ECOGRAPHY

Research Article

Enabling ecological survey data integration with the Humboldt Extension to Darwin Core

Yanina V. Sica[1,2,3], Wesley M. Hochachka[4], Robert D. Stevenson[5], Kate Ingenloff[6], Paula F. Zermoglio[7,8], John Wieczorek[9], Yi Ming Gan[10], Dmitry Schigel[6], Zachary R. Kachian[11], Steve Baskauf[12], Peter Brenton[13], Anahita J. N. Kazem[14], Walter Jetz[1,2] and Robert Guralnick[15]

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA
[2]Center for Biodiversity and Global Change, Yale University, New Haven, CT, USA
[3]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt (Main), Germany
[4]Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA
[5]Department of Biology, University of Massachusetts Boston, Boston, MA, USA
[6]Global Biodiversity Information Facility, Secretariat, Copenhagen Ø, Denmark
[7]Universidad Nacional de Río Negro, Instituto de Investigaciones en Recursos Naturales, Agroecología y Desarrollo Rural, San Carlos de Bariloche, Río Negro, Argentina
[8]Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Investigaciones en Recursos Naturales, Agroecología y Desarrollo Rural, San Carlos de Bariloche, Río Negro, Argentina
[9]Rauthiflor LLC, Seattle, WA, USA
[10]Royal Belgian Institute of Natural Sciences, Brussels, Belgium
[11]Field Museum of Natural History, Keller Science Action Center, Chicago, IL, USA
[12]Vanderbilt University Heard Libraries, Nashville, TN, USA
[13]Atlas of Living Australia, CSIRO, Canberra, Australia
[14]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
[15]Florida Museum of Natural History, University of Florida, Gainesville, FL, USA

Correspondence: Yanina V. Sica (yanina.sica@senckenberg.de)

Ecography
2026: e08223
doi: 10.1002/ecog.08223

Subject Editor: Henrique Pereira
Editor-in-Chief: Miguel Araújo
Accepted 20 October 2025

NORDIC SOCIETY OIKOS

www.ecography.org

In the face of the global biodiversity crisis, accessibility to biodiversity data that are maximally effective for downstream use in science, conservation, and policy is paramount. The Darwin Core standard has played a central role in providing a standardised structure and vocabulary for biodiversity data. However, early iterations of the standard were not optimised to capture the sampling context of biodiversity surveys – survey methods, scope, and sampling effort – which is essential for the correct interpretation and potential reuse of such data. To address this limitation, we present the Humboldt Extension to Darwin Core, a ratified standard designed to accommodate datasets that contain such contextual information. Building upon an initial, previously developed framework, we significantly improved, fully tested, and ratified a final standard, following a community process defined by biodiversity information standards (TDWG), an international standards organisation. The resulting Humboldt Extension adds 55 terms that enrich the Darwin Core, providing the terms needed to capture and share multiple types of biodiversity survey data. We illustrate the benefits of implementing the Humboldt Extension with three case studies and demonstrate how richer data can be used in research, modelling, and to inform decision-making. We urge the uptake

and use of this Extension to facilitate the reuse and synthesis of monitoring data, particularly structured surveys and inventories, for science, conservation, and policy.

Keywords: biodiversity monitoring, data standards, ecological inventories, FAIR data, sampling design

Introduction

Biodiversity loss presents a critical threat to global health and ecosystems, exacerbated by unsustainable human activities that accelerate species extinction (Wilson 1988, Cardinale et al. 2012, IPBES 2019, WWF 2024). Despite significant advancements in our understanding of biodiversity and nature's essential contributions to human well-being (Costanza et al. 1998, Daily 1999), the urgency of the crisis demands immediate action. Policymakers and governments have responded with commitments such as the United Nations Convention on Biological Diversity (CBD) and the establishment of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) to navigate these challenges. In December 2022, the CBD adopted the historic Kunming–Montreal Global Biodiversity Framework (GBF) to halt and reverse biodiversity loss by 2050, incorporating a monitoring mechanism to track progress towards its ambitious targets and rectify shortcomings of previous efforts, such as the Aichi Biodiversity Targets (Tittensor et al. 2014, Buchanan et al. 2020). The effectiveness of this framework hinges upon the availability of robust biodiversity data necessary for measuring progress (Affinito et al. 2024). Consequently, the biodiversity monitoring community should enhance data collection and sharing following FAIR principles – findability, accessibility, interoperability, and reusability (Wilkinson et al. 2016) – while fostering infrastructures that facilitate access and sharing of in situ data and local datasets to meet the GBF's monitoring requirements (Bisby 2000, Heberling et al. 2021).

The use of standardised vocabularies and common data structures underlies the success of large biodiversity data infrastructures, such as the Global Biodiversity Information Facility (GBIF 2025a, https://www.gbif.org/) and the Ocean Biodiversity Information System (OBIS 2025, https://obis.org/). Established in 2001, GBIF provides access to over 113 000 datasets comprising more than three billion species occurrences documented by more than 2400 institutions (as of May 2025). OBIS, launched in 2004, is a global open-access data and information clearing house on marine biodiversity, hosting 5719 datasets with over 137.5 million occurrence records (as of May 2025).

At the heart of these biodiversity platforms are publishing systems that use the Darwin Core standard (DwC, Wieczorek et al. 2012), which is maintained by an international organisation, the Biodiversity Information Standards (TDWG, formerly the Taxonomic Databases Working Group). DwC was originally developed to facilitate the interchange of information about the occurrence of organisms and taxa, focusing on standardising data and metadata from natural history collections. It therefore has focused on 'species occurrences' and how, when, where, and by whom those occurrences were recorded. DwC is most suitable for incidentally recorded occurrences, lacking the granular reporting of the data collection process required for semi-structured and structured surveying. Still, it has been expanded and improved to include additional key concepts and data types, and is managed by an active maintenance group that continues to improve and extend the standard (Robertson et al. 2014). DwC has proven to be the lynchpin standard for sharing species occurrence data and promoting biodiversity research because it enables interoperability of data. To date, over 13 000 peer-reviewed publications have used GBIF-mediated data to advance global biodiversity research and management (GBIF 2025b).

The 'presence-only' data standardised and made interoperable through DwC, while undoubtedly of value (e.g. for species distribution modelling), also possess significant limitations (Jetz et al. 2012, Guillera-Arroita et al. 2015). Knowing where species were detected but also where they were not detected (i.e. reported counts of zero) enables more accurate and precise estimates of realised species distributions (Elith et al. 2006, Johnston et al. 2021). Similarly, accounting for differences in the sampling or observation process leading to records enables better estimates of distribution, relative abundance (Johnston et al. 2021), and trends in abundance over time (Fink et al. 2023). In combination, records of zero detections and information about the sampling process can be used to estimate the proportion of zero counts that represent true absences, and not merely the failure to detect a taxon that was present but undetected. Some analytical methods (MacKenzie et al. 2002, Royle 2004) additionally require repeated surveys of target taxa at each of a set of locations, such that data are structured hierarchically (e.g. with individual sampling events nested within survey sites). While some of these types of information can be reported using DwC, it may not fully capture aspects such as the absence of detections, taxonomic scope, variations in the data collection process (e.g. sampling effort), or hierarchical structure. Such information may exist, but it is typically shared as part of the metadata in an unstructured manner, limiting the survey's proper interpretation and potential re-use (Guralnick et al. 2018). The lack of a standardised vocabulary to describe biodiversity surveys is a persistent barrier to their integration and broadest reuse. This limitation, for instance, constitutes a key challenge for deriving some of the essential biodiversity variables (EBV) useful for monitoring biodiversity with high accuracy (Pereira et al. 2013, Jetz et al. 2019). In particular, understanding trends in occupancy and abundance over time is critical for assessing species and population health

(Jetz et al. 2019), but in most cases requires semi-structured or structured monitoring and its associated metadata (Kelling et al. 2019).

To fill the need for a richer vocabulary by which to capture and share biodiversity survey data and metadata, Guralnick et al. (2018) proposed a new standard: the Humboldt Core. This proposal was an initial step towards standardising the capture of information about the data collection process and aimed to lay the foundations for a broader effort. Here, we describe the outcome of a subsequent community-driven process to develop a standardised vocabulary for capturing the sampling context of biodiversity survey data, specifically focusing on enabling a proper description of the data collection process such that the data are interpretable and interoperable. This formal process was accomplished by a Task Group within TDWG, following a robust procedure (Vocabulary Maintenance Specification Task Group 2017). We describe how and why the Humboldt Core evolved into the Humboldt Extension for Ecological Inventories (https://eco.tdwg.org/), and we demonstrate how its implementation supports biodiversity data sharing. A key outcome described here is a comprehensive set of terms that facilitate explicit capture of the information needed to understand a survey design (i.e. information related to survey scope, effort, and completeness) to enable proper downstream integration and analysis. The results also contain three implementation case studies, and we discuss the implications of adopting the Humboldt Extension for data collection, curation, analyses, and further applications in policy and conservation.

Material and methods

Biodiversity surveys

Biodiversity surveys and inventories (we will use these terms interchangeably) are systematic processes for identifying and cataloguing the taxa present within a specific area. Because of their systematic structure and design, surveys can effectively reveal species coexistence and absence patterns if the sampling design and effort are appropriate relative to the intended taxonomic and spatiotemporal scopes (Guralnick et al. 2018). These key descriptors of the data collection process have to be reported in a standardised and structured way if users are to interpret and reuse survey data accurately.

Guralnick et al. (2018) identified three types of survey data, and they contrasted these with 'incidental data' – obtained via opportunistic sampling, lacking structured design and taxonomic scope, with no information on non-detections (e.g. museum specimens, incidental citizen science data). An 'elementary inventory' refers to a single sampling event with a limited spatiotemporal scope (e.g. surveys composed of a single net sweep, trap, transect, or point count). An 'extended inventory' involves repeated sampling events that aggregate data over broader temporal or spatial scales (e.g. multi-point or transect sampling surveys, long-term vegetation surveys, algal sampling in water monitoring). Both survey types provide explicit details on sampling protocols, spatial boundaries, and expectations for targeted organisms. At a higher aggregation level, a 'compilation' merges studies employing different data collection processes (such as sampling events combined with literature searches), often involving larger spatial or temporal scopes (e.g. taxon checklists in protected areas, atlas grid cell summary data). The more complex the sampling design is (from elementary to extended and compilation inventories), the more information about the data collection process needs to be captured to maximise the reusability of these data (Fig. 1). Creating a standardised way of recording these complex data was the overarching goal for creating the Humboldt Extension.

Augmenting the DwC standard with the Humboldt Extension

Moving from the proposed Humboldt Core (Guralnick et al. 2018) to the ratified Humboldt Extension to Darwin Core occurred in five distinct phases during 2021–2024: 1) formation of a TDWG task group to review and progress the Humboldt Core, 2) determination of whether 'Humboldt Core' should be presented as a new standard (e.g. a 'core') or as an extension to DwC, 3) review of the Humboldt vocabulary following TDWG standards, 4) implementation and evaluation of the proposed vocabulary with real-world data, and 5) ratification following the TDWG process (Vocabulary Maintenance Specification Task Group 2017). See the Supporting information for details about the process.

Development of the Humboldt Extension began in 2021 under the auspices of TDWG, through the Humboldt Task Group (https://www.tdwg.org/community/osr/humboldt-extension/). The Task Group determined that the scope and terms of the proposed Humboldt Core were best suited as an extension to DwC (Supporting information). Hence, we reviewed the 94 terms initially proposed in the Humboldt Core (Supporting information in Guralnick et al. 2018), reformulating term definitions, comments, and examples to follow TDWG standards and current DwC patterns. Following Guralnick et al. (2018) we identified the key descriptors of the data collection process and defined six classes of information that need to be covered by at least one term to ensure the interpretability of a survey: spatial scope, temporal scope, taxonomic and organismal scope, sampling design, sampling protocols, and sampling effort. We defined new terms where gaps in coverage were identified and removed other terms to reduce redundancy and to avoid the use of terms describing summaries and/or interpretations of the data collected (Sica et al. 2022). We presented an initial set of 45 terms for implementation and evaluation with real-world datasets. Testing these datasets provided explicit feedback about vocabulary completeness and clarity, and resulted in enlarging the vocabulary to 54 terms (Hochachka et al. 2023). Three datasets were selected as case studies to illustrate the broad applicability of the Humboldt Extension (Supporting information).

To pursue ratification of the Humboldt Extension (i.e. formalisation of the extension through the TDWG process), we submitted the vocabulary for public review, in response

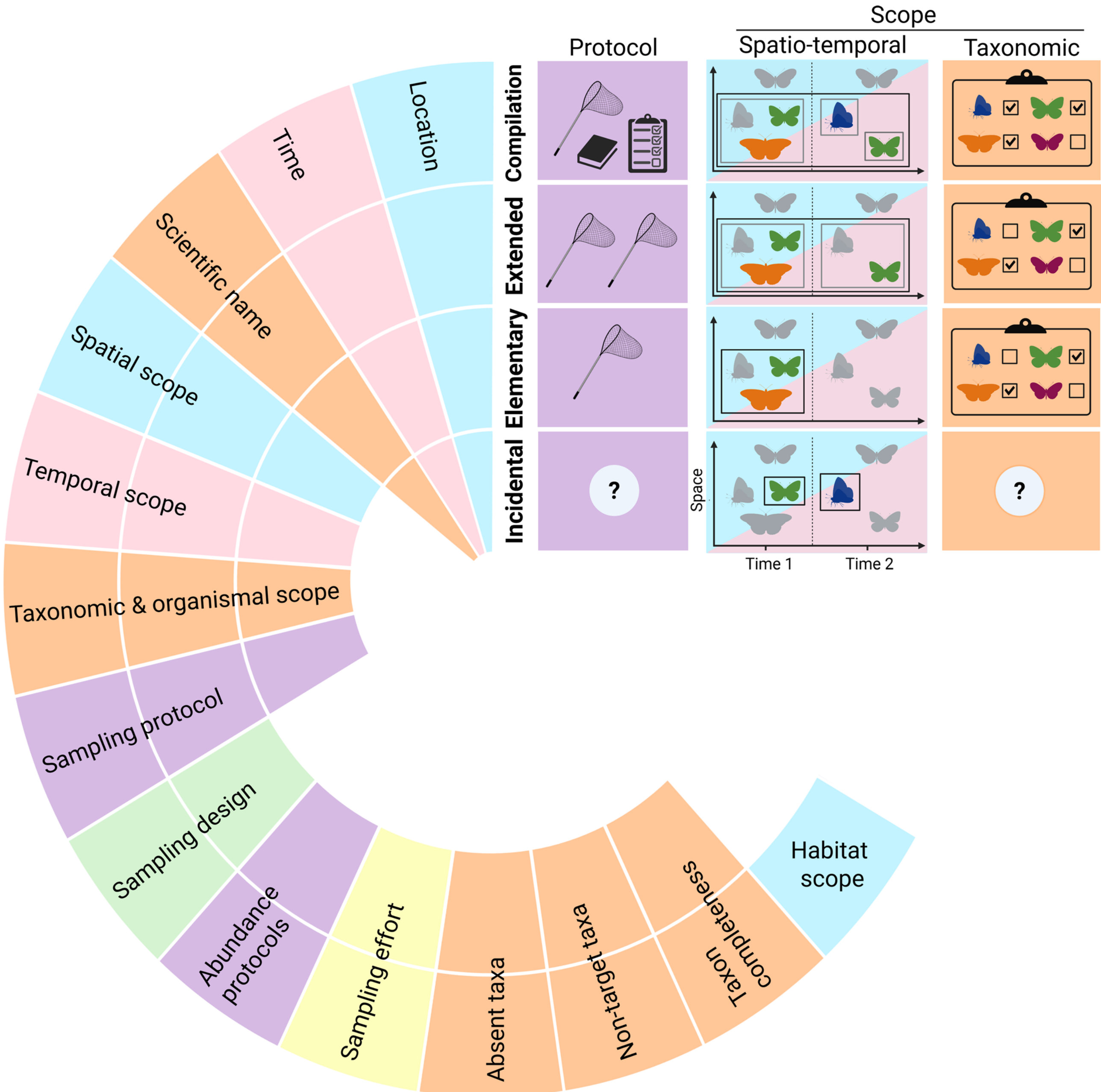

Figure 1. Overview of the four main survey types, illustrating the types of information that can now be reported using the Humboldt Extension: location and time of recorded taxa, spatiotemporal scope, expected taxa, protocol used, sampling design, counts of organisms and information about abundance, sampling effort, absent taxa, taxa out of scope, taxonomic completeness, and habitat scope. Colours represent related aspects of survey data (information on location in blue, time and duration in pink, taxa in orange, protocols in purple, sampling design in green, and effort in yellow). The 'Incidental records' example here describes presence-only data lacking contextual information (data on two incidental observations in the same area at different times). The other types of surveys represented here have information on the protocols used and possess a defined taxonomic, geospatial, and temporal scope. Our 'Elementary survey' example addresses a single sampling event, while the 'Extended survey' example combines two events. Our 'Compilation' example involves multiple data collection processes (multiple sampling events combined with a literature search and checklist data). As the amount of information in a dataset increases, the potential uses of such datasets and the inferences that can be made (e.g. inference about absences, taxon completeness, and abundance trends) also increase. Within the spatio-temporal scope, box outlines indicate the space and time effectively sampled, and grey icons denote species that were present but not detected. Question marks indicate lack of information. Created with BioRender.com.

to which several modifications were made. After a 6-month community review, the Humboldt Extension for Ecological Inventories (https://eco.tdwg.org/), encompassing a final list of 55 terms, was officially ratified in February 2024. The extension is actively maintained by the Darwin Core Maintenance Group (https://www.tdwg.org/community/dwc/), and ongoing feedback can be provided at the extension's GitHub site (https://github.com/tdwg/hc).

Results

Humboldt Extension for ecological Inventories

The Humboldt Extension for Ecological Inventories allows the reporting of six different aspects of the data collection process in a structured manner (Fig. 2). The 55 Humboldt Extension terms can be used alongside the existing DwC terms, including those from other standards and extensions (e.g. extended measurement or fact: see https://rs.gbif.org/extension/obis/extended_measurement_or_fact_2023-08-28.xml, Audiovisual Core: Audiovisual Core Maintenance Group 2023, Camtrap DP: Bubnicki et al. 2024). Twenty-three terms have equivalents that allow linking to an externally described object via an internationalised resource identifier (IRI). To avoid information loss or misinterpretation, the details of the data collection process have to be reported at all relevant levels of the hierarchical sampling design (if any). Hence, many terms can be documented at any hierarchical level. The Humboldt Extension includes guidelines on how to structure and utilise hierarchies to describe surveys (TDWG Humboldt Extension Task Group 2024, Supporting information).

Below, we describe how the Humboldt Extension enables reporting of the data collection process. Some Humboldt Extension terms partially overlap with DwC Event terms (e.g. dwc:samplingEffort) but allow more granularity in the reporting. The more specific Humboldt terms (e.g. *samplingEffortValue*, *samplingEffortUnit*, *samplingEffortProtocol*) should be used where there is such an overlap. The term descriptions presented in this paper should be complemented by the definitions and specifications of the terms included in the Humboldt Extension Quick Reference Guide (https://eco.tdwg.org/terms/). Although the Quick Reference Guide divides the vocabulary into eight groups, we have summarised the terms under six categories, which more naturally follow the logic of someone collecting data. Note that while we have organised our presentation of the Humboldt Extension's terms into categories, these categories do not represent formal DwC classes. Humboldt Extension term names below are written in italics.

Capturing the sampling design

DwC enables reporting of hierarchically nested survey designs, represented as a relational schema, characterised by parent–child relationships (e.g. a parent sampling event – plot – linked to one or more child sub-sampling events – traps) using event and parent event identifiers. Two Humboldt Extension terms provide a means by which to share additional context to describe this structure (*siteNestingDescription*; e.g. traps nested within transects nested within plots) and report the number of sites surveyed *siteCount*. These terms can be documented at any hierarchical level.

Two additional terms capture information about the survey sites, such as the site names *verbatimSiteNames* and descriptions *verbatimSiteDescriptions*.

Spatial scope

Several DwC terms already capture the location where observations are made or samples/measurements are taken (e.g. decimalLatitude, decimalLongitude, coordinateUncertaintyInMeters). The Humboldt Extension adds to these DwC terms by including two sets of paired terms that capture the geospatial scope of the survey, i.e. the area expected to be sampled (*geospatialAreaScopeValue* and *geospatialAreaScopeUnit*), and the realised survey area, i.e. the area effectively sampled (*totalAreaSampledValue* and *totalAreaSampledUnit*). An additional pair of terms can be used to report habitat types specifically targeted or excluded from sampling (*targetHabitatScope*, *excludedHabitatScope*).

Temporal scope

Several DwC terms already capture the time when observations are made or samples/measurements are taken (e.g. eventDate, eventTime). The Humboldt Extension adds a pair of terms (*eventDurationValue*, *eventDurationUnit*) that report the duration of a sampling event at any level of the hierarchy where the duration is needed.

Taxonomic and organismal scopes

Information on the taxonomic group(s) targeted for and excluded from sampling can now be captured using two terms (*targetTaxonomicScope*, *excludedTaxonomicScope*). Whether every taxon included in the taxonomic scope that was observed or sampled is also reported (i.e. completeness of reporting) can be declared using the term *isTaxonomicScopeFullyReported*.

Taxa deemed absent during a survey can be reported at any taxonomic level using two terms: *isAbsenceReported*, *absentTaxa*. Absence refers to the absence of reporting (i.e. taxa not reported), which may be the result of three separate processes: 1) true absence, definitive absence of a taxon in the specified scope at the site at the time of surveying, 2) non-detection, where the taxon was not detected by the sampling gear and effort at the site at the time of surveying, or 3) incomplete reporting, lack of reporting of a taxon that was present and detected, either because that taxon was outside the intended scope of the survey or simply because its presence was not recorded or reported. True absence of a taxon is hard to prove with complete certainty, but a combination of standardised information about taxonomic scope, sampling effort, and reporting completeness (that is now enabled by multiple terms in the Extension) can be used to make inferences about true absence. If the absent taxon is within the taxonomic scope of the survey, the sampling protocol and

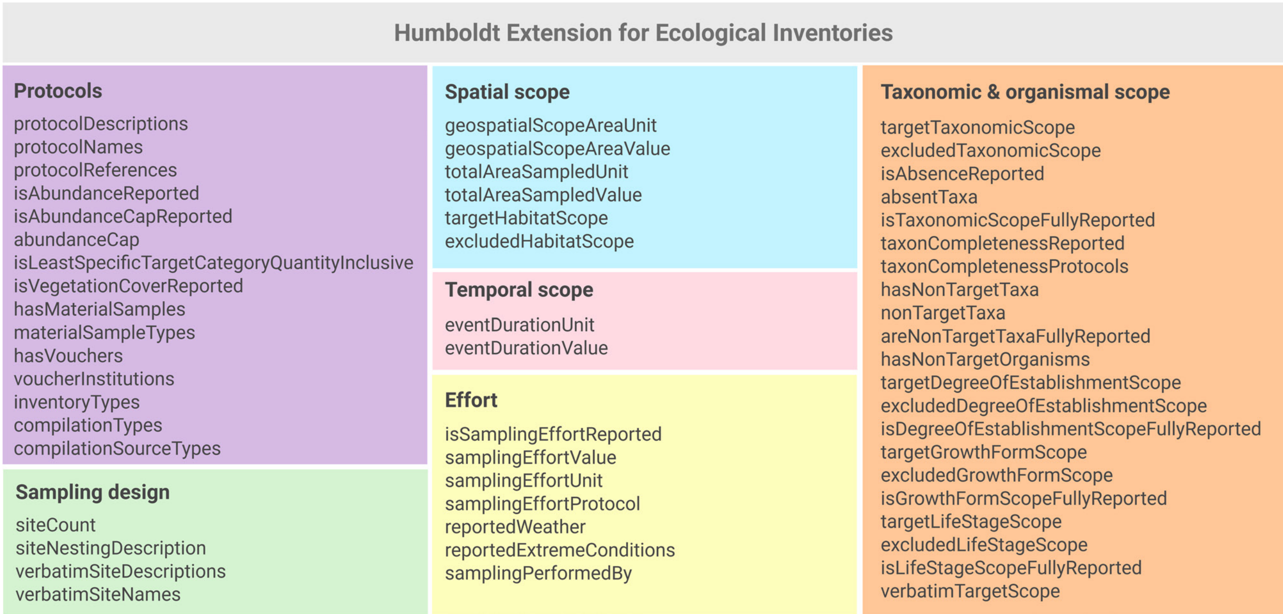

Figure 2. Fifty-five Humboldt Extension terms organised in six non-formal categories that capture different aspects of the data collection process. Created with BioRender.com.

effort are properly designed to sample the taxon, and the reporting of taxa is complete within that scope, then formal statistical methods exist for estimating the probability that a taxon is truly absent (MacKenzie et al. 2002).

If taxonomic completeness (i.e. how comprehensively an area has been surveyed) was assessed, both the results and the protocols employed to determine this can also be documented using the terms *taxonCompletenessReported* and *taxonCompletenessProtocols*.

Four terms are available to report the existence of non-target taxa and/or organisms (a.k.a. bycatch) in a dataset: *hasNonTargetTaxa*, *nonTargetTaxa*, *areNonTargetTaxaFullyReported*, and *hasNonTargetOrganisms*.

As some taxa undergo dramatic morphological and physiological transformations during their lifetime (e.g. corals, butterflies and moths, amphibians), nine terms are available to report targeted (or excluded) organismal life stages, as well as growth forms and/or degrees of establishment: *targetLifeStageScope*, *excludedLifeStageScope*, *isLifeStageScopeFullyReported*, *targetDegreeOfEstablishmentScope*, *excludedDegreeOfEstablishmentScope*, *isDegreeOfEstablishmentScopeFullyReported*, *targetGrowthFormScope*, *excludedGrowthFormScope*, *isGrowthFormScopeFullyReported*. Other target traits not captured by these three categories (e.g. freshwater macroinvertebrates or small mammals) can be captured in *verbatimTargetScope*.

Sampling protocols and survey type

Protocols and specific methodologies for sampling or collecting measurements during a survey can be reported using three terms: *protocolNames*, *protocolDescriptions*, and *protocolReferences*.

A series of terms are available to describe if abundance information was collected and reported (*isAbundanceReported*), if there was any cap on abundance value (*isAbundanceCapReported*, *abundanceCap*), how organism counts were reported when a taxon could be reported at more than one taxonomic level such as species and subspecies using *isLeastSpecificTargetCategoryQuantityInclusive* (see specific guidelines on usage of this term here: https://eco.tdwg.org/inclusive/), and if vegetation cover was reported (*isVegetationCoverReported*).

Four terms are available to report if specimen vouchers or material samples were collected (*hasMaterialSamples* and *hasVouchers*), what those samples were (*materialSampleTypes*), and where any vouchers are housed (*voucherInstitutions*).

Finally, three terms expand on the DwC Event term eventType. If the data reported are a survey or inventory, the search or data collection process employed can be documented in *inventoryTypes*. If the survey is a compilation of multiple datasets, detailed information can be captured on whether the data originate from sampling events, ancillary data compiled from other sources, or a combination of both, using the terms *compilationTypes* and *compilationSourceTypes*.

Sampling effort

Information about the sampling effort (i.e. the intensity of the sampling process) can be captured with five terms at any hierarchical level. Whether or not sampling effort is reported, and what that effort was, can be reported using *isSamplingEffortReported* and the paired terms *samplingEffortValue* and *samplingEffortUnit*. Specific details about the methods implemented can be reported in *samplingEffortProtocol*. Importantly, the people and/or organisations responsible for conducting the survey can now be captured using *samplingPerformedBy*, both as free text and, ideally, by including permanent identifiers for these entities (e.g. Open Researcher and Contributor ID – ORCID, Research Organization Registry – ROR). Since weather and other conditions can affect the probability of detecting and identifying an organism, which in turn can impact the resulting effort, two terms help describe weather conditions during the survey and extreme weather events (*reportedWeather*, *reportedExtremeConditions*).

Case studies

We illustrate the application of the Humboldt Extension using three case studies (Table 1, Supporting information). The first case study describes Antarctic marine inventory data (Van De Putte et al. 2010), which includes multiple target and organismal scopes and exemplifies the use of multiple extensions to properly describe such a complex dataset. The second case study illustrates the importance of describing variation in sampling methods among observers taking part in a global citizen/participatory science project (Sullivan et al. 2009, 2014). The third example highlights the ability of the Humboldt Extension to describe complex, hierarchical sampling designs in data collected using a diversity of planned methods to survey different taxa (https://www.rapidinventories.fieldmuseum.org). These case studies demonstrate the flexibility and comprehensiveness of the Humboldt Extension in describing real-world datasets, highlighting how survey information beyond the species occurrence can be captured in a structured way (Table 1).

Discussion

The Humboldt Extension was developed to capture and represent a broad array of biodiversity data collected through standardised, structured, or semi-structured methods, including museum expeditions, biodiversity surveys using direct observations, semi-structured monitoring such as eBird checklists, eDNA, camera traps, acoustic monitoring, or any other ecological monitoring initiatives. It also accommodates data collected in an unstructured way from citizen science efforts, bioblitzes, or other opportunistic sampling. A large amount of biodiversity survey data is produced by ecological research, but its availability is limited as it often depends on the motivations of the data collector and the constraints on reporting and sharing (Bowler et al. 2025). While some of these data are shared and compiled in biodiversity databases, such as the Living Planet Database (WWF 2024), Predicts (Hudson et al. 2014), and BioTime (Dornelas et al. 2018), the vast majority

Table 1. The Humboldt Extension enables researchers to store multiple types of information about surveys as data that can be searched and sorted. Previously, this contextual information was only contained in metadata documentation. In this table, we illustrate some important types of information that can now be reported making use of Humboldt Extension terms. Examples of data sources for which Humboldt Extension terms describe important facets of data are described as case studies in the Supporting information.

Type of information	Reason for the addition	Case studies
What taxa or organisms were the targets of the survey? Were there any taxa or organisms intentionally excluded from the scope of the survey?	The types of organisms that are in the scope of the survey can be described with Humboldt Extension terms not only taxonomically (e.g. which species?) but also based on other criteria such as life stage (e.g. larva or adult) or other defined groupings (e.g. by size or ecological functionality). It can also be important to report which types of organisms are excluded from the scope of the survey (e.g. because of limitations in the sampling instrument used).	Case study 1: OBIS
Were any non-target organisms detected and reported?	While a survey will typically be intended to detect and record specific types of organisms, potentially non-target organisms will also be detected. If desired, these non-target organisms or bycatch can be reported using the Humboldt Extension.	Case study 1: OBIS Case study 3: Field Museum rapid inventories
Which taxa were the targets of the survey, but were not detected during a survey event?	Data about the failure to detect a taxon, when that taxon was in the scope of the survey, is important for many types of analytical methods. With the Humboldt Extension, records of zero individuals do not need to be explicitly entered as such but can be indicated implicitly when the taxonomic scope is reported and when all detected taxa are indicated as being reported (i.e. completeness of reporting).	Case study 1: OBIS Case study 2: eBird
Within one sampling event, were there differences in survey protocol among records?	When sampling protocols are not tightly constrained (e.g. for data collected in citizen science surveys) some aspects of the data collection process can differ among sampling events in a survey (e.g. sampling effort). Such differences can be reported using the Humboldt Extension. Accounting for these variations improves the precision and accuracy of model estimates.	Case study 2: eBird Case study 3: Field Museum rapid inventories
Are multiple surveys, each focused on a different target organism type and with different sampling methodology, related to each other (e.g. same sampling locations, same sampling dates)?	Community and ecosystem surveys may need to employ different sampling protocols for different types of organisms or may require a structured spatial or temporal design. The Humboldt Extension allows information from different sampling events to be linked with each other in a hierarchical manner (e.g. different survey protocols conducted at the same location at the same time or a spatially nested sampling design).	Case study 3: Field Museum rapid inventories

are stored in repositories as ancillary data to scientific articles where they are typically discoverable through dataset catalogues and supplementary materials to publications, limiting their discoverability and reuse (Huang et al. 2012). Many data are not shared at all, only existing in the possession of the data collectors (Huang et al. 2012), which increases the risk of loss due to local storage failures or ageing technology (Gonzalez and Peres-Neto 2015). As policymakers increasingly rely on biodiversity data to inform decision-making and conservation, the adoption of standardised vocabularies that enhance data interpretability and reuse becomes imperative. Features of the Humboldt Extension can encourage biodiversity data reuse by: 1) enabling easier discovery and interpretability of data, 2) facilitating integration of data from multiple sources, and 3) incentivising data sharing.

Enabling biodiversity data discovery and interpretability

The terms in the Humboldt Extension provide a standardised way for researchers to describe their data and metadata, so that people have a common understanding of the dataset (e.g. what was intended to be sampled, how it was collected, and by whom) while being in a machine-readable form. Four important types of information about surveys can now be more properly structured and reported for each dataset: 1) hierarchical events and relationships among records, 2) a richer description of the data collection process and sources of variation that can affect the probability of organisms being detected, and 3) the taxonomic completeness of the data reported, all of which are only interpretable because of 4) information concerning the survey scopes (i.e. taxonomic scope, target spatial and temporal resolution). By adding terms that convey the dataset's spatiotemporal scope, sampling effort, and taxonomic completeness, the Humboldt Extension facilitates indexing of this information in databases and enables potential data users to filter, search, and use information that was previously buried in metadata text files and other ancillary documentation.

While the Humboldt Extension allows the informed integration of survey data, the expanded range of information may require more detailed steps in data discovery and analysis. For example, when searching for available data, researchers could consider restricting themselves to using data from surveys that provide complete reports of all taxa (within a defined scope) so that information on non-detections is available or can be inferred. If so, they will then need to decide whether to search for specific ranges of survey effort and protocols to aid in the process of inferring whether the non-detections indicate true absence of taxa. This added capability, enabled by the Humboldt Extension, may create more overhead for users to discover data, but it also provides needed information that was previously unavailable in the more limited vocabulary available in DwC and currently used in data discovery platforms such as GBIF and OBIS. Critically, Humboldt Extension and DwC are meant to be used together so species lists and metadata about those lists' survey protocols remain conjoined.

Facilitating the integration of data from multiple sources

The information contained using Humboldt Extension terms creates the opportunity to search for data collected within a user-specified range of protocols, allowing for the best possible approach to be used in the analysis of biodiversity data. Potentially, survey data could be directly combined if they follow similar, widely adopted data collection protocols. However, more complex integration of data following different protocols is also possible. By statistically modelling the effects of differences in data types and protocols, accuracy can be maximised and uncertainty can be appropriately captured (Fletcher et al. 2019, Adjei et al. 2023, Mäkinen et al. 2024). This should enable more comprehensive prediction of species abundance or occurrence for a full geographic and taxonomic scope, with applications such as the development of Species Population EBVs (Jetz et al. 2019).

The lack of data standards and data models for biodiversity surveys is also a barrier to integrating survey data with qualitatively different data types, such as species traits and environmental data (Huang et al. 2012, Guralnick et al. 2018). The availability and volume of remotely sensed data that measure different aspects of the environment have increased since the 1970s, with an increased resolution from pixel widths of kilometres to less than a meter (Mathieu and Aubrecht 2018). However, higher integration with in situ biodiversity surveys is needed to significantly improve the monitoring of ecosystems or species habitats (Luque et al. 2018). The spatial resolution of biodiversity data (particularly for incidental records) is generally low or not reported at all, and locations are recorded with high uncertainty (Bloom et al. 2018), which makes it difficult to align with the latest remotely sensed products. Having clearly defined terms to address the spatial extent and resolution of biodiversity data can aid in: 1) highlighting the higher resolution at which many surveys are carried out, 2) validating remotely sensed data with in situ biodiversity data, and 3) better informing models that integrate environmental and species level data about the accuracy of their results.

Incentivising data sharing

The ability to reliably credit data collectors for their work has the potential to change the incentive structure and create a cultural shift toward open science (Groom et al. 2020). Current open data infrastructures already enable data providers to be acknowledged for their contributions via dataset Digital Object Identifier (DOI) citation. The Humboldt Extension enhances this further by allowing the proper attribution of the actual data collectors (*samplingPerformedBy* term). By providing this possibility, we hope to encourage the ecological community to make their data more accessible by sharing their surveys following FAIR principles.

Community development and maintenance of the Humboldt Extension

The Humboldt Extension is already ratified as an international biodiversity standard, it is already implemented in

GBIF, and can be readily incorporated into other biodiversity platforms (e.g. OBIS, Map of Life) as well as in field data collection systems (e.g. BioCollect) or other database structures (e.g. the Field Museum database). We acknowledge that it has not been tested with all potential biodiversity survey types, nor do the terms within the Humboldt Extension cover all facets of a biodiversity survey (e.g. only some of the unbounded number of possible target scopes are covered). Expanding the Extension's capabilities remains possible by developing and adding new terms and even new structures to cover further aspects of biodiversity datasets.

With the initial version in place and validated by the wider biodiversity community, the addition of further terms to support the reporting of specific data is streamlined. Initiatives to develop standards or structures for storing other types of survey information are also underway. For example, GBIF is leading an initiative to expand the structural capabilities of data sharing via DwC (https://www.gbif.org/new-data-model). Structurally, a survey, supported by the terms of the Humboldt Extension, is an entirely new DwC class in this new data model. Further, this new data model generalises Humboldt's concept of target scopes, allowing any scope or number of scopes to be defined and included. A known shortcoming of the current state of the Humboldt Extension (and of many other extensions) is the need to develop controlled vocabularies for well-known biodiversity sampling methods and protocols, which would improve the ability of data users to filter datasets based on common protocols. To foster the development and maintenance of community-driven controlled vocabularies or other terminologies, the DwC Maintenance Group of TDWG will maintain the Humboldt Extension and, together with GBIF, will engage with the biodiversity and ecology communities to continue its ongoing evolution and fitness for purpose (Ingenloff 2025, Ingenloff et al. 2025).

Conclusion

The GBF, as well as other global treaties focused on biological diversity, emphasises the need for improved biodiversity monitoring data from local to national levels, and information sharing to support global decision-making. It is not sufficient to simply increase the quantity of monitoring data; more data may make the proper downstream processing more complicated. Instead, what is needed are ways to ensure monitoring data are in formats that require minimal effort for use in downstream analytical pipelines. The Humboldt Extension for Ecological Inventories can be a key contribution to this streamlining and improvement effort, as it can be implemented in national biodiversity data infrastructures to support their data collection systems and databases, encouraging data collectors to share their data following Open Science and FAIR principles. However, it is up to the data collectors, researchers, data managers, and aggregators to maximise its use. We call for all individuals and institutions involved in biodiversity monitoring to utilise data standards and, in particular, work with the standards community to map their monitoring schemas to this extension. This effort will ensure that biodiversity data are accessible for the benefit of the whole of society; data managers can improve the curation and maintenance of natural history collections; researchers can provide more robust assessments and modelled predictions; policymakers can make data-informed decisions; and civil society can benefit from an improved collective understanding of biodiversity status and trends.

Acknowledgements – The development of the Humboldt Extension for ecological inventories would not have been possible without the collaborative work of all members of the TDWG Humboldt Task Group within the Biodiversity Information Standards (TDWG), as well as the contributions of the larger biodiversity community. The authors are extremely thankful to all contributors who tested or mapped their datasets using the proposed vocabulary, provided feedback on the terms' definitions, provided examples of different survey types, and shared their knowledge and diverse material to broaden the data scopes that can be captured with the Humboldt Extension. Authors also appreciate the willingness of the biodiversity community to participate in the open review established by TDWG, and to reach consensus for the formal ratification of the standard. Open Access funding enabled and organized by Projekt DEAL.
Funding – The authors acknowledge the support of the Biodiversity Information Standards (TDWG) association, which hosted the Humboldt Extension Task Group for over four years. TDWG's support included community outreach, communication, and infrastructure development. The authors' institutions recognised and agreed upon the time dedicated to research, development, and engagement with the broader biodiversity community required to fulfil this study.
Conflict of interest – The authors declare no conflict of interest.

Author contributions

Yanina V. Sica: Conceptualization (equal); Investigation (equal); Methodology (equal); Project administration (lead); Visualization (lead); Writing – original draft (lead); Writing – review and editing (equal). **Wesley M. Hochachka**: Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Validation (equal); Writing – original draft (equal); Writing – review and editing (equal). **Robert D. Stevenson**: Conceptualization (equal); Methodology (equal); Writing – original draft (equal); Writing – review and editing (equal). **Kate Ingenloff**: Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Validation (equal); Writing – review and editing (equal). **Paula F. Zermoglio**: Conceptualization (equal); Methodology (equal); Writing – review and editing (equal). **John Wieczorek**: Conceptualization (equal); Methodology (equal); Software (equal); Validation (equal); Writing – review and editing (equal). **Yi Ming Gan**: Conceptualization (equal); Data curation (equal); Investigation (equal); Validation (equal); Writing – review and editing (equal). **Dmitry Schigel**: Conceptualization (equal); Methodology (equal); Writing – review and editing (equal). **Zachary R.**

Kachian: Conceptualization (equal); Data curation (equal); Investigation (equal); Validation (equal); Writing – review and editing (equal). **Steve Baskauf**: Conceptualization (equal); Methodology (equal); Software (equal); Writing – review and editing (equal). **Peter Brenton**: Conceptualization (equal); Investigation (equal); Writing – review and editing (equal). **Anahita J. N. Kazem**: Conceptualization (equal); Methodology (equal); Writing – review and editing (equal). **Walter Jetz**: Conceptualization (equal); Methodology (equal); Writing – review and editing (equal). **Robert Guralnick**: Conceptualization (equal); Funding acquisition (lead); Methodology (equal); Supervision (lead); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ecog.08223.

Data availability statement

Data are available from the Zenodo Repository: https://zenodo.org/records/17359264 (Sica et al. 2025).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Adjei, K. P., Mostert, P., Parada, J. S., Skarstein, E. and O'Hara, R. B. 2023. The point process framework for integrated modelling of biodiversity data. – arXiv, arXiv:2311.06755.

Affinito, F., Williams, J. M., Campbell, J. E., Londono, M. C. and Gonzalez, A. 2024. Progress in developing and operationalizing the monitoring framework of the global biodiversity framework. – Nat. Ecol. Evol. 8: 2163–2171.

Audiovisual Core Maintenance Group 2023. Audiovisual core. – http://www.tdwg.org/standards/638.

Bisby, F. A. 2000. The quiet revolution: biodiversity informatics and the internet. – Science 289: 2309–2312.

Bloom, T. D. S., Flower, A. and DeChaine, E. G. 2018. Why georeferencing matters: introducing a practical protocol to prepare species occurrence records for spatial analysis. – Ecol. Evol. 8: 765–777.

Bowler, D. E., Boyd, R. J., Callaghan, C. T., Robinson, R. A., Isaac, N. J. B. and Pocock, M. J. O. 2025. Treating gaps and biases in biodiversity data as a missing data problem. – Biol. Rev. Camb. Philos. Soc. 100: 50–67.

Bubnicki, J. W. et al. 2024. Camtrap DP: an open standard for the FAIR exchange and archiving of camera trap data. – Remote Sens. Ecol. Conserv. 10: 283–295.

Buchanan, G. M., Butchart, S. H. M., Chandler, G. and Gregory, R. D. 2020. Assessment of national-level progress towards elements of the Aichi biodiversity targets. – Ecol. Indic. 116: 106497.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., Wardle, D. A., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S. and Naeem, S. 2012. Biodiversity loss and its impact on humanity. – Nature 486: 59–67.

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R. V., Paruelo, J., Raskin, R. G., Sutton, P. and van den Belt, M. 1998. The value of the world's ecosystem services and natural capital. – Ecol. Econ. 25: 3–15.

Daily, G. 1999. Developing a scientific basis for managing Earth's life support systems. – Conserv. Ecol. 3: 14.10.5751/ES-00140-030214.

Dornelas, M. et al. 2018. BioTIME: A database of biodiversity time series for the Anthropocene. – Global Ecol. Biogeogr. 27: 760–786.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Fink, D., Johnston, A., Strimas-Mackey, M., Auer, T., Hochachka, W. M., Ligocki, S., Oldham Jaromczyk, L., Robinson, O., Wood, C., Kelling, S. and Rodewald, A. D. 2023. A Double machine learning trend model for citizen science data. – Methods Ecol. Evol. 14: 2435–2448.

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A. and Dorazio, R. M. 2019. A practical guide for combining data to model species distributions. – Ecology 100: e02710.

GBIF. The Global Biodiversity Information Facility. 2025a. Resource search. – https://www.gbif.org/resource/search?contentType=literature&relevance=GBIF_USED&peerReview=true.

GBIF. The Global Biodiversity Information Facility. 2025b. What is GBIF?. – https://www.gbif.org/what-is-gbif.

Gonzalez, A. and Peres-Neto, P. R. 2015. Act to staunch loss of research data. – Nature 520: 436–436.

Groom, Q., Güntsch, A., Huybrechts, P., Kearney, N., Leachman, S., Nicolson, N., ... & Haston, E. 2020. People are essential to linking biodiversity data. Database, 2020, baaa072. – https://doi.org/10.1093/database/baaa072.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R. and Wintle, B. A. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. – Global Ecol. Biogeogr. 24: 276–292.

Guralnick, R. P., Walls, R. and Jetz, W. 2018. Humboldt Core – toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. – Ecography 41: 713–725.

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B. and Schigel, D. 2021. Data integration enables global biodiversity synthesis. – Proc. Natl Acad. Sci. USA 118: e2018093118.

Hochachka, W. M., Gan, Y.-M., Kachian, Z. R., Sica, Y. V., Baskauf, S. J., Stevenson, R., Ingenloff, K., Brenton, P., Suwa, T., Putte, A. V. de and Wieczorek, J. 2023. Implementation experience report for Humboldt Extension for ecological inventories. – https://eco.tdwg.org/humboldt_extension_implementation_experience_report.pdf.

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R. and Qiao, G. 2012. Willing or unwilling to share primary biodiversity data: results and implications of an international survey. – Conserv. Lett. 5: 399–406.

Hudson, L. N. et al. 2014. The PREDICTS database: a global database of how local terrestrial biodiversity responds to human impacts. – Ecol. Evol. 4: 4701–4735.

Ingenloff, K. 2025. Survey and monitoring data quick start: a guide to updating a Darwin Core event dataset with the Humboldt extension. – GBIF Secretariat.

Ingenloff, K., Svenningsen, C., Earl, C., Shimabukuro, P. H. F., Sica, Y., Gan, Y.-M., Kachian, Z. R., Brenton, P., Hochachka, W., Wieczorek, J., Stevenson, R., Kazem, A., Baskauf, S., Zermoglio, P. F., Bloom, D., Rodrigues, A., Gamboa Martínez, J. and Schigel, D. 2025. Guide for publishing biological survey and monitoring data to GBIF. – GBIF Secretariat.

IPBES 2019. In: Brondízio, E. S., Settele, J., Díaz, S. and Ngo, H. T. (eds), Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on biodiversity and Ecosystem Services. – Zenodo, https://doi.org/10.5281/zenodo.3831673.

Jetz, W., McPherson, J. M. and Guralnick, R. P. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. – Trends Ecol. Evol. 27: 151–159.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S. and Turak, E. 2019. Essential biodiversity variables for mapping and monitoring species populations. – Nat. Ecol. Evol. 3: 539–551.

Johnston, A., Hochachka, W. M., Strimas-Mackey, M. E., Ruiz Gutierrez, V., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S. T. and Fink, D. 2021. Analytical guidelines to increase the value of community science data: an example using eBird data to estimate species distributions. – Divers. Distrib. 27: 1265–1277.

Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W. M., Julliard, R., Kraemer, R. and Guralnick, R. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. – BioScience 69: 170–179.

Luque, S., Pettorelli, N., Vihervaara, P. and Wegmann, M. 2018. Improving biodiversity monitoring using satellite remote sensing to provide solutions towards the 2020 conservation targets. – Methods Ecol. Evol. 9: 1784–1786.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J. and Langtimm, C. A. 2002. Estimating site occupancy rates when detection probabilities are less than one. – Ecology 83: 2248–2255.

Mäkinen, J., Merow, C. and Jetz, W. 2024. Integrated species distribution models to account for sampling biases and improve range-wide occurrence predictions. – Global Ecol. Biogeogr. 33: 356–370.

Mathieu, P.-P. and Aubrecht, C. (eds). 2018. Earth observation open science and innovation. – Springer.

OBIS. 2025. Ocean biodiversity information system. – Intergovernmental Oceanographic Commission of UNESCO.

Pereira, H. M. et al. 2013. Essential biodiversity variables. – Science 339: 277–278.

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L. and Desmet, P. 2014. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the Internet. – PLoS One 9: e102623.

Royle, J. A. 2004. N-mixture models for estimating population size from spatially replicated counts. – Biometrics 60: 108–115.

Sica, Y. V., Ingenloff, K., Gan, Y.-M., Kachian, Z., Baskauf, S. J., Wieczorek, J., Zermoglio, P. and Stevenson, R. 2022. Application of Humboldt extension to real-world cases. – Biodivers. Inf. Sci. Stand. 6: e91502.

Sica, Y. V., Hochachka, W. M., Stevenson, R. D., Ingenloff, K., Zermoglio, P. F., Wieczorek, J., Gan, Y. M., Schigel, D., Kachian, Z. R., Baskauf, S., Brenton, P., Kazem, A. J. N., Jetz, W. and Guralnick, R. 2025. Data from: Enabling ecological survey data integration with the Humboldt extension to Darwin Core. – Data are available from the Zenodo Repository, https://zenodo.org/records/17359264.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. and Kelling, S. 2009. eBird: a citizen-based bird observation network in the biological sciences. – Biol. Conserv. 142: 2282–2292.

Sullivan, B. L. et al. 2014. The eBird enterprise: an integrated approach to development and application of citizen science. – Biol. Conserv. 169: 31–40.

TDWG Humboldt Extension Task Group. 2024. Properties of hierarchical events in the Humboldt Extension for Ecological Inventories. – Biodiversity Information Standards (TDWG).

Tittensor, D. P. et al. 2014. A mid-term analysis of progress toward international biodiversity targets. – Science 346: 241–244.

Van De Putte, A. P., Jackson, G. D., Pakhomov, E., Flores, H. and Volckaert, F. A. M. 2010. Distribution of squid and fish in the pelagic zone of the Cosmonaut Sea and Prydz Bay region during the BROKE-West campaign. – Deep Sea Res. II 57: 956–967.

Vocabulary Maintenance Specification Task Group. 2017. Vocabulary maintenance standard. – http://www.tdwg.org/standards/642.

Wieczorek, J. R., Bloom, D., Guralnick, R. P., Blum, S., Döring, M., Giovanni, R., Robertson, T. and Vieglais, D. 2012. Darwin core: an evolving community-developed biodiversity data standard. – PLoS One 7: 29715.

Wilkinson, M. D. et al. 2016. The FAIR guiding principles for scientific data management and stewardship. – Sci. Data 3: 160018.

Wilson, E. O. 1988. Biodiversity. – National Academies Press.

WWF 2024. Living planet report 2024 – a system in peril. – WWF.