TITLE

A globally integrated structure of taxonomy to support biodiversity science and conservation

AUTHOR NAMES

Emily L. Sandall[1,2,*], Aurore A. Maureaud[1,2,3,*], Robert Guralnick[4], Melodie A. McGeoch[5], Yanina V. Sica[1,2], Matthew S. Rogan[1,2], Douglas B. Booher[1,2], Robert Edwards[1,2,6], Nico Franz[7], Kate Ingenloff[1,2,#], Maisha Lucas[1,2], Charles J. Marsh[1,2], Jennifer McGowan[1,2,8], Stefan Pinkert[1,2,9], Ajay Ranipeta[1,2], Peter Uetz[10], John Wieczorek[11], Walter Jetz[1,2,12]

(*equal contribution)

Affiliations:

[1] Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT, 06520, USA
[2] Center for Biodiversity & Global Change, Yale University, New Haven, CT, 06520, USA
[3] Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ, USA
[4] Department of Biodiversity Informatics, Florida Museum of Natural History, University of Florida, Gainesville, FL, USA
[5] Securing Antarctica's Environmental Future, Department of Environment and Genetics, LaTrobe University, Melbourne, Australia
[6] Cleveland Museum of Natural History, Cleveland, OH, USA
[7] School of Life Sciences, Arizona State University, Tempe, AZ, USA
[8] The Nature Conservancy, Arlington, VA, USA
[9] Department of Conservation Ecology, University of Marburg, Marburg, Germany
[10] Center for Biological Data Science, Virginia Commonwealth University, Richmond, VA, USA
[11] Museum of Vertebrate Zoology, University of California, Berkeley, CA, USA
[12] E.O. Wilson Biodiversity Foundation, Durham, NC, USA
[#] Current Address: Global Biodiversity Information Facility, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark

Corresponding Authors: Emily L. Sandall sandall.emily@gmail.com and Aurore A. Maureaud aurore.aqua@gmail.com

KEYWORDS

taxonomic backbone, integrative science, data linkage, social infrastructure, biodiversity conservation

ABSTRACT

All aspects of biodiversity research, from taxonomy to conservation, rely on data associated with species names. Effective integration of names across multiple fields is paramount and depends on coordination and organization of taxonomic data. We assess current efforts and find that even key applications for well-studied taxa still lack commonality in taxonomic information required for integration. We identify essential taxonomic elements from our interoperability assessment to support improved access and integration of taxonomic data. A stronger focus on these elements has the

potential to involve taxonomic communities in biodiversity science and overcome broken linkages currently limiting research capacity. We encourage a community effort to democratize taxonomic expertise and language in order to facilitate maximum interoperability and integration.

MAIN TEXT

Biodiversity and conservation sciences rely on taxonomic data

Taxonomy (see Glossary) provides the fundamental units around which we organize, assess, and mediate the components of biodiversity for research and conservation [1–4]. Both research and conservation use of taxonomic names has rapidly expanded online in recent decades [5,6]. With new monitoring frameworks such as the Post-2020 Global Biodiversity Framework [7] and technologies such as DNA barcoding [8,9], this trend is poised to continue. Differing needs and **values** between communities producing and using such data are usually centered around discipline- or funding-specific goals without necessarily a consideration of broader utility [10–12]. The resulting dynamic nature, heterogeneity, and bias in taxonomic data might not be obvious to users, but can have large effects on research, cultural, and **biodiversity conservation** outcomes [13–16].

Binomial **nomenclature** came about as a standardized and shared means to reference the identity of organisms, complementing **vernacular names** and descriptions based on appearance and cultural relevance [17]. With the vast increase of formally described species since Linnaeus' time, a key challenge has been to track the changes in the **taxon concept** delineating a taxon [18], represented by **accepted names** and their **synonyms** as well as by scientific nomenclature (and its regulation by codes [19,20]). **Scientific names** are used by researchers who typically associate them with a **species concept,** physical **specimens** [21], or other data and share those resources in databases and the literature [22]. Each of these elements—taxon or species concept, accepted name, and synonym(s)—can be subject to revision, because of new scientific evidence. Acceptance of taxonomic revisions is a scientific process based on objective data but also on a variety of cultural practices and norms. As a result, multiple taxonomic structures and concepts remain in use across domains of application and time periods.

The dynamic nature and multiplicity of taxonomic frameworks is further compounded by the different types of data associated with names in biodiversity repositories, including spatial, functional, genetic, and physical data [2,23] (Figure 1). The **taxonomic backbone**—connecting accepted names to synonyms in a taxonomic hierarchy—is often presented as a **global species list** [24]. That list of names forms the key enabler of subsequent synthesis for linking different data sources and/or types that use different names for the same species in support of **integrative science**, **interdisciplinary research**, and conservation [1,2,25].

Key to overcoming complexities in achieving **taxonomic integration** is enabling **interoperability** across disparate data sources. Here, we present how different objectives in taxonomy and biodiversity informatics have led to opportunities and challenges in interoperability across taxonomic data sources. We highlight the elements needed to support a more accessible, effective, and diverse use of taxonomic data. We suggest a combination of these elements into a framework to facilitate and assess taxonomic integration by the community of users for broad multipurpose utility, access, and longevity.

Needs for taxonomic integration

Many subdisciplines in biodiversity sciences and conservation are inherently intertwined with taxonomic data. Three examples illustrate the broad significance of taxonomic integration across data sources (Figure 1):

Human health—For zoonotic diseases and viruses such as SARS-CoV-2, accurate taxonomic names, from virus strain to host species, are key to integrating genetic, spatial and even clinical data for assessments and impact mitigation [26–28]. Quality assured taxonomic synthesis is relevant to governmental authorities across health sectors from local to national and multinational jurisdictions. This enables targeted research and communication into the origin, severity, and threats posed by such outbreaks (Figure 1A).

Species invasions—The spread of invasive species is causing long-term challenges for biodiversity and humanity. Members of *Opuntia* (Cactaceae), a widespread genus of cacti, including the common ornamental prickly pear native to the Americas, are now established across continents (Figure 1B). Differences in taxonomic treatments of *Opuntia* subspecies have significantly delayed early detection and management [29], a problem that could be overcome through robust taxonomic **harmonization** and updated rapid detection tools (e.g., field guides, phone applications).

Species assessments—Each of 19,327 currently recognized butterfly species have on average six synonyms [30], although some species such as the common palearctic butterfly *Plebejus argus*, have as many as 160 [31] (Figure 1C). Assessing distributions to track threat status and population declines often requires significant efforts combining spatial data, natural history information, and taxonomic expertise, all of which can be under a variety of names and taxon concepts in flux.

Biodiversity studies and conservation interventions increasingly rely on more than one data source or type [25,32]. The above examples illustrate the large array of questions and integrated data usage from basic to applied that rely on the common language of taxonomy and multi-source harmonization and integration.

Reliance on a complex landscape of taxonomic databases and perspectives

The current landscape of taxonomy sources in databases

Taxonomists and other key actors have addressed the need for integration through the development of taxonomic databases. Taxonomists and taxonomy users may engage with taxonomic data through many entry points and mechanisms, such as informatics innovations, computational and storage capabilities, and novel online engagement avenues like mobile apps [33]. These advances have catalyzed growing efforts in understanding semantic alignment and relationships of different taxonomic concepts, enhancing the potential for linking data across multiple sources [34,35]. In particular, stable identifiers reference taxonomic names and concepts and allow tracking of both with transparency, though there is to this day no global system for all taxa [36–38]. Initiatives to standardize, maintain, and organize relevant communities around taxonomic backbones have made important progress towards this goal. Yet, taxonomic efforts often face regional- [39], taxonomic- [40], temporal- [41], or funding-specific [42] constraints, leading to a spectrum of longevity, interoperability, and maintenance hurdles that limit effective research and conservation applications [43–45].

Broadly, we can distinguish three levels of taxonomic databases. Primary databases aim to produce a taxonomic backbone for one taxon, linking experts and the primary literature together to cover the taxonomic knowledge [46]. Secondary databases maximize the list of names through aggregating primary databases. They typically cover more than a single taxonomic group and similarly aim to be authoritative in their field. One of the leading global initiatives is The Catalogue of Life (COL), which relies on numerous primary databases (produced by many experts) to update the catalogue [47]. Third, biodiversity databases build a taxonomy with the goal of combining available biodiversity data rather than as its primary objective. Such efforts may ‘mix and match’ between primary authorities, add further harmonization, or implement customized updates to create more comprehensive species lists and taxonomic backbones. As the largest **biodiversity data aggregator**, the Global Biodiversity Information Facility (GBIF) [48] currently harmonizes over 2.3 billion occurrence records against a taxonomic backbone [49,50], informed by many dynamic taxonomic lists and biodiversity databases. GBIF and COL are now linked via ChecklistBank, created to share curated and standardized lists of names [51]. All databases are interlinked but rarely fully interoperable due to separate maintenance timetables, varying taxonomic classifications, and dynamic taxonomic advances [25,52,53].

Diverse communities and values around taxonomies

Collaboration is fundamental in interdisciplinary science, including individuals and communities with diverse perspectives, contributions, and project emphases [54,55]. Taxonomists, biodiversity researchers, and conservationists have legacies and values that position their interactions with taxonomic data. For taxonomists to successfully describe species and maintain nomenclatural continuity, they focus their work on legacy, history, and taxonomic specialty [56]. Where appropriately incentivized, experts who assemble large-scale biodiversity resources maintain data relations, harmonization, and standards while the data itself constantly changes. Researchers (including taxonomists) relying on analysis-ready, taxonomically harmonized data can lead synthesis and conduct transparent analysis to make their work broadly available as part of the scientific enterprise. **Conservation managers** and **decision-makers** use taxonomic data to inform strategies and conservation plans [57]. Integrating perspectives across communities evolving around taxonomy inherently brings a diversity of values in how data is ultimately structured and consumed. The consideration of these values is key as they can be the source of innovations, but also misunderstandings and conflicts in the community [21,58,59].

With so many taxonomic sources, governance and practices around taxonomic databases become complex. Progress towards interoperability and maintenance depend on communities coordinating efforts to reconcile different concepts and semantics across databases.

The challenge of taxonomic integration

A case study of limited taxonomic interoperability across taxa and data sources

To test and gain a deeper understanding of the interoperability challenge, we applied our experience in assembling taxonomic data for nine groups of terrestrial and marine plants, vertebrates, and invertebrates within Map of Life (MOL, Box 1) [60]. We expect MOL taxonomies to provide a reasonable snapshot across important taxonomic groups and realms because they cover multiple underlying sources with the aim of maximizing integration (see Online Supplemental Table S1 and Box 1). As a simple but straightforward measure of interoperability, we quantified the portion of species-level names from widely used databases that could not be matched against MOL taxonomies. We selected databases covering broad applications and data categories from taxonomy, biodiversity, and conservation (see Methods in Box 1).

Our analysis showed often surprisingly limited levels of interoperability that varied across sources and taxa (Box 1). For instance, we found 15% of butterfly names from both COL and the National Center for Biotechnological Information (NCBI) that did not have a match with the MOL list of names. This is substantial, and even with a large list of synonyms (Online Supplemental Table S1), the lack of a coordinated database of names in the butterfly community may be the source of such mismatch. The lowest interoperability was found with the Open Tree of Life names source, where in six of the nine taxonomic groups 12% to 47% of names could not be matched against MOL names. This is partly due to lack of interoperability and convention among higher rank name searches in the database or taxon names including gene region. This case study shows that, because interoperability is often only partially achieved, taxonomic integration partly remains the task and responsibility of the end-user by making choices in how names are matched across sources in their own work, further complicating future interoperability.

Key elements of taxonomic data structure supporting interoperability

To guide users in assessing integration and understanding the expertise involved in the list of names from each database they may need, we suggest that at least six elements are required to support successful data integration in a particular taxon (Table 1). These elements were identified building on available literature, expertise, and our experience with synthesis [18,49,58]. The first element, a "Global authoritative list", comprehensively catalogs all accepted names, like a digital taxonomic catalog. Secondly, a "Synonyms list" directly linked to accepted names in the "Global authoritative list", matches older and divergent names in spelling, subsumed rank, or no longer valid names with current data. Third, "Authorship information" comprising the author name and year of publication associates a taxon name with its original publication and description. The fourth element, "Name sources and timestamps", captures the original database source name and version. It ensures reproducibility and transparency as sources and reported names change over time. Fifth, the "Name instance", such as an observation or specimen, provides an instance of usage of the name in a data source. Finally, "Taxon concept in space and time" links names to records of the taxon concept used, documented with dates and locations to provide the associated spatio-temporal context and eventual needs for revision.

The interdependent significance of the elements is underappreciated outside the field of taxonomy, partially due to a lack of common vocabulary among users, and they are most meaningful when considered together. For instance, the content of the "Synonyms list" is dependent on which source(s) is selected for the "Global authoritative list". Similarly, "Taxon concept in space and time" may already be implicit from other elements but requires explicit tracking (e.g., revision**, splitting, lumping**). These elements—together considered as a Globally Integrated Structure of Taxonomy (GIST) framework—represent the minimum required for taxonomic integration to be accurate and complete.

Assessing the coverage of key elements

Although several elements have seen substantial development in recent years, shortcomings in a single element can constrain overall interoperability. To gauge the magnitude and variation of this issue, we conducted a simple assessment of elements coverage for the same nine taxa analyzed above (Box 2). Overall, no group seems to present a full integration elements score. "Global authoritative lists" are seemingly well-curated except for butterflies, and potentially the source of lower interoperability levels for this group (Box 1). "Synonyms lists" appear most challenging in butterflies and dragonflies, where comprehensive lists require compilation of many sources [30,61]. "Authorship information" is not consistently available across sources and species, and only reptiles received the maximum score for this element. “Name sources and timestamps” were well integrated, but the “Taxon concept in space and time” were consistently poorly available or unattainable. This low coverage highlights an avenue needing improvement, and some ongoing efforts aim to improve the integration of the “Taxon concept in space and time” [62,63].

Toward improved taxonomic data integration

The GIST elements can enhance communication and understanding of the challenges around taxonomic integration by providing a standardized vocabulary that can be readily used across databases, communities, and disciplines.

Recognize the ongoing challenge

The examples and analyses we presented above draw attention to the challenges and importance of taxonomic data integration. We provide methods for improved assessments of the interoperability and integration status, recognizing that several levels of interoperability exist. Matching between names (Box 1) might nevertheless bear inaccurate links, especially if key elements are not well integrated (Box 2). For instance, names may match, but represent different species concepts, creating uncertainty and bias in scientific and applied outcomes. Additionally, as names and concepts change, the linkages depending on them can become unstable and harder to reliably track. As taxonomic revisions and species additions create a dynamic flow of names and concepts, better ways for aggregators and users to track updates across data sources, such as stable identifiers, are lacking.

Evaluate the level of integration in your own work

Taxonomic data users should be empowered to examine their data and the decisions made based on them, even for those who may not have taxonomic expertise. We anticipate that expanding the use of the GIST elements across databases will further facilitate the implementation of the FAIR principles (findable, accessible, interoperable, reusable) [64] and improve access to information about sources (e.g., “Names Sources & Timestamps”, “Authorship Information” elements). Moving forward, we recommend documenting how available and integrated the elements are across databases by following the assessment criteria in Box 1 and 2. Even though databases may respond to distinct codes of nomenclature [19,20,65], or models of governance [35,44], the GIST elements are simple enough to be transferable across all databases and taxa and can rely on the DwC standards (Table 1) [49].

Communicate your challenges and opportunities to facilitate innovation

Both producers and users of data need to be a part of the solutions for broader integration innovations [58,66]. Considering the diversity and complexity in the landscape of taxonomic databases and contributors, it is essential to prioritize consistent engagement and shared responsibility between all actors on the integration process to identify synergies and address challenges. One important innovation enabling coalition-building will be the implementation of stable identifiers for all names and sources, building upon existing biodiversity informatics and natural history attribution infrastructure (i.e. TDWG[I] and Bionomia[II]). Complementary to stable identifiers, the GIST lays out a framework to make biodiversity data and information readily accessible to users, most of whom are not taxonomic experts. Establishing an interdisciplinary community representing all actors central to the future of interoperable taxonomically informed projects must become a priority, and this will need further coalescence to sustain it [3,10,35,55].

CONCLUDING REMARKS

Taxonomy is central for the integration of data sources informing biodiversity research and management. In a time of rapid biodiversity change and increasing data volumes, renewed vigor around valuing, funding, developing, and integrating taxonomy and its interdisciplinary community is paramount. We identified important gaps in taxonomic data integration based on a framework allowing simple and transparent assessments, which can be replicated in other cases by data aggregators and users. This framework highlights opportunities for the scientific community to realize and expand the potential of taxonomic data (see Outstanding Questions). Rather than generating novel data and social infrastructures, structuring efforts around integration has the potential to help the community in building and using taxonomic data for improved scientific and conservation outcomes.

ACKNOWLEDGMENTS

This project originated from discussions in the taxon expert group hosted at the Center for Biodiversity and Global Change at Yale University, partly funded by the E.O. Wilson Foundation, including E.L.S, A.A.M., Y.V.S., M.R., K.I, D.B.B., S.P., R.E., M.L., C.J.M., A.R., and led by C.J.M for 2020-2021 and E.L.S. for 2021-2022. We thank Scott Loarie for valuable discussions, and members of the Center for Biodiversity and Global Change and the Map of Life project for general feedback during the development of this project. We thank Mark John Costello for valuable discussions during the development of this project and his contributions to the manuscript.

DECLARATION OF INTERESTS

No interests are declared.

HIGHLIGHTS

- Taxonomic knowledge is critical to understand, catalog, and assess biodiversity and is central to measuring and achieving conservation goals, including the Post-2020 Framework of the Convention on Biological Diversity.
- Taxonomy is a centuries-old discipline, but its tools, diversity of users, and applications are constantly expanding and evolving.
- The lack of trackable and interoperable taxonomic data inhibits data integration and knowledge transfer across communities and disciplines, constraining conservation applications.
- We propose a globally integrated structure of taxonomy composed of six elements to increase understanding of taxonomic interoperability status across the fields of taxonomy, biodiversity, and conservation.
- Normalizing taxonomic integration assessments by data aggregators and users will help inform and track progress towards cross-group name integration, biodiversity synthesis science, and applications.

OUTSTANDING QUESTIONS

- Taxonomic integration across data sources and taxonomic groups is challenging. How will taxonomic integration evolve and be sustained with a continued explosion of biodiversity data and names in the coming decades, especially with the expected increase in DNA barcoding?
- Interoperability and key elements integration can be improved with the development of stable identifiers and their uptake by taxonomists and users. How will stable identifiers and other informatics solutions help integration with increasing volumes of taxonomic data?
- One of the greatest sources of improvement across important elements of taxonomic integration is the documentation of the taxon concept in space and time. While there are important advances on this subject for some taxonomic groups (e.g., mammals) and through efforts such as that of the Biodiversity Information Standards (TDWG) Taxonomic Concept Schema (TCS) Maintenance Group, will it be scalable to all other taxonomic groups and databases?
- Will organizations, such as TDWG, be enough to support the interaction and data exchange between generators, aggregators, regionalized efforts, local communities, and users of taxonomic data?
- Taxonomic communities exist all around the world, but most databases and museum collections are hosted in a few countries while most biodiversity is elsewhere. How can a more global model of participation and collaboration be facilitated to ensure different types of biodiversity knowledge and community perspectives are incorporated?

FIGURES

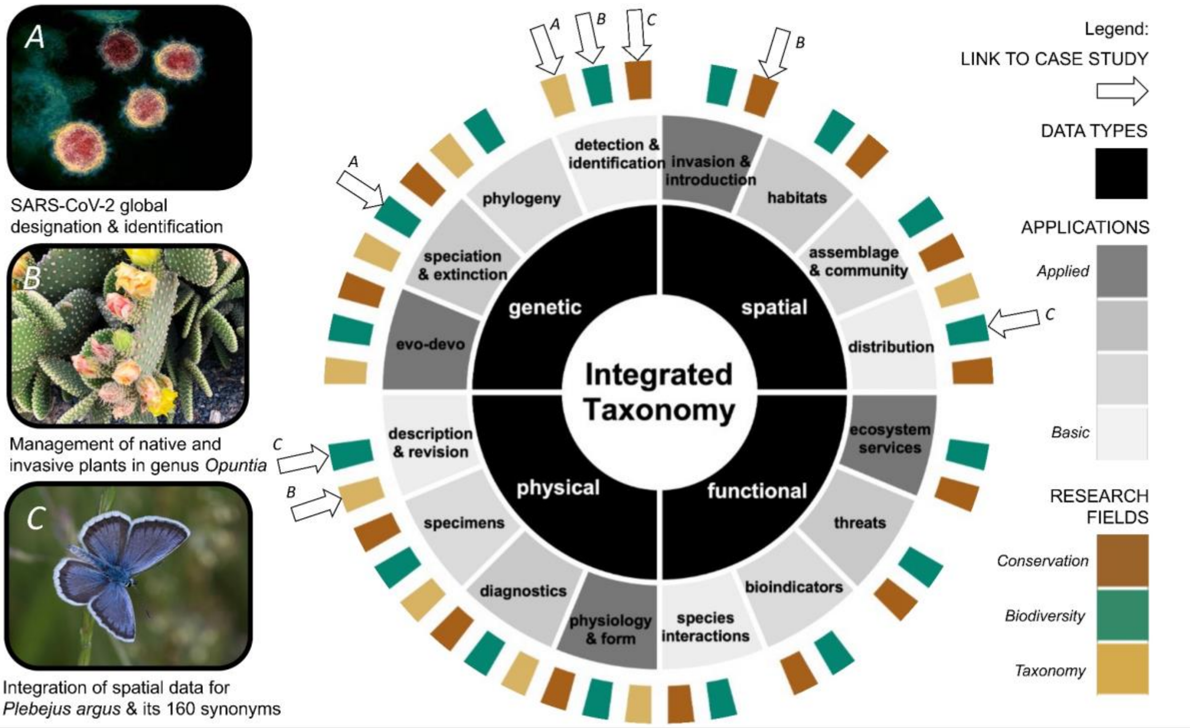


Figure 1. Research themes and examples with associated data types relying on taxonomic integration. Innermost ring (black): main data categories. Middle ring (gray gradient): data applications (foundational to applied) across the four data categories. Outer ring (color categories): example research questions and applications (from taxonomy, biodiversity, and conservation). Arrows on the outermost edge of the rings denote a linkage with one of three examples (A, B, C), illustrating how integration facilitates a transparent connection between primary data, biodiversity analysis and practice and could avoid problems downstream: (A) SARS-CoV-2, coronavirus 2 global designation and identification. (B) Management of invasive plants in *Opuntia*, a genus of cacti species. (C) Spatial range comparison of the butterfly *Plebejus argus*, characterized by 160 synonyms. Photo sources and credits are documented in Online Supplemental Information Table S3.

TABLES

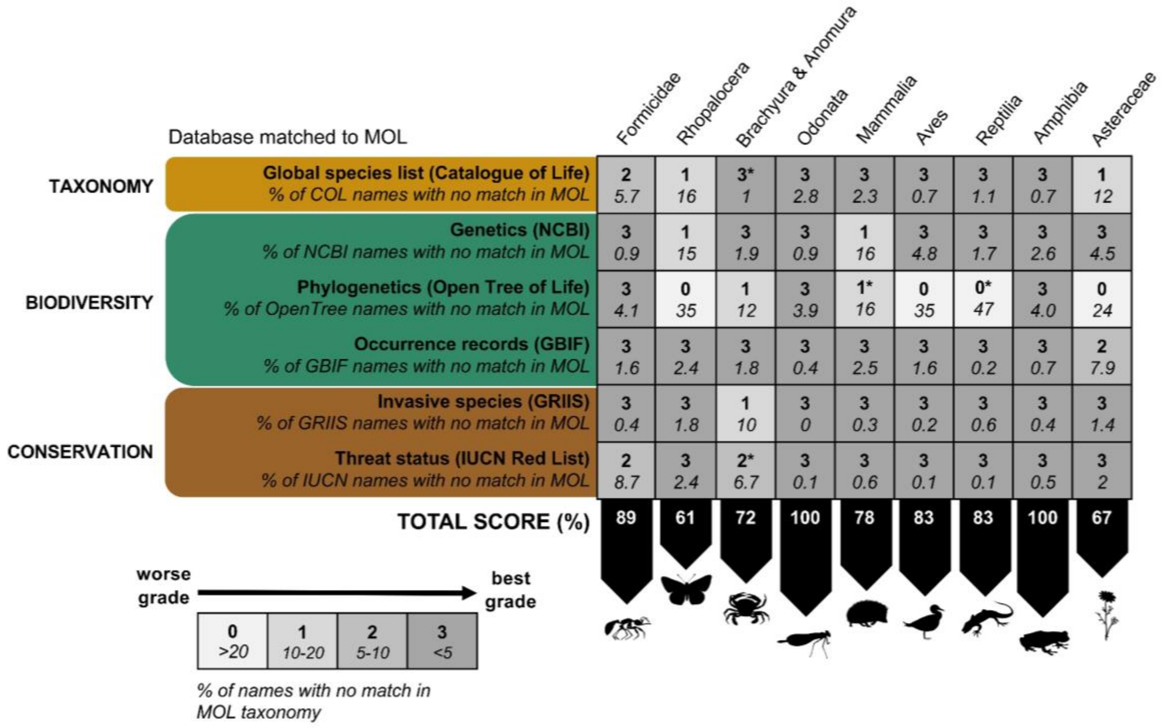
Table 1: Description of the key elements for taxonomic integration. Each element is described, illustrated by an example themed around the American bison *Bison bison*, and linked to corresponding Darwin Core (DwC) [49] standard terms. Accepted as a species in Mammal Species of the World [67] version 3 from 2005, with subspecies lumped as synonyms based on phylogenetic evidence [68], there are occurrence records in GBIF as of 25th February 2022 under ten scientific names, including: *Bos bison*, *Bison bison bison*, *Bison bison athabascae*, and *Bison bison*. The list of the DwC terms is available here: https://dwc.tdwg.org/terms/.

Element	Description	Example	DwC standard
Global authoritative list	*All accepted names for the members of a taxonomic group.* List of accepted names for a taxon defined by all species contained in a particular taxonomic rank (such as family, order, class or kingdom). This can be a preferred taxonomic authority or a compilation of accepted names in the absence of an authority.	Accepted name for the American bison is *Bison bison* from the list Mammal Species of the World (MSW, v.3)[III]	\`taxonomicStatus\`: accepted; \`taxonRank\`: species (or below). Binomial name described by a minimum of \`genus\` and \`specificEpithet\`
Synonyms list	*Other names that can be matched to the \`Global authoritative list\`.* List(s) of alternative names, including spelling differences and names that are not accepted but can be matched to the accepted names for the taxon. This list is appended to the \`Global authoritative list\` of names when clear matching can be done with accepted names.	*Bison bison* Synonym: *Bos bison*, linked to occurrence points in GBIF[IV]	\`taxonomicStatus\`: other than accepted; \`taxonRank\`: species (or below). Binomial name described by a minimum of \`genus\` and \`specificEpithet\`
Authorship information	*Author and publication year of a name to link name history.* Name of the author and year associated with the global list accepted and synonym names. This information is crucial when associating synonyms with accepted names.	For *Bison bison*: (Linnaeus, 1758)[V]	\`scientificNameAuthorship\`; \`namePublishedInYear\`
Name source & timestamp	*Source and date of an individual name or version of a list or data source.* This information facilitates our understanding of how names have changed over time and across authorities.	Name *Bison bison* is accepted by the source MSW, v.3 from 2005[III]	\`references\`;\`datasetID\`;\`datasetName\`; \`modified\` (but need to customize for the date when the data were retrieved)
Taxon concept in space and time	*Names associated with dates and locations, taxon concept history.* Track record of species name changes through time for careful use of the data. Requires the year the change was published and accepted, as well as information on the locality associated with each name used	*Bison bison bison* and *Bison bison athabascae* are subspecies of *Bison bison*.These names are linked to occurrence	Location class terms https://dwc.tdwg.org/terms/#location \`year\`

		points in GBIF[V,VI,VII]	
Name instance	*Instance of usage of a taxon name at a given time and place.* Physical observations or collections of organisms.	*Bison bison*, observed in Idaho, US 11th July 2017[VIII]	For example `basisOfRecord`;`collectionCode`;`institutionCode`;`occurrenceID`;`catalogNumber`;`associatedSequences`;`organismID`;`eventID`;`eventDate`;`verbatimEventDate`;`locationID`,`higherGeography`; `locality`;`decimalLatitude`; `decimalLongitude`;`geodeticDatum`;`coordinateUncertaintyInMeters`

Box 1. Interoperability across databases of taxonomy, biodiversity, and conservation sciences.

We used the Map of Life (MOL) [60] taxonomic database to evaluate interoperability across a range of biodiversity databases, illustrating some data categories and use cases from Figure 1.



	Database matched to MOL	Formicidae	Rhopalocera	Brachyura & Anomura	Odonata	Mammalia	Aves	Reptilia	Amphibia	Asteraceae
TAXONOMY	**Global species list (Catalogue of Life)** *% of COL names with no match in MOL*	2 5.7	1 16	3* 1	3 2.8	3 2.3	3 0.7	3 1.1	3 0.7	1 12
BIODIVERSITY	**Genetics (NCBI)** *% of NCBI names with no match in MOL*	3 0.9	1 15	3 1.9	3 0.9	1 16	3 4.8	3 1.7	3 2.6	3 4.5
	Phylogenetics (Open Tree of Life) *% of OpenTree names with no match in MOL*	3 4.1	0 35	1 12	3 3.9	1* 16	0 35	0* 47	3 4.0	0 24
	Occurrence records (GBIF) *% of GBIF names with no match in MOL*	3 1.6	3 2.4	3 1.8	3 0.4	3 2.5	3 1.6	3 0.2	3 0.7	2 7.9
CONSERVATION	**Invasive species (GRIIS)** *% of GRIIS names with no match in MOL*	3 0.4	3 1.8	1 10	3 0	3 0.3	3 0.2	3 0.6	3 0.4	3 1.4
	Threat status (IUCN Red List) *% of IUCN names with no match in MOL*	2 8.7	3 2.4	2* 6.7	3 0.1	3 0.6	3 0.1	3 0.1	3 0.5	3 2
	TOTAL SCORE (%)	89	61	72	100	78	83	83	100	67

0	1	2	3
>20	10-20	5-10	<5

Figure I. Name matches between MOL and key databases. Arrow lengths at the bottom of the matrix are proportional to the score attributed to each taxon. ‘*’ specifies when the taxon could only be partly matched due to data availability or inability to access part of the data.

Methods

The basis for comparison were nine synthesized taxonomies, compiled by taxonomic specialists to provide comprehensive lists of extant species [60]. MOL is not developing the accepted taxonomy, but rather using authoritative lists and known sources to integrate in the platform. For each taxonomic group, we drew upon established sources of accepted names in 2021, typically based on one established main name source authority when available, combined with a range of sources addressing synonyms and orthographic variants (see Online Supplemental Table S1). The nine taxa assessed include ants (Formicidae), butterflies (Rhopalocera), crabs (Brachyura and Anomura), dragonflies and damselflies (Odonata), mammals (Mammalia), birds (Aves), reptiles (Reptilia), amphibians (Amphibia), and flowering plants (Asteraceae).

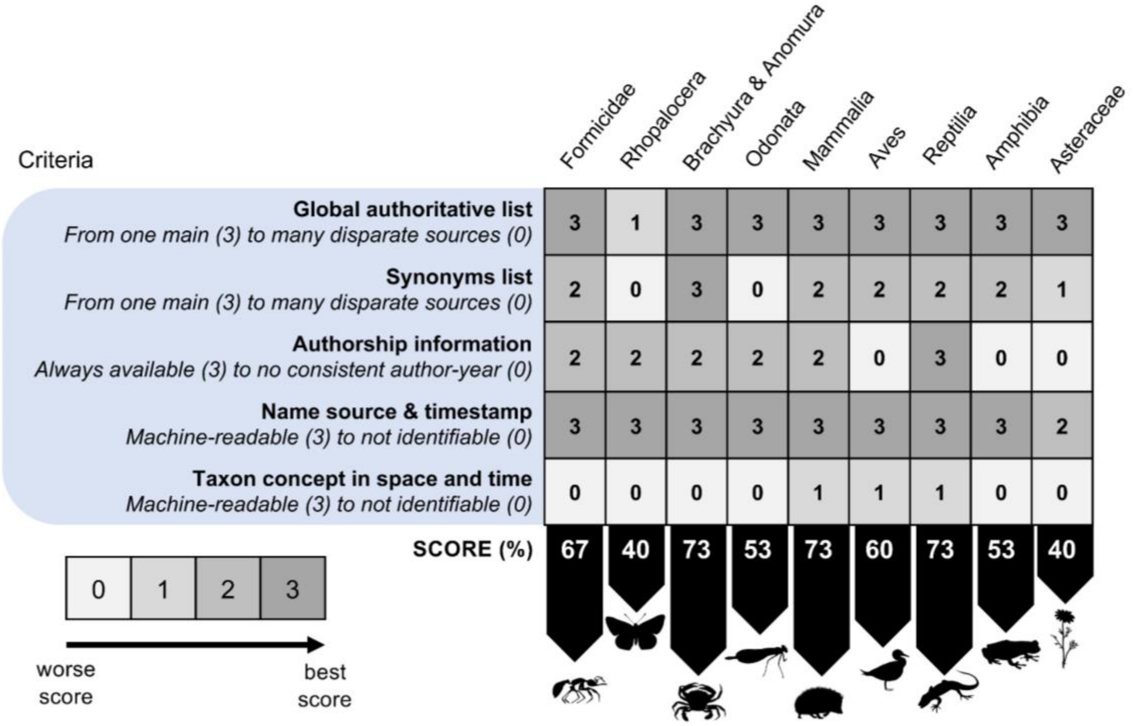
We quantified how well names from key databases match these taxonomies, either directly or via synonyms or orthographic variants. Key databases were: Catalogue of Life COL [47], Open Tree of Life [69], National Center for Biotechnology Information NCBI [70], Global Biodiversity Information Facility GBIF [48], Global Register for Introduced and Invasive Species GRIIS [71], and International Union for the Conservation of Nature IUCN Red List of Threatened Species [72]. Lists of all names present (accepted and others) from these databases were retrieved at the end of 2021 to match MOL names timestamps.

For each taxon and data source combination, we calculated the percentage mismatch in terms of names in the data source absent from the MOL taxonomy (indicated in italics in Figure I). Taxa were graded (0–3) according to the name linkage percentage: grade 0 is attributed when >20% of names in the database do not match against MOL names, 1 is given for 10-20%, grade 2 for 5-10%, and grade 3 for <5%. The grade of 3 corresponds to the most interoperable case for a taxonomic group and data source: when the percentage of names that do not match against MOL is the lowest. Each taxon is then attributed a total score as a percentage based on the sum of the individual scores relative to the potential maximum score.

The underlying data, processing steps, and the code to conduct the name match and grading are open access [73,74].

Box 2. Coverage of minimum essential elements for taxonomic integration.

Figure I. Evaluation of the integrative elements for nine taxa. Taxa were graded from 0 (worst) to 3 (best) for attributes (italics) representing five of the key elements. The length of arrows at the bottom are proportional to the total score attributed to each taxonomic group.



Criteria	Formicidae	Rhopalocera	Brachyura & Anomura	Odonata	Mammalia	Aves	Reptilia	Amphibia	Asteraceae
Global authoritative list *From one main (3) to many disparate sources (0)*	3	1	3	3	3	3	3	3	3
Synonyms list *From one main (3) to many disparate sources (0)*	2	0	3	0	2	2	2	2	1
Authorship information *Always available (3) to no consistent author-year (0)*	2	2	2	2	2	0	3	0	0
Name source & timestamp *Machine-readable (3) to not identifiable (0)*	3	3	3	3	3	3	3	3	2
Taxon concept in space and time *Machine-readable (3) to not identifiable (0)*	0	0	0	0	1	1	1	0	0
SCORE (%)	67	40	73	53	73	60	73	53	40

Methods
We assigned scores varying from 0 to 3 according to specific qualitative and semi-quantitative criteria. We derived an overall ranking from the percentage of the sum score relative to potential maximum score. The scores were attributed based on careful evaluation of the sources used for the MOL taxonomies, but remain necessarily subjective (see Online Supplemental Information Table S2):

The list of names ("Global authoritative list" and "Synonyms list") were graded as follows:

- 0 if many non-machine-readable sources in the literature without a main authority were needed
- 1 if complementary authoritative were needed and not all machine-readable
- 2 if complementary authoritative sources were needed but all machine-readable
- 3 if there is one main authoritative source that is machine readable.

The “Authorship information” was graded as follows:

- 0 if there is no consistent author and year
- 1 if no author name or no year are consistent
- 2 if no author name but year are consistently available
- 3 if author name and year are consistently available

The “Name sources & timestamp” and “Taxon concept in space and time” were graded as follows:

- 0 if it cannot be identified
- 1 if it is not clearly designated but can be identified
- 2 if it is clearly designated
- 3 if it is identified, clearly designated and machine readable.

GLOSSARY

Accepted name. The scientific name of a taxon that has been formally chosen among alternative names (synonyms) by scientific experts.

Binomial nomenclature. System of naming species using two Latin terms, genus (rank above species) and specific epithet.

Biodiversity data aggregator. A digital platform for collecting and sharing biodiversity data.

Biodiversity conservation. Scientific discipline and practice for maintaining and protecting natural resources and ecosystems.

Conservation manager. An individual responsible for actions in an organization aiming at the protection of the environment, landscape, seascape, biodiversity, and/or wildlife.

Decision-maker. An individual responsible for making strategic decisions based on multiple variables and dependent on the amount of information available.

Global species list. List of accepted names covering all species within a taxonomic group at a given rank (e.g., family, order, class or kingdom). It may be approved by a particular taxonomic authority or a compilation of accepted names in absence of a defined authority.

Harmonization. Process of joining and integrating data from multiple sources to make a unified dataset.

Integrative science. Science that brings together multiple disciplines, taxonomic groups, spatial, temporal, and organizational scales, and/or communities, and allows exploring and testing new paradigms to transform current practices.

Interdisciplinary research. Science related to more than one discipline.

Interoperability. Ability for databases or systems to exchange information with minimal manual effort by the end user.

Nomenclature. Set of rules governing the correct form of scientific names in relation to aspects such as priority and spelling.

Scientific name. Recognized scientific name of organisms, typically a binomial name including genus and species.

Species. Group of organisms that can be considered one taxonomic unit, typically as the lowest taxonomic rank that has an accepted name.

Species concept. The conceptual framework for forming hypotheses that represent species.. Biological or phylogenetic species concepts are examples. See **taxon concept**.

Species splitting and lumping. Used in the context of changing delimitation of a species due to varying taxonomic opinion, whereby one species may be divided into several species, or several species are grouped into one. This is distinct from the process of describing new species or adding new synonyms.

Specimen. Physical example of an organism.

Synonym. An alternative name to the accepted name. In a taxonomic backbone scheme, synonyms are appended under accepted names as a “child” term.

Taxon concept. Set of organisms and their characteristics that form a hypothesis representing a taxon and distinguishing it from other taxa, which can vary between particular authors and change over time with new data or specimen evidence. See **species concept**.

Taxonomic backbone. A data structure for matching taxonomic synonyms to accepted names, within a hierarchy.

Taxonomic integration. Integrative science focused on new interfaces between taxonomies, or between taxonomies and other products or disciplines.

Taxon. A term denoting a commonly recognized unit or collective of organisms. Also called a taxonomic group.

Taxonomist. An individual who identifies, classifies, and/or describes taxa.

Taxonomy. Science of the classification of organisms.

Values. The moral, societal, or epistemic basis for actions.

Vernacular name. A common, non-scientific name for an organism, which may be regional or draw on the features of an organism.

REFERENCES

1. Jetz, W. *et al.* (2021) Include biodiversity representation indicators in area-based conservation targets. *Nat. Ecol. Evol.* DOI: 10.1038/s41559-021-01620-y
2. Sigwart, J.D. *et al.* (2018) Measuring Biodiversity and Extinction—Present and Past. *Integr. Comp. Biol.* 58, 1111–1117
3. Abrahamse, T. *et al.* (2021) *The Global Taxonomy Initiative in Support of the Post-2020 Global Biodiversity Framework.*, Secretariat of the Convention of Biological Diversity
4. Deans, A.R. *et al.* (2012) Time to change how we describe biodiversity. *Trends Ecol. Evol.* 27, 78–84
5. Costello, M.J. *et al.* (2015) Conservation of biodiversity through taxonomy, data publication, and collaborative infrastructures. *Conserv. Biol.* 29, 1094–1099
6. Heberling, J.M. *et al.* (2021) Data integration enables global biodiversity synthesis. *Proc. Natl. Acad. Sci.* 118
7. Tancoigne, E. and Ollivier, G. (2017) Evaluating the progress and needs of taxonomy since the Convention on Biological Diversity: going beyond the rate of species description. *Aust. Syst. Bot.* 30, 326
8. van Klink, R. *et al.* (2022) Emerging technologies revolutionise insect ecology and monitoring. *Trends Ecol. Evol.* 37, 872–885
9. Hebert, P.D.N. and Gregory, T.R. (2005) The Promise of DNA Barcoding for Taxonomy. *Syst. Biol.* 54, 852–859
10. Wüster, W. *et al.* (2021) Confronting taxonomic vandalism in biology: conscientious community self-organization can preserve nomenclatural stability. *Biol. J. Linn. Soc.* 133, 645–670
11. Hey, J. *et al.* (2003) Understanding and confronting species uncertainty in biology and conservation. *Trends Ecol. Evol.* 18, 597–603
12. Stropp, J. *et al.* (2022) Taxonomic uncertainty and the challenge of estimating global species richness. *J. Biogeogr.* 49, 1654–1656
13. Isaac, N.J.B. *et al.* (2004) Taxonomic inflation: its influence on macroecology and conservation. *Trends Ecol. Evol.* 19, 464–469
14. Vogel Ely, C. *et al.* (2017) Implications of poor taxonomy in conservation. *J. Nat. Conserv.* 36, 10–13
15. Correia, R.A. *et al.* (2018) Nomenclature instability in species culturomic assessments: Why synonyms matter. *Ecol. Indic.* 90, 74–78
16. Montoya, R.D. (2022) *Power of position: classification and the biodiversity sciences*, The MIT Press
17. Winston, J.E. (2018) Twenty-First Century Biological Nomenclature—The Enduring Power of Names. *Integr. Comp. Biol.* 58, 1122–1131
18. Franz, N.M. and Peet, R.K. (2009) Perspectives: Towards a language for mapping relationships among taxonomic concepts. *Syst. Biodivers.* 7, 5–20
19. International Commission on Zoological Nomenclature *et al.*, eds. (1999) *International code of zoological nomenclature: Code international de nomenclature zoologique*, (4th ed.), International Trust for Zoological Nomenclature, c/o Natural History Museum
20. Turland, N. *et al.*, eds. (2018) *International Code of Nomenclature for algae, fungi, and plants*, 159, Koeltz Botanical Books
21. Thomson, S.A. *et al.* (2018) Taxonomy based on science is necessary for global conservation. *PLOS Biol.* 16, e2005075
22. Agapow, P. *et al.* (2004) The Impact of Species Concept on Biodiversity Studies. *Q. Rev. Biol.* 79, 161–179
23. König, C. *et al.* (2019) Biodiversity data integration—the significance of data resolution and domain. *PLOS Biol.* 17, e3000183
24. Bisby, F.A. (2000) The Quiet Revolution: Biodiversity Informatics and the Internet. *Science* 289, 2309–2312
25. Feng, X. *et al.* (2022) A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. *Glob. Ecol. Biogeogr.* 31, 1242–1260

26. Gorbalenya, A.E. *et al.* (2020) The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* 5, 536–544

27. Forster, P. *et al.* (2020) Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci.* 117, 9241–9243

28. Wang, C. *et al.* (2020) The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J. Med. Virol.* 92, 667–674

29. Smith, G. *et al.* (2011) South Africa's ongoing Opuntia Mill. (Cactaceae) problem: the case of O. microdasys (Lehm.) Pfeiff. *Bradleya* 29

30. Pinkert, S. *et al.* (2022) Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. *Glob. Ecol. Biogeogr.* 31, 830–839

31. Macgregor, C.J. *et al.* (2019) Climate-induced phenology shifts linked to range expansions in species with multiple reproductive cycles per year. *Nat. Commun.* 10, 4455

32. Edwards, J.L. *et al.* (2000) Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science* 289, 2312–2314

33. Di Cecco, G.J. *et al.* (2021) Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. *BioScience* DOI: 10.1093/biosci/biab093

34. Franz, N.M. *et al.* (2016) Two Influential Primate Classifications Logically Aligned. *Syst. Biol.* 65, 561–582

35. Sterner, B.W. *et al.* (2020) Decentralized but Globally Coordinated Biodiversity Data. *Front. Big Data* 3, 519133

36. Page, R.D.M. (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Brief. Bioinform.* 9, 345–354

37. Kõljalg, U. *et al.* (2020) The Taxon Hypothesis Paradigm—On the Unambiguous Detection and Communication of Taxa. *Microorganisms* 8, 1910

38. Pyle, R.L. (2016) Towards a Global Names Architecture: The future of indexing scientific names. *ZooKeys* DOI: 10.3897/zookeys.550.10009

39. Bénichou, L. *et al.* (2018) Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy. *Eur. J. Taxon.* DOI: 10.5852/ejt.2018.475

40. McClure, C.J.W. *et al.* (2020) Towards reconciliation of the four world bird lists: hotspots of disagreement in taxonomy of raptors. *Proc. R. Soc. B Biol. Sci.* 287, 20200683

41. Kindt, R. (2020) WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. *Appl. Plant Sci.* 8, e11388

42. Britz, R. *et al.* (2020) Funding, training, permits—the three big challenges of taxonomy. *Megataxa* 1

43. Conix, S. *et al.* (2021) Towards a global list of accepted species III. Independence and stakeholder inclusion. *Org. Divers. Evol.* DOI: 10.1007/s13127-021-00496-x

44. Lien, A.M. *et al.* (2021) Towards a global list of accepted species IV: Overcoming fragmentation in the governance of taxonomic lists. *Org. Divers. Evol.* DOI: 10.1007/s13127-021-00499-8

45. Uetz, P. *et al.* (2021) A Quarter Century of Reptile and Amphibian Databases

46. Frost, D.R. (1999) Amphibian Species of the World: an Online ReferenceAmerican Museum of Natural History

47. Hobern, D. *et al.* (2021) Towards a global list of accepted species VI: The Catalogue of Life checklist. *Org. Divers. Evol.* DOI: 10.1007/s13127-021-00516-w

48. What is GBIF?[Online]. Available: https://www.gbif.org/what-is-gbif. [Accessed: 16-Jun-2022]

49. Wieczorek, J. *et al.* (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE* 7, e29715

50. Leonelli, S. (2013) Classificatory Theory in Biology. *Biol. Theory* 7, 338–345

51. Döring, M. *et al.* (2022) Introducing ChecklistBank: An index and repository for taxonomic data. *Biodivers. Inf. Sci. Stand.* 6, e93938

52. Costello, M.J. *et al.* (2013) Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. *PLOS ONE* 8, e51629

53. Grenié, M. *et al.* (2023) Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods Ecol. Evol.* 14, 12–25

54. Rolin, K. (2015) Values in Science: The Case of Scientific Collaboration. *Philos. Sci.* 82, 157–177

55. Eyster, H.N. *et al.* (2022) Why People Do What They Do: An Interdisciplinary Synthesis of Human Action Theories. *Annu. Rev. Environ. Resour.* 47, 725–751

56. Wheeler, Q.D. *et al.* (2004) Taxonomy: Impediment or Expedient? *Science* 303, 285–285

57. Kühl, H.S. *et al.* (2020) Effective Biodiversity Monitoring Needs a Culture of Integration. *One Earth* 3, 462–474

58. Garnett, S.T. *et al.* (2020) Principles for creating a single authoritative list of the world's species. *PLOS Biol.* 18, e3000736

59. Raposo, M.A. *et al.* (2017) What really hampers taxonomy and conservation? A riposte to Garnett and Christidis. *Zootaxa* 4317, 179–184

60. Jetz, W. *et al.* (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* 27, 151–159

61. Sandall, E.L. *et al.* (2022) Country-level checklists and occurrences for the world's Odonata (dragonflies and damselflies). *J. Biogeogr.* 49, 1586–1598

62. Upham, N. *et al.* (2022) Extended Taxonomic Curation: Moving beyond species lists to linking species data. *Biodivers. Inf. Sci. Stand.* 6, e93670

63. Pyle, R. *et al.* (2022) Modeling Taxon Concepts: A new approach to an old problem. *Biodivers. Inf. Sci. Stand.* 6, e93927

64. Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018

65. IPNI (2023) International Plant Names Index. *R. Bot. Gard. Kew Harv. Univ. Herb. Libr. Aust. Natl. Herb. Retrieved 05 April 2023* at <Published on the Internet https://www.ipni.org/>

66. Todman, L.C. *et al.* (2023) 'Small Data' for big insights in ecology. *Trends Ecol. Evol.* 38, 615–622

67. Wilson, D.E. and Reeder, D.M. (2005) *Mammal Species of the World: A Taxonomic and Geographic Reference*, ((3rd edn)), Johns Hopkins University Press, Baltimore, Maryland

68. Cronin, M.A. *et al.* (2013) Genetic Variation and Differentiation of Bison (Bison bison) Subspecies and Cattle (Bos taurus) Breeds and Subspecies. *J. Hered.* 104, 500–509

69. Rees, J. and Cranston, K. (2017) Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodivers. Data J.* 5, e12581

70. Schoch, C.L. *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020

71. Pagad, S. *et al.* (2018) Introducing the Global Register of Introduced and Invasive Species. *Sci. Data* 5, 170202

72. IUCN The IUCN Red List of Threatened Species. Version 2022-2. *IUCN Red List Threat. Species* at <https://www.iucnredlist.org/en>

73. Maureaud, A.A. *et al.* (2023) AquaAuma/review_gist: interoperability assessment v1.0Zenodo

74. Sandall, E.L. *et al.* (2022) Project data and code access underlying the globally integrated taxonomic structure. DOI: 10.17605/OSF.IO/9RPTQ

RESOURCES LIST

[I] https://www.tdwg.org/

[II] https://bionomia.net/

[III] https://www.departments.bucknell.edu/biology/resources/msw3/browse.asp?s=y&id=14200669

[IV] https://www.gbif.org/species/2441178

[V] https://www.gbif.org/species/2441176

[VI] https://www.gbif.org/species/9182932

[VII] https://www.gbif.org/species/7194051

[VIII] https://www.gbif.org/occurrence/2832703394