Implementation Experience Report for Humboldt Extension for ecological inventories

Authors

Wesley M. Hochachka, Yi-Ming Gan, Zachary R. Kachian, Yanina V. Sica, Steven J. Baskauf, Robert Stevenson, Kate Ingenloff, Peter Brenton, Tomomi Suwa, Anton Van de Putte, and John Wieczorek

Abstract

Access to high-quality ecological data is pivotal to assessing and modeling biodiversity and its change through space and time. Biodiversity inventory data (i.e., records of species at specific places and times) are particularly relevant to monitoring species' distributions and abundance, but their reliability for use in downstream models depends on clear reporting of the methodology implemented and associated sampling effort and completeness. This critical information about the inventory processes is often either not reported or the information is described in an unstructured manner, greatly limiting potential reuse for other analyses. In order to support the reuse of inventories and to assure better standardization of newly collected data, we developed a vocabulary to standardize inventory data reporting that is designed for broad adoption and use. The Humboldt Extension for Ecological Inventories (eco), developed as a Darwin Core Extension, has been implemented and tested with real world data. This document describes the implementation and testing process of the Extension prior to the official public review. Here we, the TDWG Humboldt Task Group, present the use of the Humboldt Extension thorough different case studies, discuss it's advantages, and propose its ratification as a vocabulary enhancement to the TDWG Darwin Core standard.

Introduction and background

Note: Because this introduction provides the rationale for this enhancement to Darwin Core and describes how its necessary features were determined, it serves as the Feature Report required by Section 4.2.1 of the TDWG Vocabulary Maintenance Specification.

While mobilization (i.e., compilation, curation, and sharing) of incidental point occurrence records has been in focus for the broader biodiversity data community for many years, sharing of inventory and monitoring data is a much more recent phenomenon. The Darwin Core standard (DwC), although most commonly used to capture data at the individual specimen or observation level, does provide an avenue by which to capture some inventory-level data. For instance, DwC contains terms such as dwc:samplingProtocol, dwc:sampleSizeValue, dwc:sampleSizeUnit, and dwc:samplingEffort, and can capture some of the information related to inventories. However, DwC does not have the capacity to accommodate explicit reporting of inventory scope (spatial, temporal, taxonomic, and environmental), sampling protocol, and a whole suite of commonly measured aspects of the inventory processes (e.g., reported measures of sampling effort and completeness).

The lack of a standardized vocabulary by which to characterize biodiversity inventory data is a persistent barrier to their mobilization, integration, and broad reuse. In an effort to overcome these limitations, Guralnick, Walls, and Jetz (2018) introduced the Humboldt Core as a proof of concept and demonstrated its implementation in Map of Life. Although originally planned as a new TDWG standard, ratification was not pursued at the time, limiting adoption by the broader community.

In 2021, the TDWG Humboldt Task Group was established to explore integration of the terms proposed in the original publication with existing standards and implementation schemas. Members of Map of Life, the Global Biodiversity Information Facility (GBIF), VertNet, Atlas of Living Australia (ALA), Ocean Biodiversity Information System (OBIS), and other partners across the larger biodiversity community met regularly to define a standardized way to accommodate the information needed to describe the inventory process. Since inventories (with or without hierarchical structure) can be considered Events, it was deemed appropriate to integrate the proposed terms as an Extension to the Darwin Core Event Core. Hence, the now renamed **Humboldt Extension for Ecological Inventories (eco)**, hereafter referred to as the Humboldt Extension, aims to include the necessary terms to more fully describe the inventory process. The task group first revised the original Humboldt Core terms from Guralnick, Walls, and Jetz (2018), reformulated definitions, comments, and examples, and discarded redundant terms or added new ones as needed. The vocabulary was then implemented in a test instance of the GBIF

Integrated Publishing Toolkit (IPT) for a limited group of data publishers to test their inventory datasets. Here we present four case studies that demonstrate the advantages of using the Humboldt Extension in terms of biodiversity inventory data standardization, sharing, and reuse before it undergoes a public review as established by the TDWG process.

Through ratification of the Humboldt Extension as a TDWG vocabulary enhancement, we expect to provide the community with a solution for capturing and sharing inventory data thereby improving biodiversity data discoverability, interoperability, utility, and reusability while lowering the reporting burden. This has been clearly called out as urgently needed in order to meet the new set of goals and targets of the Convention on Biological Diversity's Kunming-Montreal global biodiversity framework.

Development of the vocabulary

Following approval of the Task Group's charter, Task Group members held weekly meetings for over two years to discuss the Humboldt Core framework (Guralnick et al. 2018) and amend existing or propose new terms for a standardized vocabulary to be used in reporting key information on biodiversity inventories in order to maximize the usability and interoperability of these data.

Following Guralnick, Walls, and Jetz (2018), the Task Group considered biodiversity inventories to be surveys set out to document and identify a particular group of organisms (**taxonomic scope**) in a specific location (**spatial scope**, e.g., an area of land or volume of water) over a defined period of time (**temporal scope**) using a specified methodological approach (**protocols**, **sampling design**; see the Biological Collections Ontology term defining a taxonomic inventory). Inventories and other monitoring efforts are performed routinely (in space or time) and offer high-quality data characterizing biodiversity patterns and trends. Inventories have the potential to inform inferences of species co-existence or absence across a given geographic space and time, but their usability depends largely on how well the inventory process is captured (Figure 1). For instance, knowledge about sampling protocol and its suitability to address a targeted taxonomic and/or spatiotemporal scope affects the 'completeness' of the inventory (the proportion of expected species successfully detected), which defines whether an inventory can help inform about potential species absences.

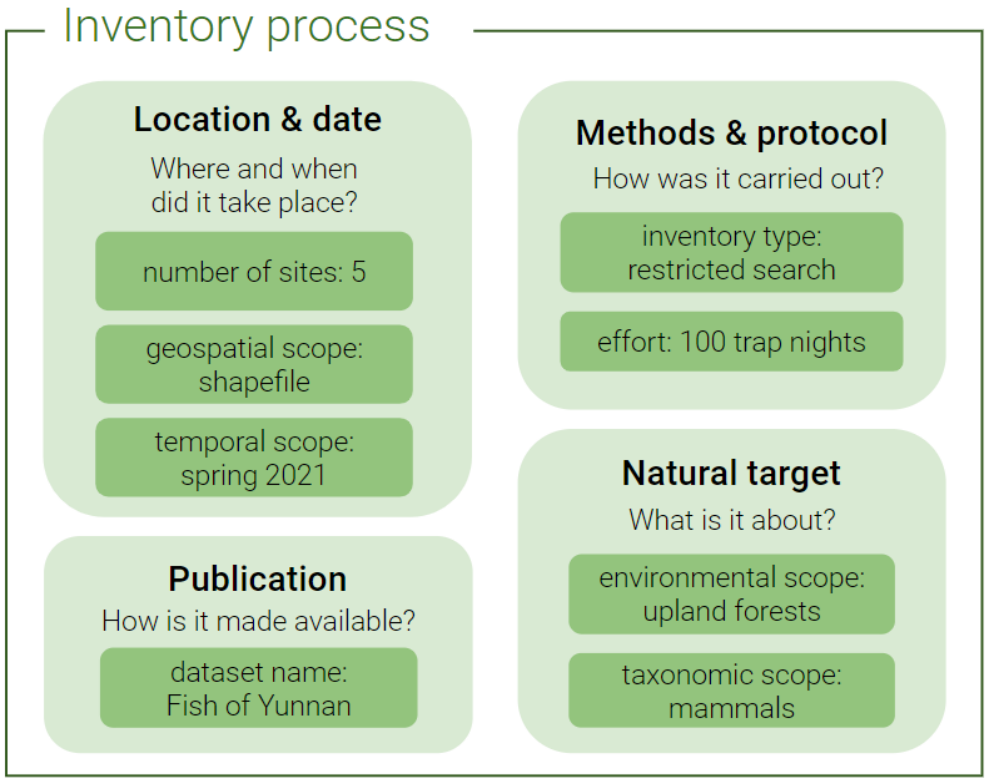

Figure 1. Elements of an inventory process that may produce a species list with examples of those processes in dark green boxes.

Within the Darwin Core framework, an inventory is considered an Event which may (or may not) have a hierarchical structure. Hence, we propose an additional 54 terms (the Humboldt Extension for Ecological Inventories, 'eco') to extend the Event class in Darwin Core and capture critical inventory process elements in a structured manner. We defined six categories (not formal classes) that describe each element of an inventory process (modified from Guralnick et al. 2018; Table 1):

- **General dataset and identification:** terms describing dataset level information.
- **Geospatial scope and habitat**: terms describing where an inventory takes place and the habitat characteristics and environmental conditions of survey sites.
- **Temporal scope**: terms describing when a survey takes place.
- **Taxonomic and organismal scope**: terms describing the target taxonomic group, life stages and/or growth forms, and degree(s) of establishment.
- **Methodology description**: terms describing the inventory methodology including details about inventory type performed, protocol(s) used, abundance, absence reporting, and presence of material samples or vouchers.
- **Completeness and effort**: terms describing inventory completeness and effort.

Within these categories, we reviewed all original terms, reformulated definitions, and discarded or added new terms where needed. The discussions were documented in a set of meeting notes and the main outcomes tracked in the Task Group GitHub repository. We also presented advances of the Task Group work at TDWG Conferences in 2021 and 2022 (Sica and Zermoglio 2021, Sica et al. 2022) and incorporated new members and feedback from participants. Proposed Humboldt Extension terms, definitions, and comments can be found in the Humboldt Extension Quick Reference Guide. All other Darwin Core terms are appropriate to include while reporting inventory data, including terms from other Extensions (for an example inventory report using dwc:, eco:, and emof: terms, see 'Use case: Distribution of squid and fish in the pelagic zone of the Cosmonaut Sea and Prydz Bay region during the BROKE-West campaign - data').

Table 1. The six main categories and relevant terms of the Humboldt Extension for Ecological Inventories (eco) proposed to extend the Darwin Core Event class to more effectively describe the biological inventory process. Terms borrowed from Darwin Core Identification class (dwc:Identification) are prefixed with "dwc:" (e.g., dwc:identifiedBy). Categories are not defined as formal classes. For updates see Note on Public Review.

Category	Humboldt Extension (eco) Terms
General Dataset and Identification	samplingPerformedBy, identifiedBy, identificationReferences
Geospatial Scope and Habitat	geospatialScopeAreaInSquareKilometers, totalAreaSampledInSquareKilometers, siteNestingDescription, siteCount, verbatimSiteNames, verbatimSiteDescriptions, targetHabitatScope, excludedHabitatScope, reportedWeather, reportedExtremeConditions
Temporal Scope	eventDuration, eventDurationUnit
Taxonomic and organismal Scope	targetTaxonomicScope, excludedTaxonomicScope, targetLifeStageScope, excludedLifeStageScope, targetDegreeOfEstablishmentScope, excludedDegreeOfEstablishmentScope, targetGrowthFormScope, excludedGrowthFormScope
Methodology Description	compilationType, compilationSourceTypes, inventoryTypes, protocolNames, protocolDescription, protocolReferences, isAbundanceReported, isAbundanceCapReported, abundanceCap, isVegetationCoverReported, hasVouchers, voucherInstitutions, hasMaterialSamples, materialSampleTypes, isLeastSpecificTargetCategoryQuantityInclusive, hasNonTargetTaxa, nonTargetTaxa, areNonTargetTaxaFullyReported, hasNonTargetOrganisms
Completeness and Effort	isSamplingEffortReported, samplingEffortProtocol, samplingEffortValue, samplingEffortUnit, taxonCompletenessReported, taxonCompletenessProtocols, isTaxonomicScopeFullyReported, isLifeStageScopeFullyReported, isDegreeOfEstablishmentScopeFullyReported , isGrowthFormScopeFullyReported, isAbsenceReported, absentTaxa

Implementation of the vocabulary: testing

In a second phase, members of the larger biodiversity community were invited to test the Humboldt Extension with real-world data. Several mapping exercises were carried out in a test instance of the GBIF Integrated Publishing Toolkit (IPT). The Task Group also aligned our efforts with other biodiversity data standards task groups (within and outside TDWG) including the Camera Trap Data Package (CamTrap DP) and the diversified GBIF Data Model). After testing numerous datasets using the proposed Extension, three case studies were selected to illustrate the broad applicability of the Extension here.

Use cases

Use case: eBird Volunteer–Collected Observations of Birds

eBird is a citizen-science project started in 2002 for the purpose of collecting, managing, and disseminating observations of birds made by bird watchers (Sulllivan et al. 2009, Sullivan et al. 2014). While data were initially collected only in the United States and Canada, the project now collects data globally on all species of wild birds. Most new records are entered via smartphone apps, although web-based data entry is also still commonly used. The Cornell Lab of Ornithology coordinates the project and maintains the database and other software infrastructure; on-staff researchers use these data for basic and applied research. Because the data are distributed at no cost, they are widely used in academic publications with over 150 publications citing eBird data in 2022 alone (https://science.ebird.org/en/research-and-conservation/publications; accessed 22 March 2023).

There are six major features of the eBird's data model worth noting:

1. All records represent observations—visual or auditory detections of wild birds (not physical specimens). In some cases, there are digital "specimens" (still photographs, video, and/or audio recordings) associated with records. While there is some retrospective correction of identification errors, the vast majority of final taxonomic identification of an observation record is created at the time of entry into the database.
2. Data are collected and stored as ***checklists***: groups of records of taxa (mostly species, but also other taxonomic ranks such as subspecies or families). Each checklist can be viewed as a dwc:Event clustering records of taxa observed by a single group of observers, during a single continuous observation period, and within some contiguous geographical region. This checklist-oriented data model differs from many datasets compiled during ecological inventories which do not emphasize the storage of records in multi-species sets based on a shared collection/observation event.

3. eBird data records for checklists contain a rich array of ancillary information about the observation event including any variation (e.g., duration of, or distance traveled during, the observation event) affecting the probability of detection of an organism. For example, assume that a species is present and detectable during an observation event. The likelihood that that species will be recorded increases as the duration of, and distance traveled during, an observation event increases. A very large proportion of variation in numbers of birds reported is the result of variation in observation effort. Thus, the ancillary information describing the observation event and reported in a structured manner is critical to controlling for variation in detection probability when fitting statistical and machine learning models to the data (Johnston et al. 2021).
4. The eBird data model contains the concept of a ***complete list***: an indicator of whether a checklist (i.e., dwc:Event) contains records of all species detected and identified during the observation event. This is a concept of completeness of reporting. eBird's data model implicitly assumes that the list of species reported will never be a complete list of species actually present. Thus, when a checklist is recorded as being a complete list, it is implied that any unreported species were not detected (i.e., counts of zero individuals can be inferred for all unreported species on complete lists), and therefore the data can be used in analyses requiring both presence and absence information. Having this detailed information in a structured manner allows more sophisticated and accurate modeling of responses such as occurrence rates and distributions (Johnston et al. 2021).
5. The potential species reported in any checklist are not constrained, except that reported species should be limited to free-living (not captive) birds. To facilitate data entry, a list of expected species in an area is provided to observers in phone apps and on the website. However, any species can be entered on any checklist and bird watchers strive to find and report locally unusual species. Records of unexpected species are flagged for human review, with assessments based on digital media, written descriptions, and correspondence with the observer(s) reporting the unusual species when necessary. The unrestricted taxonomic scope of bird species reported means that the inference of zero individuals (see point 4) can apply to any bird species not reported on a list.
6. eBird's taxonomic backbone contains species and taxonomic concepts above and below species-level designation (e.g., genus and subspecies), as well as visually recognizable hybrids. Within a given species, it is possible to report some individuals at the taxonomic level of a recognizable sub-species and other individuals at the taxonomic level of the full species. Subspecies- and species-level reports are represented as separate records (i.e., rows of data) within eBird's data model. Because of this, counts of individuals reported as taxonomic subspecies and taxonomic species must be added together by data users to ascertain the total number of individual birds reported for that species.

The eBird data model contains numerous fields not necessary for general use of the data, or that could be derived from other information stored in the eBird database (e.g., it is possible to determine which checklists contain data from any BirdLife International Important Bird area using the fields 'latitude' and 'longitude': see http://datazone.birdlife.org/site/mapsearch). Thus, the data providers started by identifying the minimal set of fields containing all of the necessary and sufficient information to estimate distributions, relative abundances, and population trends of bird species based on the experience of the Cornell Lab of Ornithology (Fink et al. 2022). They then made two tests in which they:

(1) examined whether each of these "essential" eBird data fields could be mapped to a Darwin Core or proposed Humboldt Extension for Ecological Inventories term (or terms), and
(2) experimented with fitting eBird's "essential" data fields into the current (at the time of the mapping exercise) prototype of GBIF's developing unified data model.

Most terms mapped perfectly between the two data models, however, a small number of eBird fields could not be represented exactly by terms in either the tested version of the proposed Humboldt Extension or Darwin Core. These direct mapping obstacles were associated with:

- recording information about absence of detection and completeness of reporting,
- describing observation effort, and
- the need to sum reported counts from subspecies to obtain the full count of reported individuals for species.

The Task Group discussed these concepts and deemed it appropriate to update some terms of the Humboldt Extension to account for these facets of survey effort and individual counts. Post-testing changes are described in detail in Lessons Learned and Unresolved issues/ Remaining challenges.

Use case: Field Museum Rapid Inventory Data

Field Museum of Natural History Rapid Inventories are cooperative, rigorous surveys of the biological and cultural assets of a priority landscape for conservation. Scientists and long-term residents survey plants, fishes, amphibians, reptiles, birds, and mammals to (a) identify species, natural resources, and/or landscape features with high conservation value at global, national, or local scales, (b) assess the conservation status of those assets, and (c) document threats. Since 1999, the Rapid Inventories team has conducted 31 inventories, the majority of which (24) took place in the Andes-Amazon region, six were conducted in Cuba, and one in China. Read more about the Field Museum's Rapid Inventories Program here.

Collected data are derived from specimens, tissue samples, direct observations, machine observations (camera traps or auditory recordings), and interviews with local experts. Rapid Inventories are designed such that each taxonomic group uses the same locations, the exception being fishes which use locations specific to fish surveys.

Rapid Inventory data already conform to Darwin Core standards in the Field Museum's internal database, so the testing team selected a representative subset of data from a single inventory and attempted to map it to the proposed Humboldt Extension. The selected data included 423 occurrences collected using one sampling method at 27 sites in the sampling region over 15 days. The team confirmed that the breadth of sampling methods included in the larger dataset would fit the eco:eventID and eco:parentEventID structure. The selected data were transcribed from the Field Museum's database and manually mapped to Humboldt Extension terms in a spreadsheet. Most of the data mapped perfectly with the proposed Extension, however testers noted that the definitions and comments of some terms, such as eco:samplingEffortProtocol, eco:protocolReference, and eco:isVegetationCoverReported, were unclear, requiring consultation with the Task Group to ensure correct mapping. In a few instances, data could not be mapped at all.

No unresolved issues or identified challenges remained after discussion with the TDWG Task Group. Challenges encountered during mapping are described in the Lessons Learned section.

Use case: Distribution of squid and fish in the pelagic zone of the Cosmonaut Sea and Prydz Bay region during the BROKE-West campaign - data

The Antarctic GBIF/OBIS node identified a marine species inventory dataset, consisting of catch data collected by Rectangle Mid-water Trawl (RMT) during the BROKE-West cruise in 2006 in Antarctica (Van de Putte et al. 2010), to test the proposed Humboldt Extension. The dataset includes fish, squid, and bycatch (i.e., non-target taxa) sampled using Rectangular Midwater Trawl (RMT) in the Cosmonaut Sea and Prydz Bay during the BROKE-West campaign (Figure 2).

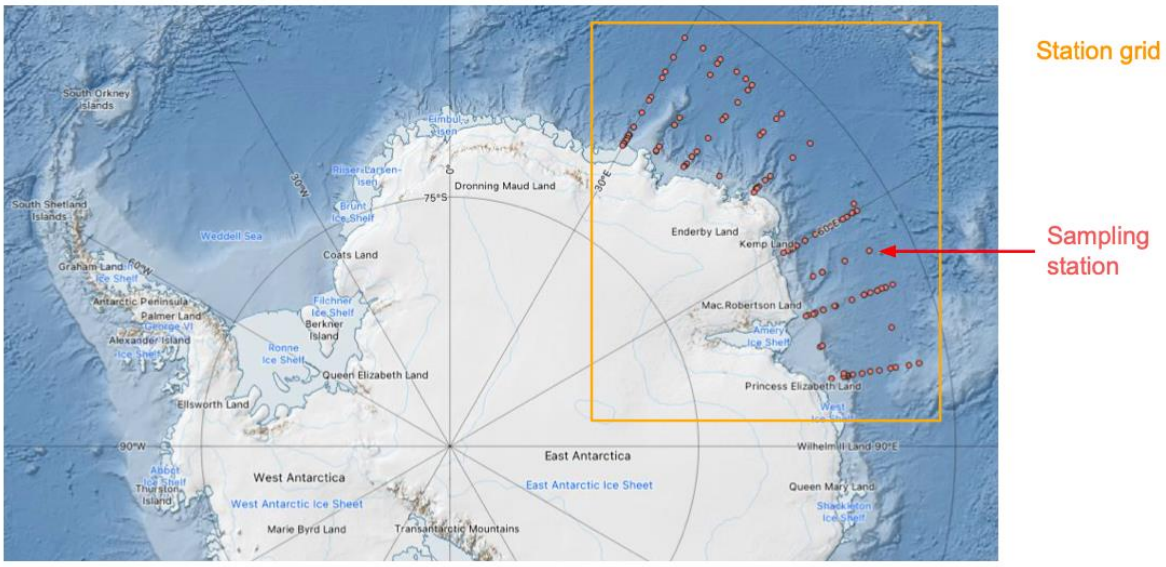

Figure 2. Rectangular Midwater Trawl (RMT) sampling sites during the BROKE-West campaign. Predetermined sampling stations (red points) are situated within the station grid (yellow square). Map of Antarctica plotted in projection EPSG:3031.

The example dataset is available at:

- https://ipt.gbif.org/resource?r=brokewest-fish
- https://www.gbif-uat.org/dataset/c60ec6f0-1089-4c1e-b25c-deb0a5a5c38b

The dataset was mapped to the testing version of the proposed Humboldt Extension terms and published on an Integrated Publishing Toolkit (IPT) test instance. See the 'published' dataset here. A visual overview of that mapping is presented in Figure 3.

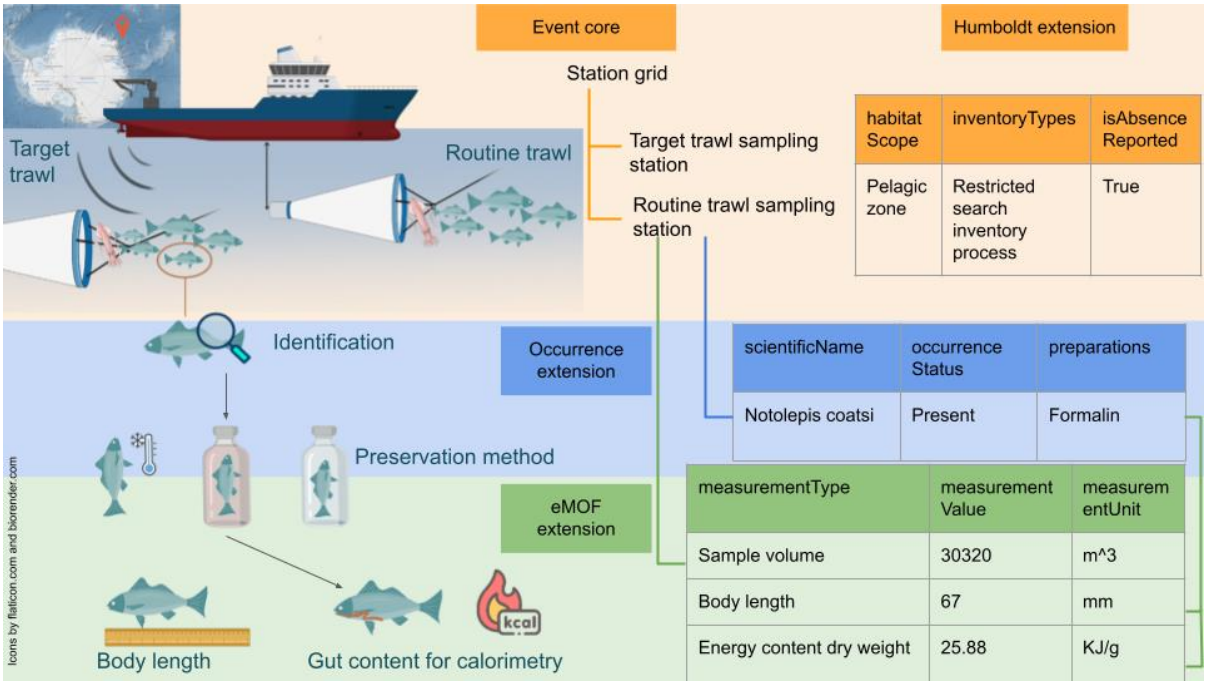

Figure 3. Schematic overview of the BROKE-West campaign's catch data using Darwin Core. The research vessel visited distinct sampling stations where two types of trawls were performed: target trawls and routine trawls. Target trawls were not pre-planned and were aimed at acoustically detected aggregations. The net was lowered to the depth of the aggregation, and opened and closed remotely. The catch was identified to the lowest taxon level possible and individuals' body lengths were measured. Specimens were either frozen or preserved in formalin or ethanol. The energy content of gut contents of formalin-preserved fish was determined using calorimetry. The spatiotemporal information is recorded using terms in the Event Core. The sampling effort and protocols are described in the Humboldt Extension. Information about the occurrences of species are described in the Occurrence Extension. Biotic (body length, energy content of gut content) and abiotic measurements (temperature, salinity) are recorded in the eMOF Extension.

Mapping of the dataset posed particular challenges. The data providers mapped the dataset to multiple Darwin Core Event Extensions beyond the proposed Humboldt Extension including the occurrence and extended measurement of fact (emof) Extensions. See the Humboldt Extension User Guide for detailed information about this mapping exercise. Identified advantages and limitations of the Humboldt Extension for ecological inventories are discussed in detail in the Lessons Learned and Unresolved issues/Remaining challenges sections below.

Use case: Hummingbirds of the Northern Andes

Map of Life (MoL; mol.org) assembles and integrates different sources of data describing species distributions worldwide. These data include expert species range maps, species occurrence

points, ecoregions, and protected areas from providers like IUCN, WWF, GBIF, and more. All data assets are stored, managed, backed up, and accessed using a hosted cloud instance. One compilation dataset, the Hummingbirds of the Northern Andes (Parra et al. 2019) was selected to test the proposed Humboldt Extension. The compilation dataset includes surveys reporting 4208 records from 162 species and 727 unique 1km sites. See the MoL dataset here.

Most data mapped readily to the DarwinCore and Humboldt Extension terms. See the published data here. The only direct mapping obstacle was associated with the number of species recorded at a given sampling site. The Task Group discussed these concepts and suggested that the species count per site be included under the OBIS (extended measurements or facts) Extension. Post-testing changes are described in detail in Lessons Learned.

Lessons learned

Multiple challenges were identified during the testing phase including:

- lack of clarity in term definition, description, or comments
- lack of appropriate term available to accommodate recorded survey data

In this section we focus on the challenges that we were able to resolve and incorporate in the current Extension.

Term descriptions, comments and examples

All three use cases identified Humboldt Extension terms requiring clarification of their definitions and/or comments, specifically eco:samplingEffortProtocol, eco:protocolReference, and eco:isVegetationCoverReported. Improvements were cover:

- *Taxon and organism target and completeness*: The Task Group improved definitions regarding taxon and organism target and completeness as a result of the eBird use case.
- *eco:samplingEffortProtocol and eco:protocolReference*: Two types of protocol documentation are generally available for the Field Museum's Rapid Inventory data. The Humboldt documentation did not clearly specify if data providers should include the citation for the protocol section of the museum's internal survey report or the citation of an external publication. The Task Group concluded that, where multiple citations are available, all should be included with a pipe separator (e.g., citation 1 | citation 2).
- *eco:isVegetationCoverReported*: The boolean term eco:isVegetationCoverReported also presented some challenges as Rapid Inventories data include information pertaining to both the status of vegetation cover but also on the values/metrics for the Events' vegetation cover. The Task Group concluded that the values for vegetation cover would fit under eco:verbatimSiteDescription.

Missing Terms

The Rapid Inventories use case illuminated several gaps in terms proposed in the Humboldt Extension.

- *Completeness*: eBird data included explicit information relating survey completeness of reporting. In order to accommodate this information and ensure maximum reusability of inventory data, the Task Group opted to add a term allowing data providers to indicate whether the dataset contains the records of all species that were detected and identified

during the Event (eco:isTaxonomicScopeFullyReported , eco:isDegreeOfEstablishmentScopeFullyReported, eco:isGrowthFormScopeFullyReported, eco:isLifeStageScopeFullyReported).

- *Conservation status and level of human impact*: The Rapid Inventory data testers also inquired about a method of denoting conservation status or level of human impact. The Task Group suggested this information could be included either under eco:habitatScope or eco:verbatimLocationDescription.
- *Migration status*: Rapid Inventory data may include information on migration status, as this can affect the sampling. The Task Group opted not to add a new term to capture this information, recommending instead that such information be included Instead, it was recommended to include this type of information under the OBIS Extended Measurement or Fact Extension (EMoF).
- *Moon phase*: Rapid Inventory data may also include information on moonphase or migration status, as this can affect the sampling. The Task Group opted not to add a new term to capture this information, recommending instead that such information be included Extended Measurement or Fact Extension (EMoF).
- *Number of species counted per site*: The Hummingbirds of Northern Andes dataset included the number of species counted at each site, which could not be mapped to either DwC Event core, occurrence Extension, or Humboldt Extension. After discussion with the Task Group, the term was mapped to the OBIS EMoF Extension.
- *Traditional knowledge*: Rapid Inventories data sometimes include local or traditional knowledge (merging different knowledge systems and survey designs). The Task Group suggested this information could reside in eco:compilationSourceTypes.

Data model and Event parent/child structure

The Field Museum Rapid Inventory data testers struggled to determine the optimal data model to follow. Specifically, there was uncertainty in how best to structure dwc:eventID and dwc:parentEventID as Rapid Inventory data could be organized by location (survey site) or by protocol. For these inventories, the same sampling locations are used for each taxonomic group (plants, birds, herps, mammals), each with its own sampling protocol. The question was: *is it best to structure the data that each sampling site be a "parent event" with a "child event" for each protocol, or the other way around?* After multiple iterations of testing different data structures, the testers and the Task Group determined that a location-first approach, such that each parent event (location) has a sub- or child Event for each protocol, was ideal. Structuring the data in this manner ensured that all lower Events inherit the same protocol. The Task Group updated the Humboldt Extension documentation based on this exercise to help capture different data

structures more readily and developed a specific document that provides guidance on the use of Humboldt Extension terms in the context of parent and child dwc:Events (see here).

Absence of detection and completeness of reporting

Following the eBird case study, it was deemed appropriate to improve the definitions regarding taxon and organism target and completeness and include a term that allows users to indicate whether the dataset contains the records of all species that were detected and identified during the Event. This concept of *completeness of reporting* implies that unreported species were not detected (i.e., counts of zero individuals can be inferred for all unreported species on complete lists), and therefore the data can be used in analyses that require both presence and absence information. Having this detailed information in a structured manner allows more sophisticated and accurate modeling of responses such as occurrence rates and distributions (Johnston et al. 2021). The ability to infer absences is necessary to maximize the capacity for reuse of shared biodiversity inventory data and simply; to do so, users need both the targeted list of species and a standard measure of completeness.

The Task Group acknowledges that inferring absences is a persistent issue in ecological problems that requires sophisticated modeling efforts after data collection. Hence, the Humboldt Extension Task Group is collaborating with GBIF, which is exploring a new data model that can accommodate information about species' absences (access detailed information on this use case here).

Inference about organism quantities and abundance

The three use cases presented here each captured different ways of entering abundance (count) data. In light of this, the Task Group determined it necessary to include a term, *isLeastSpecificTargetCategoryQuantityInclusive*, informing the user how to calculate total counts of individuals or organisms in an inventory based on dwc:organismQuantity. Additional information on this term can be found here.

Use of IRI to avoid long list of objects

Some terms proposed in the current Extension (*e.g.*, eco:samplingPerformedBy, eco:targetTaxonomicScope) either require or can accommodate entry of multiple values differentiated using a pipe separator, | . In situations where the list values are previously published, it is ideal that the IRI version of the terms be entered into the field. For example, if the

list of expected species at a particular site was previously published, or the list of individuals carrying out a survey is already detailed in some publication, the links can be used to link the information.

Non-target taxa

Occurrences of any taxa that are not included in eco:taxonTaxonomicScope are sometimes included in datasets in the form of bycatch (e.g. BROKE-West dataset). Discussion within the Task Group and with data testers indicated that the ability to explicitly indicate if out-of-scope data are included in the dataset is valuable. Thus, the Task Group added four additional terms which are essentially an invitation to data providers to include data that they might not otherwise include (i.e., bycatch). Three terms focus specifically on taxonomic scope addressing the questions:

- *Does the dataset include data outside the taxonomic scope of the survey* (Boolean; eco:hasNonTargetTaxa)*?*
- *Which non-target taxa are included* (List; eco:nonTargetTaxa)*?*
- *Was the reporting of non-target organisms comprehensive or complete* (Boolean; eco:nonTargetTaxaFullyReported)*?*

The fourth term (Boolean; eco:hasNonTargetOrganisms) allows the data provider to indicate if organisms out of scope for non-taxonomic scopes (e.g., life stage, growth form) are included. See the full documentation for these terms on the Humboldt Extension Vocabulary List of Terms.

Unresolved issues/Remaining challenges

Multiple challenges were identified during the testing phase. This section focuses on the challenges that remain unresolved as they were out of the scope of this Task Group and may require discussion in a broader forum.

Paired information of multiple fields of an Event must fit into one record

Some terms proposed in the Humboldt Extension contain multiple paired values that extend a single Event. For example, the multiple target scope terms in the Humboldt Extension might have been consolidated in a single field such as eventRemarks in the Darwin Event Core. While useful

in its own right, it is important to note that data users will need to be aware that, to interpret the data correctly, some of these fields must be read together. See Example 1 below.

In other instances, however, neither Darwin Event Core nor the proposed Humboldt Extension include all terms necessary to directly map from a data provider's template requiring data providers to consolidate multiple fields from their own data into fewer fields in the Extension. See Example 2 below.

Example 1. Antarctic GBIF/OBIS use case

Implementation of the target scope terms in the Antarctic GBIF/OBIS use case illustrate this issue well. The dataset includes information about target taxonomic scope and target life stage scope. When attempting to map the data using the Darwin Event Core terms, both sets of information would be lumped under dwc:eventRemarks for a single parent event (Table 2a). Mapping the data to the Humboldt Extension, however, means that the data are divided between two terms (eco:targetTaxonomicScope and eco:targetLifeStageScope) as a sub-Event (Table 2b). In this context, in order to fully understand the context of eco:targetLifeStageScope, the data user must also know eco:targetTaxonomicScope.

Table 2. Example of paired values for eco:targetTaxonomicScope and eco:targetLifeStageScope from the Antarctic GBIF/OBIS use case following protocol for (a) Darwin Event Core and (b) Humboldt Extension for Ecological Inventories.

a) Darwin Event Core	
eventID	dwc:eventRemarks
BROKE_WEST_RMT_006	All life stages of Myctophidae were targeted | Only larvae and juvenile of Macrouridae, Artedidraconidae, Channichthydae and Nototheniidae were targeted by the sampling protocol

b) Humboldt Extension for Ecological Inventories		
eventID	eco:targetTaxonomicScope	eco:targetLifeStageScope
BROKE_WEST		
BROKE_WEST_RMT_006	Myctophidae | Macrouridae | Artedidraconidae | Channichthydae | Nototheniidae	all | larvae and juvenile | larvae and juvenile | larvae and juvenile | larvae and juvenile

This challenge could be resolved by having two Extensions within the star schema (e.g., in addition to the Humboldt Extension, another Extension for targets could be added).

Example 2. eBird use case

In the eBird case study, the challenge of accommodating all eBird's terms describing observation effort remains because every type of inventory uses a different methodology and the relevant metrics of collection/observation effort will differ across these different methodologies. It would be infeasible for any data model to contain separate terms for every possible descriptor of effort. Some metrics, such as time in collection/observation, are ubiquitous and are represented in Darwin Core + Humboldt Extension, and in the GBIF unified model, as separate terms. Other metrics of effort, such as eBird's number-of-observers metric of effort, will need to be placed into more generic terms that describe both the units of effort and the quantity of effort. Both Darwin Core + Humboldt Extension, and the GBIF unified model, contain appropriate terms to make these generic mappings.

These challenges are due to the limitation of the current model of Darwin Core Archive. The Humboldt Extension has been used as a use case to provide feedback to and potential solutions for GBIF on questions of sharing target information and enabling abundance and absences of detection reporting to be shared explicitly.

Properties of hierarchical Events

Many of the terms in the Humboldt Extension for Ecological Inventories can be applied to any Event in a hierarchy. As a general rule, it is recommended that a parent Event reflects the characteristics of the set of all of its child Events. For example, if a parent Event has children whose eco:targetTaxonomicScope include "Canidae" and "Hyaenidae", then the eco:targetTaxonomicScope for the parent Event would also include these two scopes. Similarly, if all children of a parent Event have material samples (eco:hasMaterialSamples=true), then the parent Event also has material samples and eco:hasMaterialSamples must be true for that Event as well. Conversely, if not all children of a parent Event have material samples then the parent Event cannot have ecohasMaterialSamples=true, because that is not a characteristic of the set of all that Event's child Events. In this latter case the value for eco:hasMaterialSamples for the parent Event should remain blank because neither "true" nor "false" is correct.

The Task Group acknowledges these challenges and the solution is to document this as Comments for all of the terms whenever applicable.

Other challenges identified in broader testing

Other challenges identified during data testing and discussed at Task Group meetings. The Task Group concluded that many of the challenges encountered were out of scope for the Extension and should be submitted as an issue to other, more appropriate TDWG Task Groups. For example:

- The need to develop a more robust description for dwc:identifiedBy was forwarded to the Darwin Core maintenance group (*e.g.*, https://github.com/tdwg/dwc/issues/318).
- The need for more detailed descriptions on machines making observations and on the confidence of taxonomic identification. These issues were also shared with the Darwin Core maintenance group as they do not fall within the Event level.
- The need to specify a way in which to capture a list of values that are related to the list of values of other terms was presented with a use case to the Technical Architecture Group (TAG; see https://github.com/tdwg/tag/issues/43).

Conclusions

After two years of discussions aimed at improving the reporting of inventory data, this Task Group, composed of members of the biodiversity community, propose the addition of 54 terms to Darwin Core. These terms are grouped in the form of a vocabulary enhancement called the Humboldt Extension for Ecological Inventories. The terms have been carefully curated, mapped to different data models, and their implementation thoroughly tested. We have presented three (3) case studies representing significantly different inventory processes using a combination of dwc:, eco:, and emof: terms. Although some challenges remain, the Task Group resolved all within our scope. We believe that the Extension successfully accommodates all the metadata and dataset descriptions of multiple inventories tested and therefore propose to proceed with the TDWG ratification process.

Note on public review (Feb 2024)

In September 2023, the Humboldt Extension for Ecological Inventories went through a public review process that lasted four months. Over this period, many actors from the larger biodiversity data community reviewed the proposed terms and exchanged comments about the proposed Extension via direct communication with the Task Group or via GitHub in accordance with the protocols stated by the Darwin Core Maintenance Group (https://github.com/tdwg/dwc/wiki/Darwin-Core-Maintenance-Frequently-Asked-Questions).

All external comments were addressed by the Task Group and modifications to the Extension vocabulary were made when appropriate. From the 55 terms submitted for public review, 5 terms were added, 4 terms were removed and 12 terms were modified (edits to the definitions or comments or examples). For a detailed description of all comments and their resolution see https://github.com/tdwg/hc/issues. This process radically improved the proposed Extension; the current List of Terms can be found in https://eco.tdwg.org/list/.

References

Fink, D., Auer, T., Strimas-Mackey, M., Lignocki, S., Robinson, O.J., Hochachka, W.M., Jaromczyk, L.O., Rodewald, A.D., Wood, C., Davies, I., and Spencer, A. (2022) eBird Status and Trends, Data Version: 2021; Released: 2022 Cornell Lab of Ornithology. Ithaca, New York. https://doi.org/10.2173/ebirdst.2021

Guralnick, R., 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. https://doi.org/10.1111/ecog.02942

Johnston, A., Hochachka, W.M., Stimas-Mackey, M., 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 eBird data to estimate species occurrence. Diversity and Distributions, 27, 1265–1277. https://doi.org/10.1111/ddi.13271

Sica, Y.V. and Zermoglio, P.F. (2021) Unlocking Inventory Data Capture, Sharing and Reuse: The Humboldt Extension to Darwin Core. Biodiversity Information Science and Standards 5: e74275. https://doi.org/10.3897/biss.5.74275

Sica, Y.V., Ingenloff, K., GAN, Y-M, Kachian, Z., Baskauf, S.J., Wieczorek, J., Zermoglio, P.F., and Stevenson, R.D. (2022) Application of Humboldt Extension to Real-world Cases. Biodiversity Information Science and Standards 6: e91502. https://doi.org/10.3897/biss.6.91502

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. Biological Conservation, 142, 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J.W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.A., Merrifield, M., Morris, W., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K.V., Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W., Wood, C.L., Yu, J. and Kelling, S. (2014) The eBird enterprise: an integrated approach to development and application of citizen science. Biological Conservation, 169, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

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 Research Part II: Topical Studies in Oceanography, 57(9–10), 956–967. https://doi.org/10.1016/j.dsr2.2008.02.015