nature ecology & evolution

Analysis

https://doi.org/10.1038/s41559-026-02985-8

The billion-dollar case for sustaining palaeontology's digital databases

Received: 15 September 2025

Accepted: 13 January 2026

Published online: 10 February 2026

Check for updates

Elizabeth M. Dowding[1] ✉, Emma M. Dunne[1,2] ✉, Katie S. Collins[3], Katheryn Cryer[1], Kenneth De Baets[4], Danijela Dimitrijević[1], Stewart M. Edie[5], Seth Finnegan[6,7], Wolfgang Kiessling[1], Kari Lintulaakso[8], Lee Hsiang Liow[9,10], Holly Little[5], Lin Na[11], Shanan E. Peters[12], Johan Renaudie[13], Erin E. Saupe[14], Barbara Seuss[1], Jocelyn A. Sessa[15], Jansen A. Smith[16], Mark D. Uhen[17], John W. Williams[18] & Ádám T. Kocsis[1] ✉

The digital revolution has transformed palaeontology through the development of openly accessible, community-driven databases that underpin some of the most complex and large-scale empirical studies of the history of life on Earth. These systems safeguard high-effort, volunteered data and have revealed major macroevolutionary patterns, including the 'Big 5' mass extinctions. These efforts also represent remarkable global scientific and financial investment, which is continually required to support the next generation of databases and associated research. Here we conducted a survey of 118 palaeontological and allied Earth science databases, analysing their diversity dynamics, including origination and extinction rates. We show that approximately 85% of all community-curated databases have lifespans of less than 15 years, putting decades of investment at risk. We show that database creation effort has increased in the past 30 years, with peaks in database loss related to 5-year funding cycles. 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.

The study of the history of life on Earth is inherently multidisciplinary and conducted at various scales from local to global. This scientific inquiry draws from geology, biology, chemistry, archaeology and mathematics, among others, to breach human temporal perspectives, reconstruct ancient ecosystems, investigate the drivers of biodiversity and forecast how life will respond to today's changing environments[1–3]. The fossil record is essential for understanding biodiversity and Earth system processes operating at timescales beyond the twentieth and twenty-first century window of instrumental observations. The past also provides examples of Earth system states with instructive analogies to the societally novel climates that are now emerging[4,5]. From their very beginning, palaeontological databases (see 'Glossary' in Supplementary Table 1) have played pivotal roles in enabling the field to scale up from site-level studies to global-scale research. These databases were founded by scientists seeking to address questions beyond the scope of any individual palaeontological dataset, including identifying global mass extinctions and their roles in macroevolution[6] and the earliest evidence of climate-driven species' range shifts and ecosystem transformations[7,8]. The subsequent migration of palaeontological databases to open-access online platforms and data systems (encompassing the database, its system for community governance and data curation, and any associated software services) increased their accessibility

A full list of affiliations appears at the end of the paper. ✉e-mail: dowding.e.m@gmail.com; dunne.emma.m@gmail.com; adam.kocsis@fau.de

https://doi.org/10.1038/s41559-026-02985-8

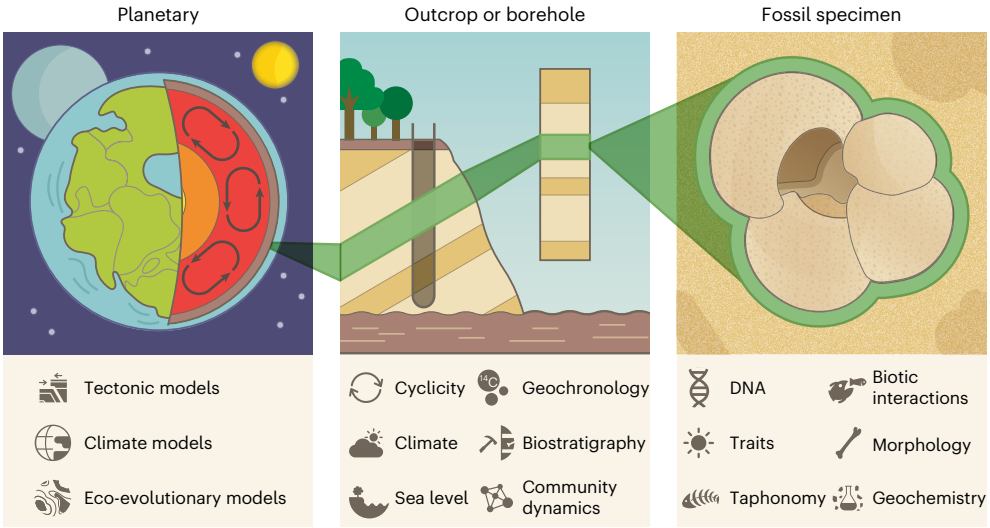

Fig. 1 | Palaeontological information in an Earth system context. From left to right, planetary or global-level information can be used to understand tectonic processes, climate and landscape evolution, and eco-evolutionary processes across timescales ranging from the present to billions of years. Outcrop- or borehole-level data provide 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 are the foundational unit in palaeobiology for analyses of, for example, taxonomy, biotic interactions, geochemistry, functional ecology, behaviour and taphonomic processes. Credit: Science Graphic Design.

and amplified their impact by enabling new questions to be explored, broader collaboration and reproducibility[9,10].

Today, openly accessible, community-run data systems function as collective archives for scientific data and knowledge about the history of life on Earth[11] (Fig. 1). These databases are invaluable for quantitatively reconstructing ancient ecosystems[12], tracing evolutionary pathways[13], studying climate- and human-driven eco-evolutionary dynamics at continental to global scales[5,14–16], and predicting future biological and geological changes—or at least assessing the limits to predictability in an increasingly novel world[17,18]. By integrating these palaeontological databases with other open data systems, scientists can tackle increasingly complex, multifaceted questions that are top priorities in global change research[3,19,20].

Representing developers, leaders, curators and users of 15 community-run palaeontological databases (Supplementary Table 2), we review the current landscape of palaeontological data systems to assess the volume, variety and value of data held in these community-curated, openly accessible databases, their diversity dynamics and longevity, the challenges faced and the opportunities for sustainable growth and scientific discovery. We close by providing recommendations for continued investment from researchers, maintainers, developers and funders.

An overview and history of palaeontological data and databases

Key concepts

Palaeontology aims to reconstruct the history of life across the broadest possible range of spatiotemporal scales and throughout the geological record (Fig. 1). Here, palaeontology encompasses closely related fields, including but not limited to palaeobiology, biostratigraphy and palaeoecology. As our collective 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 (for example, fieldwork) and the reinterpretation and reanalysis of existing data. Examples include taxonomic reidentification of old fossils following new finds[21], redating of core samples and refinement of the geological timescale using newer and improved methods and data (for example, ref. 22), re-interpretation of environmental/depositional contexts (for example, ref. 23) and incorporation of palaeobiogeographic patterns into tectonic models (for example, ref. 24).

Palaeontologists 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 (for example, 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, palaeontological databases should maintain the strongest links to fundamental data and the associated physical samples or specimens. A focus on fundamental data and well-established provenancing is essential for reproducibility and resampling efforts (for example, when palaeontologists remeasure a fossil or reassess its taxonomic identity). A focus on fundamental data also reduces database maintenance costs, because of the frequent revisions associated with processed data. Lastly, good provenancing can ensure against corollary risks such as 'data cannibalism'[25,26] when databases are used as data sources for other, secondary databases without proper attribution and dataset-level provenancing, which can violate the standard CC-BY licences that accompany most open-access data resources.

Database development history

The past—first-generation research databases. First-generation compilations of palaeontological data usually were launched with a specific research question or other purpose and focussed on the collation of processed data. For example, the collation inferred temporal (that is, stratigraphic) distribution of fossil taxa using harmonized taxonomic lists across sites, which are the minimum requirement for assessing the history of biodiversity (for example, refs. 27,28) and the shifting distribution of taxa across space and time[8,29]. These were initially collated as physical repositories (for example, the John Williams

Index of Palaeopalynology[30]) or as offline digital entities (for example, Sepkoski's Compendium[28,31] and the first version of the Neptune Sandbox database (NSB)[30]). These first-generation databases were often built either by individual scientists over their careers or by small research teams.

The present—second-generation multipurpose and community data systems. As the field advanced, palaeontologists gained further understanding of the various factors that distort the structure of the fossil record (for example, ref. 32), new research questions emerged (for example, reconstruction of past biomes and terrestrial carbon sequestration[33]), and palaeontologists developed new quantitative methods to address emergent questions. These, in turn, led to new efforts to reanalyse existing databases. For example, in deep-time biodiversity studies, the field progressed from recording observed first- and last-appearance dates of taxa[31] to the recording of fossil occurrences from the entire stratigraphic record (for example, refs. 13,34). As the breadth of questions increased, second-generation data systems (for example, Paleobiology Database (PBDB)[34] and Neotoma[10]) also began to store an increasing variety of fundamental data types, including the geographic coordinates of fossil sites, taxon abundance and traits, stratigraphic position, lithological characteristics and environmental covariates.

In parallel to this expansion of database scope, the leadership and development of these databases increasingly shifted from a few individual experts to community-curated data systems. For example, in Quaternary palynology, individual efforts to build databases and map continental-scale plant distributions for North America and Europe[8,29] expanded to continental-scale databases around the world, each with their own data leaders and stewards[10,35–37].

The data structures of current, second-generation databases vary substatially, reflecting their founding aims and user communities. As examples, PBDB was originally developed to enable, among other things, sampling-standardized estimates of Phanerozoic diversity[13]; NOW (New and Old Worlds database of fossil mammals) focused on Cenozoic mammal macroevolution[38]; Neotoma was designed to study species range shifts during the Quaternary glacial–interglacial cycles across multiple taxonomic groups[10,36–38]; and the Geobiodiversity Database (GBDB) was designed to support high-resolution stratigraphic data by linking fossil occurrences to detailed geological sections[39].

All these databases continue to grow in scope and incorporate new kinds of data. As new questions emerge and data continue to diversify and increase in accessibility[25,40], the range of scientific applications of these second-generation databases far surpass their original scope and yield input for thousands of scientific studies (see 'Database use' in Supplementary Data). For example, PBDB occurrence data have been used for climatic modelling[41,42], landscape evolution[43] and palaeogeographic models[24]. Similarly, NOW data have been used to study macroevolutionary expansion[44] and Neotoma data for reconstructing past climates[45], constraining past land cover dynamics and the terrestrial carbon cycle[46], and documenting cross-continental species invasions[47]. The scientific utility and applications of these databases thus continue to grow and diversify, as do the databases themselves.

The near future—from databases to third-generation data systems

Palaeontology is poised for its next transformative phase, in which second-generation databases will better integrate with each other and with other components of the palaeontological, Earth and life sciences data infrastructures, to address more integrative, cross-disciplinary and multiscalar questions. The transition to the third-generation database systems has already begun, with cross-database integration a core focus of backend development, using, for example, the linking capabilities provided by new data structures such as LinkML (https://linkml.io/). Other efforts are focusing on improved efficiency and more sustainable codebases through modular design[48–50]. The development of integrative platforms, such as Deep-time Digital Earth[51], and the continued growth of existing databases to support new data types, such as ancient environmental DNA[52], are striking movements towards third-generation databases.

These efforts towards integration and efficiency will enable new scientific questions to be answered at increasing power. For example, in the Big Questions in Paleontology Project[53], representative questions include 'How do external environmental drivers (for example, plate tectonics, global temperature and 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 (for example, collapse and recovery) during extinctions consistent across different biotic crises?' Addressing these integrative questions requires scalable, connected data that capture, for example, phenotypic variation among individuals in a population, in conjunction with high-stratigraphic-resolution, palaeoenvironmental and specimen-level information. These scientific needs demand further advances in how palaeontological data are reported, structured, integrated, managed and sustained. Cross-institutional aggregation of museum specimen information into iDigBio[54] and the Global Biodiversity Information Facility (GBIF[55]) provide models of how biodiversity databases can grow and be enhanced by ever-improving biodiversity data standards, such as the Darwin Core[56] and ABCDEFG[57], featuring a stronger focus on available metadata[58].

With the growth of these databases has come an awareness of their importance and impact beyond simply answering scientific research questions. Careers of an entire generation of scientists are now influenced by publicly accessible, interoperable data, and access to international, high-quality data has led to the rise of quantitative subdisciplines in palaeobiology[59,60]. Similarly, in allied fields such as geochemistry, the advent of open scientific databases PetDB and GEOROC has enabled the rise of statistical geochemistry[61]. At the same time, new concerns have arisen about whether these databases encode and perpetuate past and present inequities, such as parachute science[62], and how best to reduce these inequities to truly fulfil the deeper mission of these databases to ensure democratized data access for all[61–65]. To address these issues, the concept of community governance of palaeontological databases must be further broadened to include additional voices and to develop more effective, context-sensitive strategies that address issues of access, reciprocal research and data equity[51,53,66–68].

Landscape survey of the current state and valuation of palaeodata

Survey and assessment of palaeodata

An online survey of available palaeontological and Earth science databases was conducted using search terms in multiple languages to identify 'community-run' databases (Supplementary Table 4). Community-run databases were required not to be affiliated with any state governing body, including state-funded museums or geological surveys, and are considered 'open access' by virtue of making their data freely available 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. Among the 171 palaeontological and Earth systems databases identified, 118 were open access and community-run, and their extinction rate, origination rate and diversity dynamics were assessed (Fig. 2).

We assessed the replacement value of the data stored in three databases: PBDB, GBDB and Neotoma. We followed Thomer et al.[69] and estimated conservatively a value of US$3,000 per collection (Methods). This valuation is clearly an underestimate, not only because it does not cover all costs, but also because it assumes that the sample localities are still accessible and have not been destroyed by human

https://doi.org/10.1038/s41559-026-02985-8

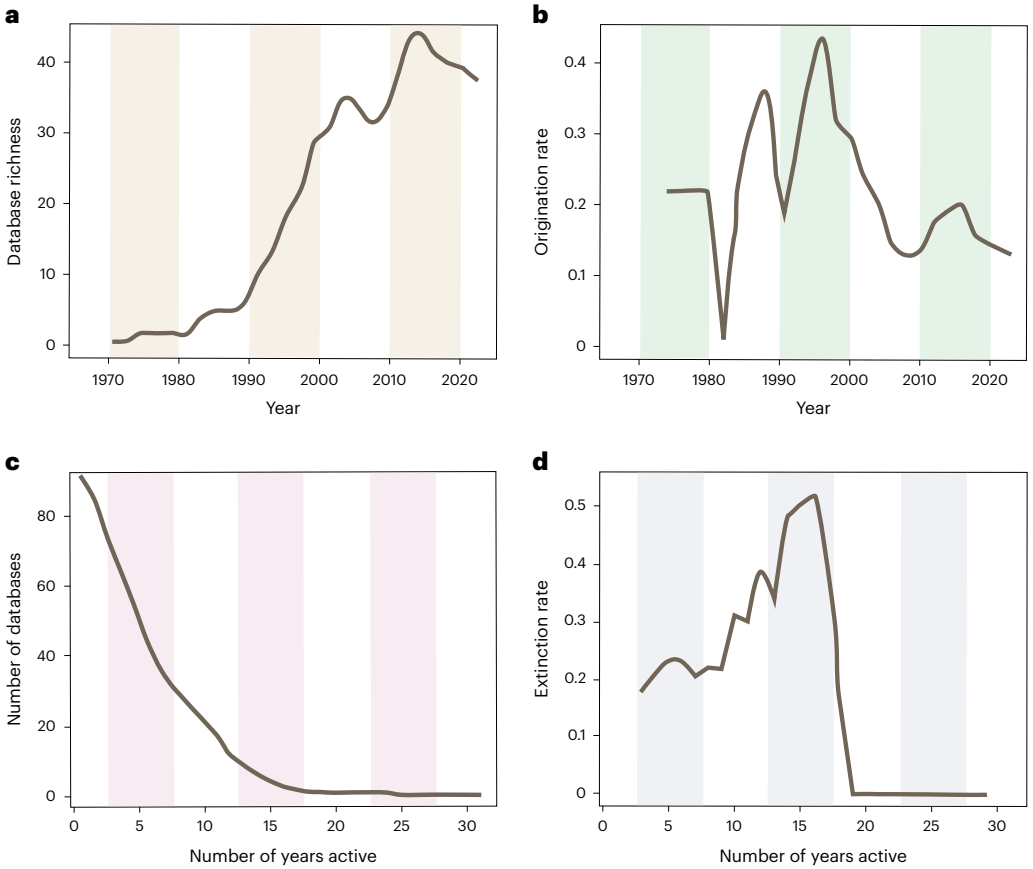

Fig. 2 | Diversity dynamics of 118 community-developed palaeontological databases from the 1970s to 2024. a, The range-through richness of databases by year. **b**, The origination rate of databases through time, indicating areas of peak activity for novel database development between 1995 and 2005. **c**, Diversity of databases as a function of years active (that is, database survivorship) showing the loss of >80% of database diversity by 10 years of activity. **d**, The rolling mean per-capita extinction rate of databases as a function of years active since inception, with peaks at 5, 15 and 25 years of activity.

land use or natural processes such as earthquakes. The data from inaccessible locations are therefore irreplaceable and priceless[70–72]. 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. They additionally do not cover the costs of labour, server hosting and infrastructure development that go into setting up and sustaining databases. Furthermore, results do not cover the article processing costs to publish a scientific paper (mean US$2,300)[73] or the value of papers published (estimated at over US$5,000 per item)[74].

Historical trends in results

Based on our web search, database origination rates peaked in the 1970s and 1990s, with a tertiary peak in the 2010s (Fig. 2). Nearly 50% of databases (n = 118) became inactive within just 5 years, and fewer than 15% survived a full decade. Only a rare 5% remained active for over 15 years (Fig. 2). This 5-year interval coincides with the standard competitive funding cycles of many large research grants from wealthy international unions or countries with a high gross domestic product, such as through the European Research Council or the US National Science Foundation, respectively. 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.

For example, recent attacks on the cyber infrastructure of the Museum für Naturkunde Berlin resulted in the loss of community access to the NSB. The NSB was intermittently funded (16 funded years since 1990) and maintained by an individual expert (see 'Curator review' in Supplementary Data). The NSB held hundreds of thousands of marine plankton microfossil species that are used to research marine community responses to climate change and has been a data contributor to other databases, including BioDeepTime[75], Microtax[76], GBDB[39] and Triton[77] among others. This attack not only impacted a key resource for microfossil taxonomists, evolutionary (palaeo)biologists and palaeoceanographers, but the data provenance of the dependent databases was compromised. The lack of funding and dedicated technical support resulted in insufficient failsafes at the museum. Instead, through community activity, external versions of the NSB—for example, those hosted on Zenodo[78]—are contributing to database recovery, further highlighting the value of community contributions in sustaining data resources.

Although 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 substantial economic waste (Fig. 2 and Table 1). The cost of allowing valuable data infrastructure to degrade is not conceptual but quantifiable and substantial. The best-case scenario for at-risk databases, as illustrated by strategy 2, involves integration with larger data systems—an example being the current incorporation of 34 constituent databases into Neotoma. The data protected and expanded by Neotoma were recently estimated to cost over US$1.5 billion to replace[69]. Despite the valuation and proven utility to the community[59,68,79], even long-lasting success stories such as Neotoma are at risk due to reliance on grant-based funding. Consequently, sustainable data infrastructure requires treating data contribution not as an obligation, but as a scholarly practice, and databases should be thought of not as products, but as commons

https://doi.org/10.1038/s41559-026-02985-8

Table 1 | The value of the samples and collections (sites) stored within three active palaeo databases in US dollars

DB	Samples (*n*)	Collections (*n*)	Sample value (US$)	Collections value (US$)	Total (US$)
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

Conservative value estimates are taken from the valuation framework[69] 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[69] has been expanded to include PBDB and GBDB. Samples refer to individual records, for example, species occurrence in PBDB. A collection refers to a grouping of samples—for example, a geographical site such as an outcrop in GBDB, or a field location in Neotoma.

sustained by collective stewardship. Volunteer labour in data contribution and backend development is often invisible and rarely credited, yet it is what has kept these long-lived databases active.

Databases achieving longevity exceeding 20 years tended to use one or more of three distinct strategies. First, some databases have relied on dedicated volunteer maintenance by one or two individuals with free or cheap institutional hosting support (for example, NSB). This solution can extend database longevity and sustainability through ties to the career of individuals but faces challenges when those individuals retire or shift positions. Second, some databases have enhanced their sustainability and achieved economies of scale by joining together and integrating into larger cooperative structures. For example, Neotoma was first formed as the union of FAUNMAP, the Latin American Pollen Database and other continental-scale databases, and new Constituent Databases continue to form and join Neotoma to leverage its data model and services[10,37]. Third, direct community-driven data contribution: PBDB has grown through primary data uploads from hundreds of volunteer contributors (see 'Curator review' in Supplementary Data). Within community initiatives, both the cooperative-database model and the direct-volunteer model leverage international research communities to build and grow their databases, while the first three strategies all rely on competitive grant funding to sustain and develop their data infrastructure. This community effort was also supported by the introduction of novel funding systems, such as the US National Science Foundation (NSF) Geoinformatics programme, which shifted its support for scientific databases from a traditional 3-year model to a development-dependent model. This new model includes a 3-year ramp-up stage for new resources, a primary database support stage lasting up to 10 years (divided into 3- to 4-year competitive awards) and a 3-year ramp-down stage. Database longevity is thus linked to both sustained community investment in volunteered time, experience and data contributions and to new funding models that support sustained, community-led database growth.

Towards the third generation of palaeontological databases

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

Modular, interoperable and community-led data systems

The scientific community and governing bodies (for example, funders) must move away from the current trend of creating standalone databases that are not interoperable and either too small or too disconnected from their communities to achieve longer-term sustainability. Instead, they must design for integration and community engagement, 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[64,80–82]. Examples such as Neotoma, NOW and 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[9,37,83]. Notable features of long-lived databases include international collaboration in data stewardship, critical community contributions by way of volunteered data, and efficient data ingestion (see 'Historical trends in Results' section; see also 'Curator review' in Supplementary Data). However, databases and related resources such as the Biodiversity Heritage Library (https://www.biodiversitylibrary.org/) remain vulnerable to 'extinction', such as through cyberattacks, but more commonly due to funding termination.

By prioritizing interoperability, modularity and close engagement between databases and their supporting communities, we can build a resilient and pluralistic community of data systems that safeguard multiple dimensions of scientific data and ensure its continued relevance to scientists, external stakeholders and the general public[66,84]. To this end, we recommend the transition to a decentralized modular data network (Fig. 3), where core components, such as those responsible for taxonomy, stratigraphy and specimen provenance, are built with a flexible scope and in a way that minimizes duplicative curatorial effort. In this vision, each part of the scientific community would be responsible for curating a specific area of data and knowledge (for example, age models and time inferences; stratigraphy and lithology; taxonomies and phylogenies; organismal abundance and occurrence; fossil morphologies; and ecological traits), and these modular, interconnected systems would integrate data and knowledge across these domains. This system would function by transitioning from the fragmented and uncoordinated data landscape (Fig. 3a) to pooled, pluralistic frameworks[66] (Fig. 3b). Pluralistic approaches to data pooling maintain domain independence and flexibility, permitting field-specific misalignment (for example, the unit differences in terminology and grouping seen between core-based micropalaeontology and global macrofossil biogeography in terms of spatial and temporal binning; Fig. 3b). Modules within this system serve as interlocking elements, offering researchers the foundation to develop extension structures necessary for addressing novel scientific questions within a broader, connected data landscape (Fig. 3b). For example, to answer questions about fossil biotic interactions (for example, BITE[85]), a new data structure is required, developed specifically to tie one or more biotic interactions and the organisms to a rock specimen. This novel database element would then be integrated with pre-existing core elements such as taxonomy, stratigraphy and geography (Fig. 3b), meaning the only new element to be constructed is the one that captures explicitly biotic interaction data. This approach saves time on database construction, reduces duplicative effort, ensures interoperability and safeguards against the loss of data from novel databases. In this way, the data from 'extinct' databases can be conserved and reintegrated, either by adding them to existing core modules or by creating new modules. The suggested solution mimics the general tendency of some corporations that move from large monolithic applications to interconnected microservices to meet the demands of scalability and a fast development cycle[84,86], and is particularly suited to scientific research that is globally distributed in nature[48–50,80].

https://doi.org/10.1038/s41559-026-02985-8

Table 2 | A roadmap to sustainable funding

Action	Description
(1) Embed sustainability from inception	Design databases with modular architecture and interoperability in mind. Incorporate regional and linguistic equity in API development. This enables future integration into broader infrastructures and reduces redundancy, lowering long-term maintenance costs.
(2) Establish core infrastructure grants	Advocate for dedicated infrastructure funding schemes for domestic and international initiatives, distinct from research project grants, which support long-term maintenance, technical upgrades and data curation. Prioritize capacity building within the community in both database curation and database use.
(3) Develop cross-sector partnerships	Collaborate with museums, universities, government agencies and industry partners to co-invest in shared data resources.
(4) Quantify and communicate value	Systematically assess the scientific and economic value of databases to demonstrate return on investment and attract strategic funding.
(5) Adopt attribution standards	Promote data citation, DOI assignment and recognition mechanisms to incentivize community data contributions and support funding applications that highlight demonstrable use.
(6) Foster community governance	Create steering bodies or consortiums to coordinate long-term strategy, technical development and funding pipelines across institutions and borders.

The community enthusiasm for developing shared resources and initiatives is evident in our data landscape. The proposed roadmap relies on structured communication of the value and importance of community-developed databased, while developing cross-sector relationships and expanding community buy-in.

To realize their full potential, third-generation databases must, whenever possible, maintain direct links to physical specimens and samples (for example, through the International Generic Sample Numbers; https://ev.igsn.org/), originators and users (for example, persistent identifier through ORCID; https://orcid.org/), and usage (for example, DATACITE for DOI mining; https://datacite.org/), while also improving linkages to other databases (see 'The near future—from databases to third-generation data systems' section). Museums, research institutes and public collections are a foundation of this system, providing crucial metadata that tie scientific conclusions based on digital data to the primary physical evidence[68,87,88]. Strengthening links between physical specimens and their digital representations will ensure long-term data accessibility, foster interdisciplinary research and empower the next generation of large-scale palaeontological analyses by upholding scientific transparency and rigour[48,65,89].

Developing application programming interfaces (APIs)—which enable one software program to request services or data from another without needing to know the other system's internal workings and that adhere to open science standards[19,90–92]—is crucial for ensuring seamless exchange of information between data systems, regardless of their underlying technologies. In addition, data harmonization tools[93] can streamline data integration and scientific workflows by automatically reconciling differences in data formats, units of measurement and terminologies. For example, the fossilpoll workflow[94] (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, allowing scientists outside the database leaders/curators to add value while establishing strong provenance between these downstream research analyses and the databases. Furthermore, such tools can leverage artificial-intelligence and machine-learning algorithms to help data stewards identify and merge duplicate records, standardize taxonomic names and align stratigraphic information, reducing the manual effort required for data integration. Because of the complexity of fossil data and the implicit knowledge often embedded in palaeontological datasets, we recommend that analytical and curatorial workflows use human-in-the-loop approaches rather than fully automated systems, to avoid 'garbage in, garbage out' situations and ensure accuracy and reliability.[26].

Financial support

Using palaeontological 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. While we focus here on open digital resources for the democratization of science, investment in such resources must be accompanied by stronger linkages to, and explicit support for, museums and physical repositories[64,65,95–97].

Long-lived databases have been developed and maintained through a combination of sporadic funding, international support and unfunded volunteer/service work[61,69,98]. 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[61,98]. Funding is essential for ensuring that community-curated data continue to inform cutting-edge science well into the future.

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) such as those offered by the NSF Geoinformatic programme model. Advocating for policy support at institutional, national and international levels is required to create an environment for these systems to thrive[97]. 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[69,98].

Engaging policymakers and funding agencies in discussions about the importance of Earth science and palaeontological community data networks can help to secure the necessary support and resources (for example, 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 (for example, the Australia Data Research Commons (https://ardc.edu.au/) and the Chinese National Natural Science Fund Key Basic Research Infrastructure programme (nsfc.gov.cn/english/site_1/funding/E1/2024/06-12/364.html), which supports geo-data infrastructure). Within our proposed funding roadmap (Table 2 and Fig. 3), 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 such as Ozboneviz[97]. Ultimately, we recommend the establishment of a dedicated international non-profit organization, akin to the European Organization for Nuclear Research (CERN) or GBIF, which would advance the financial sustainability of the life and Earth science data landscape.

These two organizations, among others, provide a useful template for palaeontology and Earth systems science more broadly. The success of the GBIF, for example, lies in its clear governance, strategic coordination and stable funding model—elements that palaeontological and Earth science databases are yet to fully achieve (Fig. 3). GBIF operates as a community-governed, multinational consortium supported by member countries, each contributing financially and through data provision, underpinned by a robust strategic framework that ensures long-term stability, interoperability and open access[55,99,100]. Its structure—from local and national nodes to global coordination—promotes accountability and sustained collaboration, while its standards (for example, Darwin Core[56]) have become a foundation for data integration across the life sciences. The proposed framework for palaeontology (Table 2 and Fig. 3) builds upon this foundation while simultaneously

https://doi.org/10.1038/s41559-026-02985-8

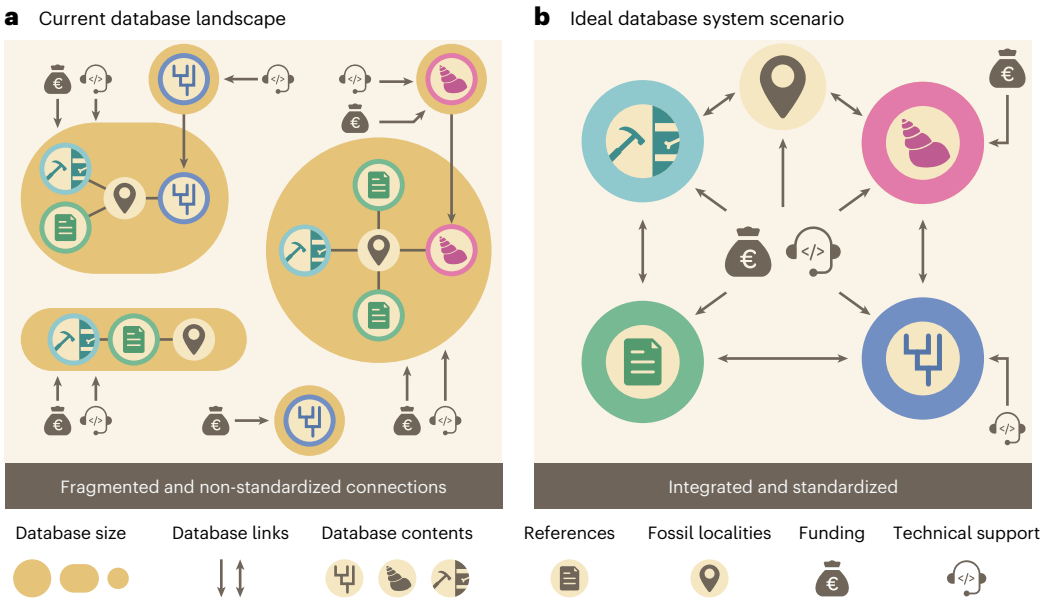

Fig. 3 | Graphical representation of the current database landscape and a possible idealized scenario for the structure of the palaeobiological database landscape. a, The current data landscape consists of disparate databases of varying size, scope and resolutions that are connected through limited links. This means that core elements (such as fossil localities) each have independent and therefore repeated solutions for each database. **b**, An idealized database system consists of a decentralized network of interconnected independent subsystems (nodes) that benefit from collective standards and potential sharing of financial and technical support. Core data elements, such as fossil localities, are accessible as a standardized module that each database uses when developing domain-specific data structures (for example, for phylogenetic matrices or stratigraphy). A central support system, without compromising data sovereignty, can be put in place to decrease funding volatility and ensure network-level standards and integrity. The maintainers of individual databases can then apply for specific external grants and technical support to develop tailored data solutions to address novel scientific questions. Credit: Science Graphic Design.

acknowledging that pluralist approaches to establishing and formalizing collaboration are essential. Stable financial support is required to bridge the domain-specific knowledge and data structures required to hold information from the multiple subdisciplines of palaeontology (for example, palaeobiology, biostratigraphy, ichnology and palaeobotany) and related Earth sciences (for example, sedimentology, geodynamics, geochemistry and climatology). This inclusivity ensures that the framework not only supports technical interoperability but also fosters equitable participation and long-term sustainability across the full spectrum of palaeontological and Earth science research (Fig. 3). For example, a recent review of geochemical databases highlights similar trends in data lifecycles, while highlighting that geochemical data require tailormade data structures to host and develop them[61]. In understanding the data requirement, and if led by those creating and using the data, a GBIF-like framework (that is, formalized international partnerships, strategic cooperative leadership, modular infrastructure and clear attribution systems) can secure sustainable data management, enhance interoperability and ensure the long-term preservation and growth of palaeontology and Earth science's collective digital resources.

Community governance and goals

We propose a phased, community-guided transition towards a sustainable, interconnected, and explicitly modular data infrastructure—one that is grounded in equitable practice and ensures proper attribution[40,64,91,101–103]. As artificial intelligence, large-scale web scraping and automated data aggregation become increasingly common tools, the palaeontological community must actively shape how its openly accessible data are structured, cited and reused[25,40,98,102]. 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[66]. By harmonizing efforts and redistributing responsibilities through open consultation, we can build an equitable and future-ready infrastructure that supports both innovation and accountability in palaeoscience.

Promising steps are already underway. Initiatives such as 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 palaeodiversity data, and open infrastructure. In 2023, CABAH produced 127 journal articles and welcomed over 60,000 attendees to its public programmes 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 trust in open data resources[90–92]. True integration goes beyond technical aspects and requires active collaboration between scientists and technical experts from varied disciplines (Table 3 and Fig. 3). Establishing interdisciplinary data standards, training programmes, research teams and projects can facilitate this collaboration (Table 2). Through this effort, we can develop common research frameworks and questions to guide data integration efforts, aligning the objectives of different disciplines[63].

Conclusions

Palaeontological data systems are critical resources for the advancement of Earth system research and the training and development of Earth scientists. By committing to the development and maintenance of decentralized, interconnected, modular data systems, we can address pressing questions about the history of life on Earth, ensure the

Table 3 | Recommendations for the sustainable development of community-developed data resources and the related benefits derived from their implementation

Recommendation	Stakeholders	Details	Benefits
(A) Incentivize data contributions	Researchers, data curators, database developers, policymakers	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	Researchers, data curators, database developers, funders	Develop a standardized framework for integrating diverse Earth system databases, ensuring interoperability and data quality transparency.	1, 2, 5, 6
(C) Secure sustainable funding	Researchers, database developers, funders, policymakers, institutions	Advocate for dedicated funding streams to support the development, maintenance and enhancement of modular data systems.	All
(D) Promote open science practices	All	Encourage the adoption of open science practices, including open data, open-access publications and collaborative research initiatives.	All
(E) Invest in technology and innovation	Funders, policymakers, institutions	Leverage technological advancements to enhance data integration, analysis and visualization capabilities.	1, 2, 6
(F) Build and foster global collaborations	Researchers, funders, policymakers, institutions	International collaborations and partnerships create a comprehensive and diverse global network of palaeontological data.	2, 4, 6
(G) Ensure ethical and legal compliance	All	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	All	Advocating for policy support at institutional, national and international levels is required to create an environment for these systems to thrive.	All

The benefits are rigour 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 tasks (6) and open publication for public benefit (7; see Supplementary Table 3 for expansion and descriptions).

longevity of our shared resources and create a more interconnected scientific community. This effort is already underway, building on the success of first- and second-generation data systems that have advanced our understanding and technical capacity. Developing integrated support systems will protect, sustain and enhance these valuable community-driven 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 palaeontological data remain valuable resources for future generations.

Methods

Database survey

To assess the temporal dynamics and sustainability of palaeontological databases, we systematically searched Web of Science and Google Scholar between November 2024 and March 2025. Search terms combined multilingual instances of 'database' (Supplementary Table 4) with 'palaeontology', 'geology', 'fossil' and 'Earth science'. Web of Science searches were restricted to 'Physical', 'Chemical & Earth Sciences' and 'Life Sciences' categories. Languages were selected on the basis of distribution by official or co-official status: English, Spanish, Arabic and French, following country counts from the South Australian Government[104]. Additional major languages (for example, Mandarin) were searched for up to five pages, with searches terminated when no new non-governmental databases were identified.

We inspected the first 10 result pages per aggregator (100 results for Google Scholar; 250 for Web of Science). Each result was examined to distinguish presentations of new databases from studies citing existing databases. Results were recorded following standardized definitions (Supplementary Table 5).

Temporal and funding data

Inception or start date (Supplementary Table 5) was defined as the year a database became publicly available, determined by the earliest of associated publication date or website launch. End date (last reference/update; Supplementary Table 5) was the most recent documented update, identified hierarchically from: (1) database website update information, (2) publications documenting database state, (3) most recent citation in scientific literature or (4) last confirmed year of public accessibility. Databases with identical start and end dates were included in diversity metrics but excluded from longevity and extinction analyses as they represent point occurrences rather than temporal spans. Funding information was recorded from database websites or associated publications when available. Records lacking start or end dates were omitted from analyses.

Database analysis

We identified 171 palaeontological and Earth science databases. After removing governmental databases (see rationale in Supplementary Table 5), 125 remained, of which 118 met the inclusion criteria (see 'Richness' in Supplementary Data). Summary statistics on database duration excluded same-year databases ($n = 30$), yielding 88 databases with temporal ranges (Supplementary Tables 6 and 7). An additional analysis excluded the top 15% longest-lived databases (approximately 25 years; Supplementary Table 6) to examine diversity dynamics, because the diversity and number of databases through time and associated changes affect the data landscape and its stability (Fig. 2), representative of typical community-maintained databases.

The per-capita extinction and origination rates were analysed using a rolling mean of year-to-year database activity, while the sampled-in-bin diversity used an extended decadal time series to account for boundary conditions.

To mitigate edge effects, we extended end dates of active databases to 2027 and truncated analyses at 2024. This approach addresses the pull-of-the-recent bias affecting terminal rates. The time series start was extended to the 1970s to include early static datasets that predate digital database proliferation.

Analyses were conducted in RStudio (4.5.0)[105,106] using DivDyn[107]. To address the question of database diversity dynamics, we calculated:

- Richness: total number of databases active within a time (divSIB);
- Diversity by duration: distribution of databases by years active (divRT);
- Origination rate: the rate at which new databases are established per year (2-year rolling mean; PC:oriPC);
- Extinction rate: the rate at which databases cease activity per year (2-year rolling mean; PC: extPC).

Rolling means used a 2-year window to smooth interannual variation. All raw data, including point occurrences (same-year databases), were included in rolling mean calculations to capture complete database origination dynamics. The following metrics

were considered in both raw and rolling mean treatments for origination, extinction and diversity used in Fig. 2: sampled-in-bin diversity, range-through diversity, per-capita extinction and per-capita origination. A full list of 12 metrics (Supplementary Table 8) was calculated.

Author survey on database maintainers, curators and data contributors

The authors of this paper, who were also database maintainers and/or developers, volunteered information about the backend, data volume and support structures (Supplementary Table 2). The authors presented data across 68 categories, including 'History and funding management' (7 categories), 'Scope' (3 categories), 'Software and maintenance' (16 categories), 'Data contained' (5 categories), and 'Entity feature' coverage (37 categories; see 'Curator review' in Supplementary Data). These descriptions informed benefits and recommendations (Tables 2 and 3) and present a clear synthesis of the variability in database structure and maintenance. The provided database ages were incorporated into the database survey (Fig. 2), in addition to funding and technical support summaries.

The citation count for each database was also requested from the database maintainers. PBDB was selected for full consideration, while Neotoma, the Geobiology Database, Triton, Neptune and NOW are present for completeness (see 'Palaeodatabase publication products' in Supplementary Data). The published literature (>1,800 papers) that cited PBDB as a data contributor were each tagged using 15 categories (palaeobiogeography, diversity, taxonomy, morphology, phylogeny, palaeoecology, environment, taphonomy, palaeoclimate, conservation, geochemistry, sedimentology, sedimentology, stratigraphy, evolution and other) to capture the diversity of topics PBDB data are used for. Owing to citation practices, the number of formal citations gained by PBDB is notably lower than the citing literature or 'mentions' the database when querying aggregators (>34,000 from Google Scholar, February 2025).

Methods for financial valuation

We used a financial valuation framework[69] on the data volume that was provided either by the database maintainers (see 'Curator review' in Supplementary Data) or the most recent version of the database as of June 2025. The rationale valuation centres on the cost of replacing the data assuming only labour, expertise and institutional overhead are required[69]. The rationale also assumes that the data can be collected again, that the sites are still accessible, and that equal quality specimens can be obtained. Within palaeontology and Earth sciences, this is often not the case. We elected to focus on only two of the options: sample value (US$150) and site value (US$3,000).

Additional costs for publication, data hosting, hiring database maintainers and developers, and curatorial labour were not included in the valuation (and also were not listed in ref. 69 the valuation formula.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data generated for the author survey, publication, products and analysis are available within the Supplementary Information and via Zenodo at https://doi.org/10.5281/zenodo.17828000 (ref. 108). Source data are provided with this paper.

Code availability

All code and required data are available via CodeOcean at https://doi.org/10.24433/CO.1586965.v1 (ref. 109). All analyses were conducted using publicly available R packages, and the links have been provided in the Methods.

References

1. Dietl, G. P. & Flessa, K. W. Conservation paleobiology: putting the dead to work. *Trends Ecol. Evol.* **26**, 30–37 (2011).
2. Dillon, E. M. et al. What is conservation paleobiology? Tracking 20 years of research and development. *Front. Ecol. Evol.* **10**, 1031483 (2022).
3. Kiessling, W., Smith, J. A. & Raja, N. B. Improving the relevance of paleontology to climate change policy. *Proc. Natl Acad. Sci. USA* **120**, e2201926119 (2023).
4. Burke, K. D. et al. Pliocene and Eocene provide best analogs for near-future climates. *Proc. Natl Acad. Sci. USA* **115**, 13288–13293 (2018).
5. Pandolfi, J. M., Staples, T. L. & Kiessling, W. Increased extinction in the emergence of novel ecological communities. *Science* **370**, 220–222 (2020).
6. Raup, D. M. & Sepkoski, J. J. Mass extinctions in the marine fossil record. *Science* **215**, 1501–1503 (1982).
7. Davis, M. B. Pleistocene biogeography of temperate deciduous forests. *Geosci. Man* **13**, 13–26 (1976).
8. Bernabo, J. C. & Webb, T. Changing patterns in the Holocene pollen record of Northeastern North America: a mapped summary. *Quat. Res.* **8**, 64–96 (1977).
9. Uhen, M. D. et al. From card catalogs to computers: databases in vertebrate paleontology. *J. Vertebr. Paleontol.* **33**, 13–28 (2013).
10. Williams, J. W. et al. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quat. Res.* **89**, 156–177 (2018).
11. Guo, H. Big Earth data: a new frontier in Earth and information sciences. *Big Earth Data* **1**, 4–20 (2017).
12. Cribb, A. T. & Darroch, S. A. F. How to engineer a habitable planet: the rise of marine ecosystem engineers through the Phanerozoic. *Palaeontology* **67**, e12726 (2024).
13. Alroy, J. et al. Phanerozoic trends in the global diversity of marine invertebrates. *Science* **321**, 97–100 (2008).
14. Mottl, O. et al. Global acceleration in rates of vegetation change over the past 18,000 years. *Science* **372**, 860–864 (2021).
15. Lang, G. et al. *Regional Vegetation History* (Univ. Bern, 2024); https://doi.org/10.48350/185289
16. Gordon, J. D., Fagan, B., Milner, N. & Thomas, C. D. Floristic diversity and its relationships with human land use varied regionally during the Holocene. *Nat. Ecol. Evol.* **8**, 1459–1471 (2024).
17. Fitzpatrick, J. L. & Lüpold, S. Sexual selection and the evolution of sperm quality. *MHR Basic Sci. Reprod. Med.* **20**, 1180–1189 (2014).
18. Stern, R. J. & Gerya, T. V. in *Dynamics of Plate Tectonics and Mantle Convection* (eds. Duarte, J. C.) 295–319 (Elsevier, 2023); https://doi.org/10.1016/B978-0-323-85733-8.00013-5
19. National Academies of Sciences, Engineering, and Medicine. *Open Science by Design: Realizing a Vision for 21st Century Research* (The National Academies Press, 2018).
20. Wang, H., Han, D., Blom, H., Dupret, V. & Pan, Z. Evolving trends in vertebrate palaeontology (2013–2022): a bibliometric analysis using DeepBone and Web of Science databases. *Hist. Biol.* https://doi.org/10.1080/08912963.2024.2330075 (2024).
21. Godfrey, S. J. & Collareta, A. A new ichnotaxonomic name for burrows in vertebrate coprolites from the Miocene Chesapeake Group of Maryland, U.S.A. *Swiss J. Palaeontol.* **141**, 9 (2022).
22. Dai, X. et al. Geochronology of the Early Triassic based on coupled Bayesian zircon eruption age and Bayesian age–depth models. *Proc. Natl Acad. Sci. USA* **122**, e2509247122 (2025).
23. Stiles, E., Montes, C., Jaramillo, C. & Gingras, M. K. A shallow-water depositional interpretation for the upper Miocene Chagres Formation (Caribbean coast of Panama). *GSA Bull.* https://doi.org/10.1130/B36291.1 (2022).

24. Torsvik, T. H., Cocks, L. R. M., Domeier, M., Marcilly, C. M. & Dowding, E. M. Devonian paleogeography and environmental change: an incomplete chronicle. *Z. Dtsch. Ges. Geowiss.* **1**, 1–26 (2025).
25. Shumailov, I. et al. AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
26. Bircan, T. & Özbilgin, M. F. Unmasking inequalities of the code: Disentangling the nexus of AI and inequality. *Technol. Forecast. Soc. Change* **211**, 123925 (2025).
27. Phillips, J. *Life on the Earth: Its Origin and Succession* (Macmillan and Company, 1860).
28. Sepkoski, J. J. A compendium of fossil marine families. *Journal Milwaukee Public Mus. Contrib. Biol. Geol.* **51**, 1–125 (1982).
29. Huntley, B. & Birks, H. J. B. *An Atlas of Past and Present Pollen Maps for Europe: 0–13000 Years Ago* (Cambridge Univ. Press, 1983).
30. Lazarus, D. Neptune: a marine micropaleontology database. *Math. Geol.* **26**, 817–832 (1994).
31. Sepkoski, J. J., Jablonski, D. & Foote, M. J. *A Compendium of Fossil Marine Animal Genera* (Paleontological Research Institution, 2002).
32. Raup, D. M. Taxonomic diversity during the Phanerozoic: The increase in the number of marine species since the Paleozoic may be more apparent than real. *Science* **177**, 1065–1071 (1972).
33. Prentice, I. C. et al. Modelling global vegetation patterns and terrestrial carbon storage at the Last Glacial Maximum. *Glob. Ecol. Biogeogr. Lett.* **3**, 67 (1993).
34. Alroy, J. et al. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proc. Natl Acad. Sci. USA* **98**, 6261–6266 (2001).
35. Uhen, M. D. et al. Paleobiology Database User Guide Version 1.0. *PaleoBios* https://doi.org/10.5070/P9401160531 (2023).
36. Grimm, E. C. et al. Pollen databases and their application. in *Encyclopedia of Quaternary Science* (ed Elias, S. A.) 831–838 (Elsevier, 2013); https://doi.org/10.1016/B978-0-444-53643-3.00174-6
37. Grimm, E. C. et al. Constituent databases and data stewards in the Neotoma Paleoecology Database: history, growth, and new directions. *Past Glob. Change Mag.* **26**, 64–65 (2018).
38. Jernvall, J. & Fortelius, M. Common mammals drive the evolutionary increase of hypsodonty in the Neogene. *Nature* **417**, 538–540 (2002).
39. Fan, J. et al. Geobiodiversity Database: a comprehensive section-based integration of stratigraphic and paleontological data. *Newsl. Stratigr.* **46**, 111–136 (2013).
40. Smith, J. A. et al. Increasing the equitability of data citation in paleontology: capacity building for the big data future. *Paleobiology* https://doi.org/10.1017/pab.2023.33 (2023).
41. Marcilly, C. M., Torsvik, T. H. & Jones, M. T. Late Paleozoic climate transition from a long-term carbon cycle modeling perspective. *Glob. Planet. Change* **253**, 104984 (2025).
42. Marcilly, C. M., Torsvik, T. H. & Conrad, C. P. Global Phanerozoic sea levels from paleogeographic flooding maps. *Gondwana Res.* **110**, 128–142 (2022).
43. Fernandes, V. M., Roberts, G. G., White, N. & Whittaker, A. C. Continental-scale landscape evolution: a history of North American topography. *J. Geophys. Res. Earth Surf.* **124**, 2689–2722 (2019).
44. Žliobaitė, I. Laws of macroevolutionary expansion. *Proc. Natl Acad. Sci. USA* **121**, e2314694121 (2024).
45. Chevalier, M. et al. Pollen-based climate reconstruction techniques for late Quaternary studies. *Earth Sci. Rev.* **210**, 103384 (2020).
46. Blarquez, O. & Aleman, J. C. Tree biomass reconstruction shows no lag in postglacial afforestation of eastern Canada. *Can. J. For. Res.* **46**, 485–498 (2016).
47. Alverson, A. J. et al. 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* **23**, 2191–2204 (2021).
48. Deng, M. & Di, L. Building open environments to meet big data challenges in Earth sciences. in *Big Data Techniques and Technologies in Geoinformatics* (eds Karimi, H. A.) 67–88 (CRC Press, 2024).
49. Deng, Y. et al. Paleontology knowledge graph for data-driven discovery. *J. Earth Sci.* **35**, 1024–1034 (2024).
50. Kaufman, D. S. & PAGES 2k special-issue editorial team Technical note: Open-paleo-data implementation pilot—the PAGES 2k special issue. *Clim. Past* **14**, 593–600 (2018).
51. Wang, C. et al. The Deep-Time Digital Earth program: data-driven discovery in geosciences. *Natl Sci. Rev.* **8**, nwab027 (2021).
52. Williams, J. W. et al. Strengthening global-change science by integrating aeDNA with paleoecoinformatics. *Trends Ecol. Evol.* **38**, 946–960 (2023).
53. Smith, J. A. et al. Identifying the Big Questions in paleontology: a community-driven project. *Paleobiology* https://doi.org/10.1017/pab.2025.10042 (2025).
54. Nelson, G. & Paul, D. L. DiSSCo, iDigBio and the future of global collaboration. *Biodivers. Inf. Sci. Stand.* **3**, e37896 (2019).
55. Telenius, A. Biodiversity information goes public: GBIF at your service. *Nord. J. Bot.* **29**, 378–381 (2011).
56. Wieczorek, J. et al. Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE* **7**, e29715 (2012).
57. Petersen, M. et al. History and development of ABCDEFG: a data standard for geosciences. *Foss. Rec.* **21**, 47–53 (2018).
58. Hardisty, A. R. et al. Digital extended specimens: enabling an extensible network of biodiversity data records as integrated digital objects on the Internet. *BioScience* **72**, 978–987 (2022).
59. Li, X. et al. Big Data in Earth system science and progress towards a digital twin. *Nat. Rev. Earth Environ.* **4**, 319–332 (2023).
60. Koch, A. et al. Open-data practices and challenges among early-career paleo-researchers. *Past Glob. Change Mag.* **26**, 54–55 (2018).
61. Klöcking, M., Lehnert, K. A. & Wyborn, L. Geochemical databases. in *Treatise on Geochemistry* (eds Anbar, A., & Weis, D.) 97–135 (Elsevier, 2025); https://doi.org/10.1016/B978-0-323-99762-1.00123-6
62. Raja, N. B. et al. Colonial history and global economics distort our understanding of deep-time biodiversity. *Nat. Ecol. Evol.* **6**, 145–154 (2021).
63. Rolin, K. Values in science: the case of scientific collaboration. *Philos. Sci.* **82**, 157–177 (2015).
64. Dunne, E. M. et al. Data equity in paleobiology: progress, challenges, and future outlook. *Paleobiology* **51**, 237–249 (2025).
65. McManimon, S. K. & Natala, A. in *Theorizing Equity in the Museum: Integrating Perspectives from Research and Practice* (eds Bevan. B & Ramos, B) 141–158 (Routledge, 2021).
66. Sterner, B., Elliott, S., Gilbert, E. E. & Franz, N. M. Unified and pluralistic ideals for data sharing and reuse in biodiversity. *Database* **2023**, baad048 (2023).
67. Hurst, S. et al. More than museums: care for natural and cultural heritage in Australia. *Geoconserv. Res.* **7**, 072405 (2024).
68. Boldgiv, B. et al. Global natural history infrastructure requires international solidarity, support, and investment in local capacity. *Proc. Natl Acad. Sci. USA* **122**, e2411232122 (2025).
69. Thomer, A., Williams, J., Goring, S. & Blois, J. The valuable, vulnerable, long tail of earth science databases. *Eos* https://doi.org/10.1029/2025EO250107 (2025).
70. Cyranoski, D. Mining threatens Chinese fossil site that revealed planet's earliest animals. *Nature* **544**, 403–403 (2017).

71. Kumar, S. India's paleontologists fight destruction of its fossil riches. *Sci. Advis*. https://doi.org/10.1126/science.aat7646 (2021).
72. King, L. M. & Halpenny, E. A. Communicating the World Heritage brand: visitor awareness of UNESCO's World Heritage symbol and the implications for sites, stakeholders and sustainable management. *J. Sustain. Tour.* **22**, 768–786 (2014).
73. Pinfield, S., Salter, J. & Bath, P. A. The "total cost of publication" in a hybrid open-access environment: institutional approaches to funding journal article-processing charges in combination with subscriptions. *J. Assoc. Inf. Sci. Technol.* **67**, 1751–1766 (2016).
74. Rousseau, S., Catalano, G. & Daraio, C. Can we estimate a monetary value of scientific publications?. *Res. Policy* **50**, 104116 (2021).
75. Smith, J. et al. BioDeepTime: a database of biodiversity time series for modern and fossil assemblages. *Glob. Ecol. Biogeogr.* **32**, 1680–1689 (2023).
76. Huber, B. T. et al. Pforams@microtax. *Micropaleontology* **62**, 429–438 (2016).
77. Fenton, I. S. et al. Triton, a new species-level database of Cenozoic planktonic foraminiferal occurrences. *Sci. Data* **8**, 160 (2021).
78. Renaudie, J., Lazarus, D. & Diver, P. Archive of Neptune (NSB) database backups. *Zenodo* https://doi.org/10.5281/ZENODO.10063218 (2023).
79. Adeleye, M. A., Haberle, S. G., Gallagher, R., Andrew, S. C. & Herbert, A. Changing plant functional diversity over the last 12,000 years provides perspectives for tracking future changes in vegetation communities. *Nat. Ecol. Evol.* **7**, 224–235 (2023).
80. Yiying, D. et al. Current status of paleontological databases and data-driven research in paleontology. *Geol. J. China Univ.* **26**, 361–383 (2020).
81. Ramachandran, R., Bugbee, K. & Murphy, K. From open data to open science. *Earth Space Sci.* **8**, e2020EA001562 (2021).
82. Ross-Hellauer, T. et al. Dynamics of cumulative advantage and threats to equity in open science: a scoping review. *R. Soc. Open Sci.* **9**, 211032 (2022).
83. Casanovas-Vilar, I. *Evolution of Cenozoic Land Mammal Faunas and Ecosystems: 25 Years of the NOW Database of Fossil Mammals* (Springer, 2023).
84. Majeed, A. & Hwang, S. O. The data island problem and its mitigation: are we there yet?. *Computer* **57**, 95–103 (2024).
85. Huntley, J. et al. Biotic Interactions in Deep Time (BITE): developing a specimen-level database to address fundamental questions. In *Geological Society of America Abstracts with Programs* https://doi.org/10.1130/abs/2023AM-390127 (2023).
86. Ponce, F., Marquez, G. & Astudillo, H. Migrating from monolithic architecture to microservices: a rapid review. In *2019 38th International Conference of the Chilean Computer Science Society* 1–7 (IEEE, 2019); https://doi.org/10.1109/SCCC49216.2019.8966423
87. Schriml, L. M. et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data* **7**, 188 (2020).
88. Johnson, K. R., Owens, I. F. P. & the Global Collection Group A global approach for natural history museum collections. *Science* **379**, 1192–1194 (2023).
89. DeMiguel, D. et al. Linking geological heritage and geoethics with a particular emphasis on palaeontological heritage: the new concept of 'palaeontoethics. *Geoheritage* **13**, 69 (2021).
90. Lin, D. et al. The TRUST principles for digital repositories. *Sci. Data* **7**, 144 (2020).
91. Carroll, S. R. et al. The CARE principles for Indigenous data governance. *Data Sci. J.* **19**, 43 (2020).
92. Jacobsen, A. et al. FAIR principles: interpretations and implementation considerations. *Data Intell.* **2**, 10–29 (2020).
93. Grenié, M. et al. Harmonizing taxon names in biodiversity data: a review of tools, databases and best practices. *Methods Ecol. Evol.* **14**, 12–25 (2023).
94. Flantua, S. G. A. et al. A guide to the processing and standardization of global palaeoecological data for large-scale syntheses using fossil pollen. *Glob. Ecol. Biogeogr.* **32**, 1377–1394 (2023).
95. Allmon, W. D., Dietl, G. P., Hendricks, J. R. & Ross, R. M. in *Museums at the Forefront of the History and Philosophy of Geology: History Made, History in the Making* (eds Rosenberg, G. D. & Clary, R. M.) 35–44 (Geological Society of America, 2018); https://doi.org/10.1130/2018.2535(03)
96. Marshall, C. R. et al. Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biol. Lett.* **14**, 20180431 (2018).
97. Weisbecker, V. et al. Ozboneviz: an Australian precedent in FAIR 3D imagery and extended biodiversity collections. *BioScience* **75**, 747–756 (2025).
98. Hooft, R. W. W. & Roos, M. Financing models for sustainable data reuse infrastructure. *Data Sci. J.* **24**, 29 (2025).
99. GBIF Science Review 2020. *Global Biodiversity Information Facility* https://doi.org/10.35035/BEZP-JJ23 (2021).
100. Berendsohn, W. G., Chavan, V. & Macklin, J. Summary of recommendations of the GBIF task group on the global strategy and action plan for the digitisation of natural history collections. *Biodivers. Inform*. https://doi.org/10.17161/bi.v7i2.3989 (2010).
101. Jennings, L. et al. Governance of Indigenous data in open Earth systems science. *Nat. Commun.* **16**, 572 (2025).
102. Lin, Y., Guan, Y., Asudeh, A. & Jagadish, H. V. Identifying insufficient data coverage in databases with multiple relations. *Proc. VLDB Endow.* **13**, 2229–2242 (2020).
103. Jennings, L. et al. Applying the 'CARE Principles for Indigenous Data Governance' to ecology and biodiversity research. *Nat. Ecol. Evol.* **7**, 1547–1551 (2023).
104. Fact Sheet 3: language list by country and place. *Government of South Australia* https://www.dpc.sa.gov.au/responsibilities/multicultural-affairs/policy/interpreting-and-translating-policy (2022).
105. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2025).
106. Posit Team *RStudio: Integrated Development Environment for R* (Posit Software, 2025).
107. Kocsis, ÁT., Reddin, C. J., Alroy, J. & Kiessling, W. The R package divDyn for quantifying diversity dynamics using fossil sampling data. *Methods Ecol. Evol.* **10**, 735–743 (2019).
108. Dowding, E. M. et al. Fossils for Future and the billion dollar case for palaeontology's digital infrastructure. *Zenodo* https://doi.org/10.5281/zenodo.17828000 (2025).
109. Dowding, E. M. et al. Fossils for Future: the billion dollar case for palaeontology's digital infrastructure. *CodeOcean* https://doi.org/10.24433/CO.1586965.v1 (2025).

Acknowledgements

This Analysis was written as part of the 'Integrated Record of Ancient Life' (IRAL) working group with support from the Paleosynthesis Project. E.M.Do. thanks C. Piper, Western Australia Biodiversity Information Office and Dandjoo Biodiversity Data Repository, for her insight into public policy, curation and physical collections. Funding sources for individual investigators include the following: E. M. Dowding, D.D., W.K., Á.T.K. and B.S.—Volkswagen Stiftung, Az 96 796; E. M. Dunne—FAU Emerging Talents Initiative; E.E.S.—Leverhulme Prize, NERC grant NE/V011405/1, NERC grant NE/C001340/1; K.D.B—I.3.4 Action of the Excellence Initiative – Research University Programme at the University of Warsaw (project: PARADIVE); L.N.—National Natural Science Foundation of China (42372039); J. A.

Sessa—NSF 1928362; M.D.U.—NSF EAR 1948831; J.W.W.—NSF 2410961. Figures 1 and 3 were designed with M. Kouvari and N. M. Morales Garcia from Science Graphic Design (sciencegraphicdesign.com).

Author contributions

E. M. Dowding conceived and designed the study. E. M. Dunne and Á.T.K. acquired funding and framed the workshop. E. M. Dowding and K.C. collected the data for database diversity survey analyses. K.S.C., K.D.B., D.D., S.M.E., S.F., W.K., K.L., L.H.L., H.L., L.N., S.E.P., J.R., E.E.S., J. A. Sessa, J. A. Smith, M.D.U., J.W.W. and Á.T.K. provided technical information about database curation and software. E. M. Dowding led the writing with E. M. Dunne, J. A. Smith and J.W.W. All authors contributed reviewed, edited and approved the final version.

Funding

Open access funding provided by Friedrich-Alexander-Universität Erlangen-Nürnberg.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-026-02985-8.

Correspondence and requests for materials should be addressed to Elizabeth M. Dowding, Emma M. Dunne or Ádám T. Kocsis.

Peer review information *Nature Ecology & Evolution* thanks Oskar Hagen, Marthe Klöcking and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

[1]GeoZentrum Nordbayern, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [2]School of Natural Sciences, Geology, Trinity College Dublin, Dublin, Ireland. [3]Natural History Museum, London, UK. [4]Institute of Evolutionary Biology, Faculty of Biology, University of Warsaw, Warsaw, Poland. [5]Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA. [6]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA. [7]Smithsonian Tropical Research Institute, Panama City, Panama. [8]Natural Sciences Unit, Finnish Museum of Natural History, Helsinki, Finland. [9]Natural History Museum, University of Oslo, Oslo, Norway. [10]Centre for Planetary Habitability, Department of Geosciences, University of Oslo, Oslo, Norway. [11]State Key Laboratory of Paleobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology, Chinese Academy of Sciences, Nanjing, China. [12]Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA. [13]Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany. [14]Department of Earth Sciences, University of Oxford, Oxford, UK. [15]Academy of Natural Sciences of Drexel University, Philadelphia, PA, USA. [16]Department of Earth and Environmental Sciences, University of Minnesota Duluth, Duluth, MN, USA. [17]Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, VA, USA. [18]Department of Geography, University of Wisconsin-Madison, Madison, WI, USA. ✉e-mail: dowding.e.m@gmail.com; dunne.emma.m@gmail.com; adam.kocsis@fau.de

nature portfolio

Corresponding author(s): Dowding Elizabeth M

Last updated by author(s): 2025/11/12

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Curator review data was collected from a survey of the author team and involved reporting information as of Nov 2024 about the back-end development, data volume, and funding of the databases the author represented. Publication products was supplied by the curators who collated formal citations of their work (the database publication) in other published literature; these were then assigned keywords on broad topics within palaeontology. Database diversity dynamics data was collected from a web survey of databases first publication, most recent publication, and whether the database was still actively available and maintained. These were conducted in the last quarter of 2024 and the first of 2025.
Data and code are available at CodeOcean, Zenodo and GIThub [https://github.com/dowdingem/IRAL]. The study has no restriction on data availability.

All data, code and supplementary material:
Dowding, E. M., et al. Fossils for Future and the billion dollar case for palaeontology's digital infrastructure. [Dataset] Zenodo. (2025). https://doi.org/10.5281/zenodo.17828000

Code and relevant data:
Dowding, E. M., et al. Fossils for Future: the billion dollar case for palaeontology's digital infrastructure. [Codebase] CodeOcean. (2025) https://doi.org/10.24433/CO.1586965.v1

Data analysis

Analysis was run in R v4.5.0 using the DivDyn Rpackage (Kocsis et al 2019) which contains metrics for richness, origination, extinction, and range through diversity amongst others. The information is available on CodeOcean and the GIThub links provided. In the time series analyses,

the durations of active and recent <5 years old) databases were extended into the future and the analytical frame was truncated at 2024 to address edge effects.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data generated for the author survey, publication, products, and analysis are available within the Supplementary Material and the stable Zenodo repository Dowding, E. M., et al. Fossils for Future and the billion dollar case for palaeontology's digital infrastructure. [Dataset] Zenodo. (2025). https://doi.org/10.5281/zenodo.17828000
All code and required data are available through CodeOcean. All analyses were conducted using publicly available R packages, and the links have been provided in Methods.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study focuses on the diversity trends of community run Earth science databases, centred around Palaeontology. The study is both qualitative and quantitative, descriptions of the databases both in terms of history, publication product, and technical development were provided by the curators through survey of the author team. Quantitative analysis of the diversity dynamics (richness, origination, cumulative sum, and extinction) were run using first and last appearance data of databases that were open access and community run. Both were completed to show the diversity of databases, both in terms of their proliferation, but also

	into terms of the variation in their volume, and technical development (back end construction). These together for the basis for conclusions and recommendations.
Research sample	Open access data bases were determined by both self description and the ability of a general member of the public to gain access to the data on their own or by request to the data administrator. Community run databases were identified by negative conditions, i.e. by not being managed by industry or government.
Sampling strategy	No sample size estimates were required as the analysis focuses on trends in time series and does no significance testing. Sampling strategy was by survey of the authorship team who represent the community run databases, and the diversity dynamics was by consistent search of free online aggregators using search terms in languages with the broadest geographic spread.
Data collection	Data was collected by Dowding using surveys which consisted both of a questionnaire and a spreadsheet for database curators to enter information about the database they maintain.
Timing and spatial scale	The information on databases first appearance (publication) and last point of activity (update publication or statement) was from the first appearance of large scale compilations that were published to 2024. Databases that fit the criteria of being open access and community run globally were included.
Data exclusions	Databases that were not open access or were governmental/industry maintained were excluded. Failing either or both of these categories resulted in exclusion from analysis.
Reproducibility	The data survey techniques are available in the supplementary information in addition to the raw data, clean data, and code. R ver. 4.5.0
Randomization	N/A
Blinding	Blinding was not required as there are no experimental constraints and information from the curators were self reported and they are named authors.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*