TITLE

A Globally Integrated Structure of Taxonomy supporting 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 towards systematic assessments of interoperability and integration.

MAIN TEXT

1. 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 and variation in the **taxon concept** delineating a taxon [18], as well as its **scientific nomenclature** (both formally represented by **accepted names** and their **synonyms**, and related codes [19,20]). Scientific names are used by researchers who associate them with physical **specimens** [21] or other data and share those resources in databases and literature [22]. Each of these elements—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 names with several data sources and/or types 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 tentatively 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.

2. 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 taxonomically named identities, 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.

3. Reliance on a complex landscape of taxonomic databases and perspectives

3.1. 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 concepts and allow tracking names and concepts 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 initiatives is the Catalogue of Life+ (COL), which relies on numerous primary databases (produced by many experts) to update the catalogue [45]. 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].

3.2. 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 knowledge, they focus their work on legacy, history, and 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** further 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. Efforts dedicated to interoperability and maintenance depend on communities coordinating efforts to reconcile different concepts and semantics across databases.

4. The challenge of taxonomic integration

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

To gain a deeper understanding of the challenge, we used 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 species groups and realms because it covers multiple underlying sources with the aim of maximizing integration (see Online Supplemental Table S1 and Box 1 Methods). As a simple but straightforward measure of interoperability, we quantified the portion of names from widely used databases that could not be matched against MOL taxonomies. We selected databases covering broad use 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 Catalogue of Life (COL) and the National Center for Biotechnological Information (NCBI) that did not have a match with the MOL list of names. This is substantial considering that the butterfly MOL taxonomy is characterized by more than >113,000 synonyms and alternative names for >19,000 accepted butterfly names (Online Supplemental Table S1). 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 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, complicating future interoperability in taxonomic relationships of further utility and published data.

4.2. 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 [18,49,61]. 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 species 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, "Species concept in space and time" links names to records of the species 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, "Species concept in space and time" may already be implicit from other elements but require 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.

4.3. 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. "Synonyms lists" appear most challenging in butterflies and dragonflies, where comprehensive lists require compilation of many sources [30,62]. "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 "Species concept in space and time" were poorly available or accomplished and represent an avenue for improving taxonomic integration. However, some ongoing efforts in this area aim to improve the integration of the "Species concept in space and time" [63,64].

5. 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.

5.1. 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 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.

5.2. 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 identifying and expanding the elements of the GIST across databases will further facilitate the implementation of the FAIR principles (findable, accessible, interoperable, reusable) [65] by improving access to information about sources (e.g., "Names Sources & Timestamps", "Authorship Information" elements). Moving forward, we recommend documenting how available and integrated the GIST elements are across databases and communities by following the assessment criteria in Box 1 and 2. Even though databases may respond to distinct codes of nomenclature [19,20,66], 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].

5.3. Communicate your challenges and opportunities to ease innovation

Both producers and users of data need to be a part of the solutions for broader integration potential directions and innovations [61]. One important innovation will be the development of common identifiers for all names and sources. Considering the diversity and complexity in the landscape of taxonomic databases and contributors, it is essential to prioritize consistent engagement between all actors on the integration process and data governance, and to share responsibility and create synergies. 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 network will need collective coalition-building 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 fragmented data and social infrastructures, mechanisms for tracking progress towards integration have the potential to help the community in building and using taxonomic data.

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 and inputs on the manuscript.

DECLARATION OF INTERESTS

No interests are declared.

HIGHLIGHTS

- Taxonomic knowledge is critical to understand, catalog, and assess biodiversity and central to measuring and achieving conservation goals, including the Post-2020 Framework of the Convention on Biological Diversity.
- Taxonomy is a centuries-old discipline, and 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 (GIST) 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

- While we are proposing to assess taxonomic integration across data sources and taxonomic groups, the continued addition of more biodiversity data is hard to manage, especially with the addition of DNA barcoding. How will taxonomic integration evolve and be sustained with a continued explosion of biodiversity data and names in the coming decades?
- Interoperability and key elements integration can be improved with the development of unique identifiers. How will persistent 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 species concept in space and time. While there are important advances on this subject for some taxonomic groups (e.g., mammals), will it be scalable to all other species groups and databases?
- Will the current community organizations in place, 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

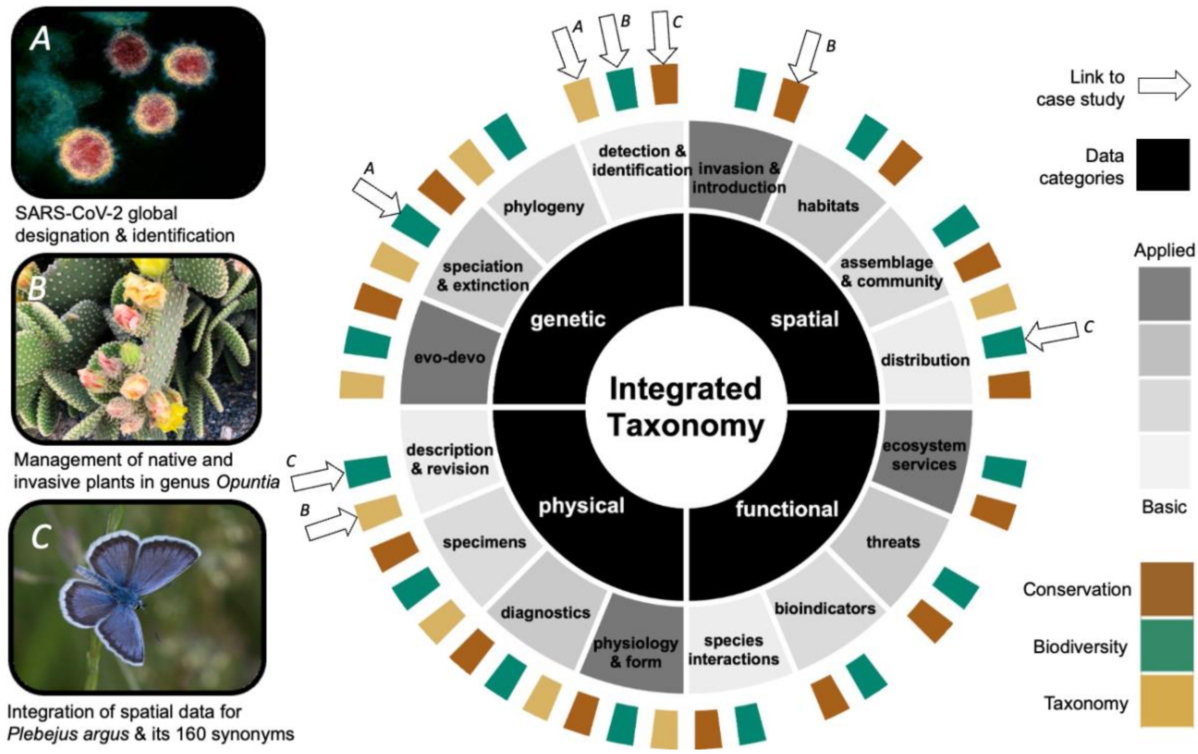


Figure 1. Research themes and examples with associated data types relying on taxonomic integration. Innermost ring (black): main data categories. Center ring (gray gradient): data applications (foundational to applied) across the four data categories. Outer ring (color categories): example research questions and applications (from taxonomy, biodiversity, and conservation). Arrows on the outermost edge of the rings denotes 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 global designation and identification. (B) Management of invasive plants in the genus *Opuntia*. (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 *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 synonyms lumped 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	Ref
Global authoritative list	*All accepted names for a taxon.* List of accepted names for a taxon defined by all species contained in a taxonomic rank (family, order, class, kingdom). This can be a preferred taxonomic authority or a compilation of accepted names in the absence of an authority.	Accepted name *Bison bison* from the list Mammal Species of the World (MSW, v.3)	\`taxonomicStatus\`: accepted; \`taxonRank\`: species (or below). Binomial name described by a minimum of \`genus\` and \`specificEpithet\`	[69]
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	\`taxonomicStatus\`: other than accepted; \`taxonRank\`: species (or below). Binomial name described by a minimum of \`genus\` and \`specificEpithet\`	[70]
Authorship information	*Author and publication years 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)	\`scientificNameAuthorship\`; \`namePublishedInYear\`	[69]
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	\`references\`; \`datasetID\`;\`datasetName\`; \`modified\` (but need to customize for the date when the data were retrieved)	[69]
Species concept in space and time	*Names associated with dates and locations, species 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 points in GBIF	Location class terms https://dwc.tdwg.org/terms/#location \`year\`	[71–73]

Name instance	*Instance of usage of a taxon name.* Physical observations of organisms and their relationships with names.	*Bison bison*, observed in Idaho, US 11th July 2017	`collectionID`;`institutionID`;`datasetID`;`catalogNumber`;`occurrenceID`	[74]

BOXES

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

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

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.

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

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 (Compositae).

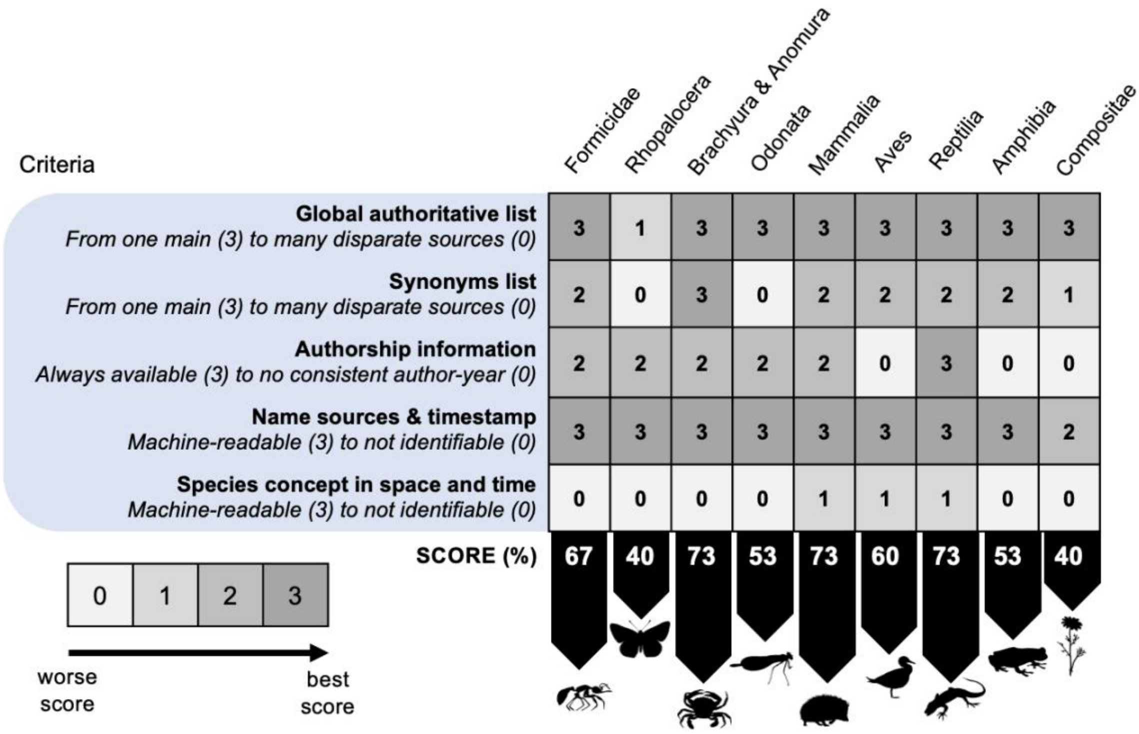
We quantified how well names from key databases match these taxonomies, either directly or via synonyms or orthographic variants. Key databases cover: Catalogue of Life COL [47], Open Tree of Life [75], National Center for Biotechnology Information NCBI [76], Global Biodiversity Information Facility GBIF [48], Global Register for Introduced and Invasive Species GRIIS [77], and International Union for the Conservation of Nature IUCN Red List of Threatened Species [78]. Lists of names 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 from the data source absent from MOL taxonomies (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. Taxa are then attributed a percentage based on the sum score relative to potential maximum score across all grades.

Data download, processing steps, and the code to conduct the name match and grading are open access (https://github.com/AquaAuma/review_gist).

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.



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 "Species 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 (family, order, class, kingdom or alike). 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.

Scientific nomenclature. Recognized scientific names 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. 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. Biological or phylogenetic species concepts are examples.

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

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

Taxonomic integration. Integrative science focused on new inferences 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, McGeoch MA, Guralnick R, Ferrier S, Beck J, Costello MJ, et al. Essential biodiversity variables for mapping and monitoring species populations. Nat Ecol Evol. 2019 Apr;3(4):539–51.
2. Sigwart JD, Bennett KD, Edie SM, Mander L, Okamura B, Padian K, et al. Measuring Biodiversity and Extinction—Present and Past. Integr Comp Biol. 2018 Dec 1;58(6):1111–7.
3. Abrahamse T, Andrade-Correa MG, Arida C, Galsim R, Hauser C, Price M, et al. The Global Taxonomy Initiative in Support of the Post-2020 Global Biodiversity Framework. Montreal: Secretariat of the Convention of Biological Diversity; 2021 p. 103 pages. Report No.: No. 96.
4. Deans AR, Yoder MJ, Balhoff JP. Time to change how we describe biodiversity. Trends Ecol Evol. 2012 Feb 1;27(2):78–84.
5. Costello MJ, Vanhoorne B, Appeltans W. Conservation of biodiversity through taxonomy, data publication, and collaborative infrastructures. Conserv Biol. 2015;29(4):1094–9.
6. Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. Data integration enables global biodiversity synthesis. Proc Natl Acad Sci [Internet]. 2021 Feb 9 [cited 2021 Feb 9];118(6). Available from: https://www.pnas.org/content/118/6/e2018093118
7. Tancoigne E, Ollivier G. Evaluating the progress and needs of taxonomy since the Convention on Biological Diversity: going beyond the rate of species description. Aust Syst Bot. 2017;30(4):326.
8. van Klink R, August T, Bas Y, Bodesheim P, Bonn A, Fossøy F, et al. Emerging technologies revolutionise insect ecology and monitoring. Trends Ecol Evol. 2022 Oct 1;37(10):872–85.
9. Hebert PDN, Gregory TR. The Promise of DNA Barcoding for Taxonomy. Syst Biol. 2005 Oct 1;54(5):852–9.
10. Wüster W, Thomson SA, O'shea M, Kaiser H. Confronting taxonomic vandalism in biology: conscientious community self-organization can preserve nomenclatural stability. Biol J Linn Soc. 2021 Jul 1;133(3):645–70.
11. Hey J, Waples RS, Arnold ML, Butlin RK, Harrison RG. Understanding and confronting species uncertainty in biology and conservation. Trends Ecol Evol. 2003 Nov 1;18(11):597–603.
12. Stropp J, Ladle RJ, Emilio T, Lessa T, Hortal J. Taxonomic uncertainty and the challenge of estimating global species richness. J Biogeogr. 2022;49(9):1654–6.
13. Isaac NJB, Mallet J, Mace GM. Taxonomic inflation: its influence on macroecology and conservation. Trends Ecol Evol. 2004 Sep;19(9):464–9.
14. Vogel Ely C, Bordignon SA de L, Trevisan R, Boldrini II. Implications of poor taxonomy in conservation. J Nat Conserv. 2017 Apr 1;36:10–3.
15. Correia RA, Jarić I, Jepson P, Malhado ACM, Alves JA, Ladle RJ. Nomenclature instability in species culturomic assessments: Why synonyms matter. Ecol Indic. 2018 Jul 1;90:74–8.
16. Montoya RD. Power of position: classification and the biodiversity sciences. Cambridge, Massachusetts: The MIT Press; 2022. 255 p. (History and foundations of information science).
17. Winston JE. Twenty-First Century Biological Nomenclature—The Enduring Power of Names. Integr Comp Biol. 2018 Dec 1;58(6):1122–31.
18. Franz NM, Peet RK. Perspectives: Towards a language for mapping relationships among taxonomic concepts. Syst Biodivers. 2009 Mar 1;7(1):5–20.
19. International Commission on Zoological Nomenclature, Ride WDL, International Trust for Zoological Nomenclature, Natural History Museum (London, England), International Union of Biological Sciences, editors. International code of zoological nomenclature: Code international de nomenclature zoologique. 4th ed. London: International Trust for Zoological Nomenclature, c/o Natural History Museum; 1999. 306 p.
20. Turland N, Wiersema J, Barrie F, Greuter W, Hawksworth D, Herendeen P, et al., editors. International Code of Nomenclature for algae, fungi, and plants [Internet]. Koeltz Botanical Books; 2018 [cited 2022 Oct 20]. (Regnum Vegetabile; vol. 159). Available from: https://www.iapt-taxon.org/nomen/main.php
21. Thomson SA, Pyle RL, Ahyong ST, Alonso-Zarazaga M, Ammirati J, Araya JF, et al. Taxonomy based on science is necessary for global conservation. PLOS Biol. 2018 Mar 14;16(3):e2005075.

22. Agapow P, Bininda-Emonds ORP, Crandall KA, Gittleman JL, Mace GM, Marshall JC, et al. The Impact of Species Concept on Biodiversity Studies. Q Rev Biol. 2004 Jun;79(2):161–79.
23. König C, Weigelt P, Schrader J, Taylor A, Kattge J, Kreft H. Biodiversity data integration—the significance of data resolution and domain. Mace GM, editor. PLOS Biol. 2019 Mar 18;17(3):e3000183.
24. Bisby FA. The Quiet Revolution: Biodiversity Informatics and the Internet. Science. 2000 Sep 29;289(5488):2309–12.
25. Feng X, Enquist BJ, Park DS, Boyle B, Breshears DD, Gallagher RV, et al. A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. Glob Ecol Biogeogr. 2022;31(7):1242–60.
26. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020 Apr;5(4):536–44.
27. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci. 2020 Apr 28;117(17):9241–3.
28. Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol. 2020;92(6):667–74.
29. Smith G, Figueiredo E, Crouch N, Boatwright J. South Africa's ongoing Opuntia Mill. (Cactaceae) problem: the case of O. microdasys (Lehm.) Pfeiff. Bradleya. 2011 Jan 1;29.
30. Pinkert S, Barve V, Guralnick R, Jetz W. Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. Glob Ecol Biogeogr. 2022;31(5):830–9.
31. Macgregor CJ, Thomas CD, Roy DB, Beaumont MA, Bell JR, Brereton T, et al. Climate-induced phenology shifts linked to range expansions in species with multiple reproductive cycles per year. Nat Commun. 2019 Oct 24;10(1):4455.
32. Edwards JL, Lane MA, Nielsen ES. Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science. 2000 Sep 29;289(5488):2312–4.
33. Di Cecco GJ, Barve V, Belitz MW, Stucky BJ, Guralnick RP, Hurlbert AH. Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. BioScience [Internet]. 2021 Sep 8 [cited 2021 Sep 13];(biab093). Available from: https://doi.org/10.1093/biosci/biab093
34. Franz NM, Pier NM, Reeder DM, Chen M, Yu S, Kianmajd P, et al. Two Influential Primate Classifications Logically Aligned. Syst Biol. 2016 Jul 1;65(4):561–82.
35. Sterner BW, Gilbert EE, Franz NM. Decentralized but Globally Coordinated Biodiversity Data. Front Big Data. 2020 Oct 23;3:519133.
36. Page RDM. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Brief Bioinform. 2008 Sep;9(5):345–54.
37. Kõljalg U, Nilsson HR, Schigel D, Tedersoo L, Larsson KH, May TW, et al. The Taxon Hypothesis Paradigm—On the Unambiguous Detection and Communication of Taxa. Microorganisms. 2020 Dec;8(12):1910.
38. Pyle RL. Towards a Global Names Architecture: The future of indexing scientific names. ZooKeys. 2016 Jan 7;(550):261–81.
39. Bénichou L, Gérard I, Laureys É, Price M. Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy. Eur J Taxon [Internet]. 2018 Nov 13 [cited 2021 Nov 24];(475). Available from: http://www.europeanjournaloftaxonomy.eu/index.php/ejt/article/view/606
40. McClure CJW, Lepage D, Dunn L, Anderson DL, Schulwitz SE, Camacho L, et al. Towards reconciliation of the four world bird lists: hotspots of disagreement in taxonomy of raptors. Proc R Soc B Biol Sci. 2020 Jun 24;287(1929):20200683.
41. Kindt R. WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. Appl Plant Sci. 2020;8(9):e11388.
42. Britz R, Hundsdörfer A, Fritz U. Funding, training, permits—the three big challenges of taxonomy. Megataxa [Internet]. 2020 Jan 31 [cited 2021 Nov 24];1(1). Available from: https://www.mapress.com/mt/article/view/megataxa.1.1.10
43. Conix S, Garnett ST, Thiele KR, Christidis L, van Dijk PP, Bánki OS, et al. Towards a global

list of accepted species III. Independence and stakeholder inclusion. Org Divers Evol [Internet]. 2021 Jul 13 [cited 2021 Jul 16]; Available from: https://doi.org/10.1007/s13127-021-00496-x

44. Lien AM, Conix S, Zachos FE, Christidis L, van Dijk PP, Bánki OS, et al. Towards a global list of accepted species IV: Overcoming fragmentation in the governance of taxonomic lists. Org Divers Evol [Internet]. 2021 Jul 23 [cited 2021 Jul 26]; Available from: https://doi.org/10.1007/s13127-021-00499-8
45. Uetz P, Koo MS, Aguilar R. A Quarter Century of Reptile and Amphibian Databases. 2021;10.
46. Frost DR. Amphibian Species of the World: an Online Reference. American Museum of Natural History; 1999.
47. Hobern D, Barik SK, Christidis L, T.Garnett S, Kirk P, Orrell TM, et al. Towards a global list of accepted species VI: The Catalogue of Life checklist. Org Divers Evol [Internet]. 2021 Oct 11 [cited 2021 Nov 15]; Available from: https://doi.org/10.1007/s13127-021-00516-w
48. Global Biodiversity Information Facility (GBIF) [Internet]. [cited 2022 Oct 5]. Available from: https://www.gbif.org/
49. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. Sarkar IN, editor. PLoS ONE. 2012 Jan 6;7(1):e29715.
50. Leonelli S. Classificatory Theory in Biology. Biol Theory. 2013 Jun 1;7(4):338–45.
51. Döring M, Jeppesen T, Bánki O. Introducing ChecklistBank: An index and repository for taxonomic data. Biodivers Inf Sci Stand. 2022 Aug 24;6:e93938.
52. Costello MJ, Bouchet P, Boxshall G, Fauchald K, Gordon D, Hoeksema BW, et al. Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. PLOS ONE. 2013 Jan 9;8(1):e51629.
53. Grenié M, Berti E, Carvajal-Quintero J, Dädlow GML, Sagouis A, Winter M. Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. Methods Ecol Evol. 2023;14(1):12–25.
54. Rolin K. Values in Science: The Case of Scientific Collaboration. Philos Sci. 2015 Apr;82(2):157–77.
55. Manolis JC, Chan KM, Finkelstein ME, Stephens S, Nelson CR, Grant JB, et al. Leadership: a New Frontier in Conservation Science. Conserv Biol. 2009;23(4):879–86.
56. Wheeler QD, Raven PH, Wilson EO. Taxonomy: Impediment or Expedient? Science. 2004 Jan 16;303(5656):285–285.
57. Kühl HS, Bowler DE, Bösch L, Bruelheide H, Dauber J, Eichenberg David, et al. Effective Biodiversity Monitoring Needs a Culture of Integration. One Earth. 2020 Oct;3(4):462–74.
58. Garnett ST, Christidis L. Science-based taxonomy still needs better governance: Response to Thomson et al. PLoS Biol. 2018 Mar 14;16(3):e2005249.
59. Raposo MA, Stopiglia R, Brito GRR, Bockmann FA, Kirwan GM, Gayon J, et al. What really hampers taxonomy and conservation? A riposte to Garnett and Christidis. Zootaxa. 2017 Sep 1;4317(1):179–84.
60. Jetz W, McPherson JM, Guralnick RP. Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol Evol. 2012 Mar 1;27(3):151–9.
61. Garnett ST, Christidis L, Conix S, Costello MJ, Zachos FE, Bánki OS, et al. Principles for creating a single authoritative list of the world's species. PLOS Biol. 2020 juil;18(7):e3000736.
62. Sandall EL, Pinkert S, Jetz W. Country-level checklists and occurrences for the world's Odonata (dragonflies and damselflies). J Biogeogr. 2022;49(8):1586–98.
63. Upham N, Powell C, Prado L, Franz N, Sterner B. Extended Taxonomic Curation: Moving beyond species lists to linking species data. Biodivers Inf Sci Stand. 2022 Aug 23;6:e93670.
64. Pyle R, Bailly N, Remsen D. Modeling Taxon Concepts: A new approach to an old problem. Biodivers Inf Sci Stand. 2022 Aug 24;6:e93927.
65. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018.
66. IPNI. International Plant Names Index. R Bot Gard Kew Harv Univ Herb Libr Aust Natl Herb Retrieved 05 April 2023 [Internet]. 2023 [cited 2023 Apr 27]; Available from: Published on the Internet https://www.ipni.org/

67. Wilson DE, Reeder DM. Mammal Species of the World: A Taxonomic and Geographic Reference. 3rd ed. Johns Hopkins University Press, Baltimore, Maryland; 2005. 2142 p.
68. Cronin MA, MacNeil MD, Vu N, Leesburg V, Blackburn HD, Derr JN. Genetic Variation and Differentiation of Bison (Bison bison) Subspecies and Cattle (Bos taurus) Breeds and Subspecies. J Hered. 2013 Jul 1;104(4):500–9.
69. Mammal Species of the World - Browse: bison [Internet]. [cited 2022 Oct 6]. Available from: https://www.departments.bucknell.edu/biology/resources/msw3/browse.asp?s=y&id=14200669
70. Bos bison Linnaeus, 1758 [Internet]. [cited 2023 Apr 22]. Available from: https://www.gbif.org/species/2441178
71. Bison bison bison [Internet]. [cited 2022 Oct 6]. Available from: https://www.gbif.org/species/7194051
72. Bison bison (Linnaeus, 1758) [Internet]. [cited 2022 Oct 6]. Available from: https://www.gbif.org/species/2441176
73. Bison bison athabascae Rhoads, 1898 [Internet]. [cited 2022 Oct 6]. Available from: https://www.gbif.org/species/9182932
74. Occurrence Detail 2832703394 [Internet]. [cited 2022 Oct 6]. Available from: https://www.gbif.org/occurrence/2832703394
75. Rees J, Cranston K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodivers Data J. 2017 May 22;5:e12581.
76. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database [Internet]. 2020 Jan 1 [cited 2021 Oct 4];2020(baaa062). Available from: https://doi.org/10.1093/database/baaa062
77. Pagad S, Genovesi P, Carnevali L, Schigel D, McGeoch MA. Introducing the Global Register of Introduced and Invasive Species. Sci Data. 2018 Jan 23;5(1):170202.
78. IUCN. The IUCN Red List of Threatened Species. Version 2022-2. IUCN Red List Threat Species [Internet]. [cited 2022 Oct 5]; Available from: https://www.iucnredlist.org/en