using fossil sampling data. Methods in Ecology and Evolution, 10(5), 735–743. https://doi.org/10.1111/2041-210X.13161
49. Lang, G., Ammann, B.,van der Knaap, W.O., Morales-Molino, C, Schwörer, C. & Tinner, W. (2023). Regional vegetation history in eds Lang, G., Ammann, B., Behre, K.-E., Tinner, W., *Quaternary Vegetation Dynamics in Europe*. Haupt Verlag, Bern, Switzerland. pp.150-248.
50. Lazarus, D. (1994). Neptune: A marine micropaleontology database. Mathematical Geology, 26, 817–832. https://doi.org/10.1007/BF02083119
51. Lendemer, J., Thiers, B., Monfils, A. K., Zaspel, J., Ellwood, E. R., Bentley, A., LeVan, K., Bates, J., Jennings, D., Contreras, D., Lagomarsino, L., Mabee, P., Ford, L. S., Guralnick, R., Gropp, R. E., Revelez, M., Cobb, N., Seltmann, K., & Aime, M. C. (2020). The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. BioScience, 70(1), 23–30. https://doi.org/10.1093/biosci/biz140
52. Li, X., Feng, M., Ran, Y., Su, Y., Liu, F., Huang, C., Shen, H., Xiao, Q., Su, J., Yuan, S., & Guo, H. (2023). Big Data in Earth system science and progress towards a digital twin. Nature Reviews Earth & Environment, 4(5), 319–332. https://doi.org/10.1038/s43017-023-00409-w
    a. **Reviews and presents possible structures of Earth Systems data, offering language and frameworks for the transition to Big Data frameworks within Earth Sciences.**
53. Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J.

(2020). The TRUST Principles for digital repositories. Scientific Data, 7(1), 144. https://doi.org/10.1038/s41597-020-0486-7

54. Majeed, A., & Hwang, S. O. (2024). The Data Island Problem and Its Mitigation: Are We There Yet? Computer, 57(12), 95–103. https://doi.org/10.1109/MC.2024.3454937
55. Marcilly, C. M., Maffre, P., Hir, G. L., Pohl, A., Fluteau, F., Goddéris, Y., Donnadieu, Y., Heimdal, T. H., & Torsvik, T. H. (2022). Understanding the early Paleozoic carbon cycle balance and climate change from modelling. Earth and Planetary Science Letters, 594, 117717. https://doi.org/10.1016/j.epsl.2022.117717
56. Marshall, C. R., Finnegan, S., Clites, E. C., Holroyd, P. A., Bonuso, N., Cortez, C., Davis, E., Dietl, G. P., Druckenmiller, P. S., Eng, R. C., Garcia, C., Estes-Smargiassi, K., Hendy, A., Hollis, K. A., Little, H., Nesbitt, E. A., Roopnarine, P., Skibinski, L., Vendetti, J., & White, L. D. (2018). Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. Biology Letters, 14(9), 20180431. https://doi.org/10.1098/rsbl.2018.0431
    a. **The quantification of the volume of data not currently represented in online open access databases held within the Museums of the USA, and traces the digital development of the field.**
57. McManimon, S. K., & Natala, A. (2021). Embodied theory and lived experience: Museums are burning: Dare we engage a liberatory imagination in practice and research? In Theorizing equity in the museum: Integrating perspectives from research and practice (pp. 141–158). Routledge.
58. Monarrez, P.M., Zimmt, J.B., Clement, A.M., Gearty, W., Jacisin, J.J., Jenkins, K.M., Kusnerik, K.M., Poust, A.W., Robson, S.V., Sclafani, J.A., Stilson, K.T., Tennakoon, S.D., Thompson, C.M., 2022. Our past creates our present: a brief overview of racism and colonialism in Western paleontology. Paleobiology 48, 173–185. https://doi.org/10.1017/pab.2021.28
59. Mottl, O., Flantua, S.G.A., Bhatta, K.P., Felde, V.A., Giesecke, T., Goring, S., Grimm, E.C., Haberle, S., Hooghiemstra, H., Ivory, S., Kuneš, P., Wolters, S., Seddon, A.W.R., Williams, J.W., 2021a. Global acceleration in rates of vegetation change over the last 18,000 years. Science. https://doi.org/10.1126/science.abg1685
60. National Academies of Sciences. (2018). Open Science by Design: Realizing a vision for 21st century research (A Consensus Study Report of the National Academies of Sciences Engineering Medicine, pp. 1–232). The National Academies press.
61. Nelson, G., & Paul, D. L. (2019). DiSSCo, iDigBio and the Future of Global Collaboration. Biodiversity Information Science and Standards, 3, e37896. https://doi.org/10.3897/biss.3.37896

62. Niebuhr, B., & Wilmsen, M. (2023). The transgression history of the Saxonian Cretaceous revisited or: The imperative for a complete stratigraphic reappraisal (Cenomanian, Elbtal Group, Germany). Zeitschrift Der Deutschen Gesellschaft Für Geowissenschaften, 174(1), 69–118. https://doi.org/10.1127/zdgg/2023/0376
63. Payne, J. L., Smith, F. A., Kowalewski, M., Krause, R. A. Jr., Boyer, A. G., McClain, C. R., Finnegan, S., Novack-Gottshall, P. M., and Sheble, L.. (2012). A lack of attribution: closing the citation gap through a reform of citation and indexing practices. TAXON 61:1349–1351.10.1002/tax.616030
64. Pandolfi, J.M., Staples, T.L., Kiessling, W., 2020. Increased extinction in the emergence of novel ecological communities. Science 370, 220–222. https://doi.org/10.1126/science.abb3996
65. Peters, S. E., Husson, J. M., & Czaplewski, J. (2018). Macrostrat: A platform for geological data integration and deep-time earth crust research. Geochemistry, Geophysics, Geosystems, 19(4), 1393-1409.
66. Petersen, M., Glöckler, F., Kiessling, W., Döring, M., Fichtmüller, D., Laphakorn, L., Baltruschat, B., & Hoffmann, J. (2018). History and development of ABCDEFG: A data standard for geosciences. Fossil Record, 21(1), 47–53. https://doi.org/10.5194/fr-21-47-2018
67. Phillips, J. (1860). Life on the Earth: Its Origin and Succession. Macmillan and Company.
68. Pinfield, S., Salter, J., & Bath, P. A. (2016). The “total cost of publication” in a hybrid open-access environment: Institutional approaches to funding journal article-processing charges in combination with subscriptions. Journal of the Association for Information Science and Technology, 67(7), 1751-1766.
69. Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375. https://doi.org/10.7717/peerj.4375
70. Ponce, F., Marquez, G., & Astudillo, H. (2019). Migrating from monolithic architecture to microservices: A Rapid Review. 2019 38th International Conference of the Chilean Computer Science Society (SCCC), 1–7. https://doi.org/10.1109/SCCC49216.2019.8966423
71. Prentice, I.C., Sykes, M.T., Lautenschlager, M., Harrison, S.P., Denissenko, O., Bartlein, P.J., 1993. Modelling global vegetation patterns and terrestrial carbon storage at the last glacial maximum. Global Ecology and Biogeography Letters 3, 67–76.

72. Raja, N. B., Dunne, E. M., Matiwane, A., Khan, T. M., Nätscher, P. S., Ghilardi, A. M., & Chattopadhyay, D. (2021). Colonial history and global economics distort our understanding of deep-time biodiversity. Nature Ecology & Evolution, 6(2), 145–154. https://doi.org/10.1038/s41559-021-01608-8
73. Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From Open Data to Open Science. Earth and Space Science, 8(5), e2020EA001562. https://doi.org/10.1029/2020EA001562
    a. **Expanded framework for the transition of Earth scientists from a focus on just open access data and toward Open Science practices at all levels of research from person to product.**
74. Raup, D. M. (1972). Taxonomic Diversity during the Phanerozoic: The increase in the number of marine species since the Paleozoic may be more apparent than real. Science, 177(4054), 1065–1071. https://doi.org/10.1126/science.177.4054.1065
75. Raup, D. M., & Sepkoski, J. J. (1982). Mass Extinctions in the Marine Fossil Record. Science, 215(4539), 1501–1503. https://doi.org/10.1126/science.215.4539.1501
76. Renaudie, J., Lazarus, D., & Diver, P. (2023). Archive of Neptune (NSB) database backups (2023-06-05) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10063218
77. Riding, J. B., Pound, M. J., Hill, T. C. B., Stukins, S., & Feist-Burkhardt, S. (2012). The John Williams Index of Palaeopalynology. *Palynology*, 36(2), 224–233. https://doi.org/10.1080/01916122.2012.682512
78. Rolin, K. (2015). Values in Science: The Case of Scientific Collaboration. Philosophy of Science, 82(2), 157–177. https://doi.org/10.1086/680522
79. Ross-Hellauer, T., Reichmann, S., Cole, N. L., Fessl, A., Klebel, T., & Pontika, N. (2022). Dynamics of cumulative advantage and threats to equity in open science: A scoping review. Royal Society Open Science, 9(1), 211032. https://doi.org/10.1098/rsos.211032
80. Rousseau, S., Catalano, G., & Daraio, C. (2021). Can we estimate a monetary value of scientific publications?. Research Policy, 50(1), 104116. https://doi.org/10.1016/j.respol.2020.104116
81. Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., Hunter, C. I., Hurwitz, B. L., Kyrpides, N. C., Meyer, F., Mizrachi, I. K., Sansone, S.-A., Sutton, G., Tighe, S., & Walls, R. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. Scientific Data, 7(1), 188. https://doi.org/10.1038/s41597-020-0524-5
82. Sepkoski, J. J. (1982). A compendium of fossil marine families. Journal of the Milwaukee Public Museum Contributions in Biology and Geology, 51, 1–125.

83. Sepkoski, J. J. (1992). A compendium of fossil marine animal families. In Milwaukee Public Museum Contributions in Biology and Geology (2nd ed., Vol. 83, pp. 1–156).
84. Sepkoski Jr, J. J. (2002). A compendium of fossil marine animal genera. Bulletins of American paleontology, 363, 1-560.
85. Smith, J., Rillo, M.C., Kocsis, Á.T., Dornelas, M., Fastovich, D., Huang, H.-H.M., Jonkers, L., Kiessling, W., Li, Q., Liow, L.H., others, (2023a). BioDeepTime: A database of biodiversity time series for modern and fossil assemblages. Global Ecology and Biogeography 32, 1680–1689. https://doi.org/10.1111/geb.13735
86. Smith, J. A., Raja, N. B., Clements, T., Dimitrijević, D., Dowding, E. M., Dunne, E. M., Gee, B. M., Godoy, P. L., Lombardi, E. M., Mulvey, L. P. A., Nätscher, P. S., Reddin, C. J., Shirley, B., Warnock, R. C. M., & Kocsis, Á. T. (2023b). Increasing the equitability of data citation in paleontology: Capacity building for the big data future. Paleobiology, 1–12. https://doi.org/10.1017/pab.2023.33
    a. **Details the impact of databases in the citation outcomes for scientists, highlighting that unequal citation of data contributors versus data users undercuts open science and scientists, whilst providing recommendations.**
87. Smith, J. A., Dowding, E. M., Abdelhady, A., Abondio, P., Araújo, R., Aze, T., Balisi, M., Buatois, L., …. Kiessling, W. (2025). Identifying the “Big Questions” in paleontology. Paleobiology, In press.
88. a community-driven project
89. Stern, R. J., & Gerya, T. V. (2023). Co-Evolution of Life and Plate Tectonics: The Biogeodynamic Perspective on the Mesoproterozoic-Neoproterozoic Transitions. In Dynamics of Plate Tectonics and Mantle Convection (pp. 295–319). Elsevier. https://doi.org/10.1016/B978-0-323-85733-8.00013-5
90. Stiles, E., Montes, C., Jaramillo, C., & Gingras, M. K. (2022). A shallow-water depositional interpretation for the upper Miocene Chagres Formation (Caribbean coast of Panama). Bulletin, 134(11-12), 2971-2985.
91. Telenius, A. (2011). Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany, 29(3), 378–381.
92. Thomer, A., Williams, J., Goring, S., & Blois, J. (2025). The Valuable, Vulnerable, Long Tail of Earth Science Databases. Eos, 106. https://doi.org/10.1029/2025EO250107
    a. **Provides the monetary framework for estimating data replacement value and argues for increased advocacy for collaborative data platforms.**

93. Torsvik, T. H., Cocks, L. R. M., Domeier, M., Marcilly, C. M., & Dowding, E. M. (2025). Devonian paleogeography and environmental change: An incomplete chronicle. In ZDGG Special Issue commemorating Leopold von Buch (Vol. 1, pp. 1–26).
94. Uhen, M.D., Barnosky, A.D., Bills, B., Blois, J., Carrano, M.T., Carrasco, M.A., Erickson, G.M., Eronen, J.T., Fortelius, M., Graham, R.W., Grimm, E.C., O'Leary, M.A., Mast, A., Piel, W.H., Polly, P.D., Säilä, L.K. (2013). From card catalogs to computers: databases in vertebrate paleontology. Journal of Vertebrate Paleontology, 33, 13–28. https://doi.org/10.1080/02724634.2012.716114
    a. **Pathways for the physical repositories of evidence (museum collections for example) to increase access to the vast stores of information they hold.**
95. Wang, C., Hazen, R. M., Cheng, Q., Stephenson, M. H., Zhou, C., Fox, P., Shen, S., Oberhänsli, R., Hou, Z., Ma, X., Feng, Z., Fan, J., Ma, C., Hu, X., Luo, B., Wang, J., & Schiffries, C. M. (2021). The Deep-Time Digital Earth program: data-driven discovery in geosciences. National Science Review, 8(9). https://doi.org/10.1093/nsr/nwab027
96. Webb, I., T., Bartlein, P.J., Harrison, S.P., Anderson, K.H., 1993. Vegetation, lake levels, and climate in eastern North America for the past 18,000 years, in: Wright, Jr., H.E., Kutzbach, J.E., Webb, I., T., Ruddiman, W.F., Street-Perrott, F.A., Bartlein, P.J. (Eds.), Global Climates Since the Last Glacial Maximum. University of Minnesota Press, Minneapolis, MN, pp. 415–467.
97. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE, 7(1), e29715. https://doi.org/10.1371/journal.pone.0029715
98. Williams, J.W., Grimm, E.G., Blois, J., Charles, D.F., Davis, E., Goring, S.J., Graham, R., Smith, A.J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A.C., Betancourt, J.L., Bills, B.W., Booth, R.K., Buckland, P., Curry, B., Giesecke, T., Hausmann, S., Jackson, S.T., Latorre, C., Nichols, J., Purdum, T., Roth, R.E., Stryker, M., Takahara, H., (2018). The Neotoma Paleoecology Database: A multi-proxy, international community-curated data resource. Quaternary Research 89, 156–177. https://doi.org/10.1017/qua.2017.105
99. Williams, J.W., Blois, J.L., Capo, E., Goring, S., Heintzman, P.D., Monchamp, M.-E., Parducci, L., Spanbauer, T.L., Von Eggers, J., Alsos, I.G., Bowler, C., Coolen, M.J.L., Crump, S., Epp, L.S., Fernandez-Guerra, A., Grimm, E., Herzschuh, U., Mereghetti, A., Meyer, R., Nota, K., Pedersen, M.W., Perez, V., Shapiro, B., Stoof-Leichsenring, K.R., Wood, J.R., (2023). Strengthening global-change science by integrating aeDNA

with paleoecoinformatics. Trends in Ecology & Evolution. https://doi.org/10.1016/j.tree.2023.04.016

100. Žliobaitė, I. (2024). Laws of macroevolutionary expansion. Proceedings of the National Academy of Sciences, 121(33), e2314694121.