



Publishing data on disease vectors, hosts and pathogens through biodiversity data platforms

Paloma Shimabukuro • Lindsay Campbell • Florence Fouque • Josiane Etang • Soledad Ceccarelli • Quentin Groom • Kate Ingenloff • Cecilie Svenningsen • Marie Grosjean • Javier Gamboa Martínez • Dmitry Schigel

Version 51aeeca, 2025-09-10 15:39:55 UTC

Table of Contents

Colophon	1
Suggested citation	1
Authors	1
Licence	1
Persistent URI	1
Document control	1
Cover image credit	1
1. Introduction	1
1.1. Rationale	1
1.2. Target audiences	2
1.3. Introduction to vector data	2
1.4. Steps to publish a dataset in GBIF	3
1.4.1. Choose a dataset class	4
1.4.2. Data cleaning and mapping the data to the DwC standard	4
1.4.3. Publishing the dataset with the IPT	6
1.4.4. Metadata	6
2. Data categorization	11
2.1. Vector data mapping: single file	14
2.1.1. Mapping sampling events	14
2.1.2. Mapping occurrences	21
2.2. Vector data mapping using extensions	23
2.2.1. Extended Measurement or Facts Extension	23
2.2.2. Measurement or Facts Extension	28
2.2.3. Occurrence Extension	30
2.2.4. Humboldt Extension for Ecological Inventories	30
2.2.5. Resource Relationship	33
2.2.6. Species Profile	35
2.2.7. Species Distribution	37
2.3. Specific requirements for publishing vector data	44
2.3.1. How to better describe species complexes/assemblages or sibling species with the DwC standard	44
2.3.2. How to best describe the native status of vectors	45
2.3.3. How to handle data that does not fit any DwC term available	45
2.3.4. Controlled vocabulary	46
2.3.5. Unique identifiers within datasets	46
2.3.6. Additional recommendations on relevant terms	46
2.3.7. How to handle sensitive data	53
3. Future prospects	54
Acknowledgements	54
Glossary	55
References	58
Appendices	61
Vocabularies	61

Useful links to resources	65
General information and guides.....	65
Coordinates, mapping and georeferencing	65
Controlled vocabulary and ontologies	66
Taxonomy and date issues	66
Quality control tools and resources	66
Publishing tools	66

Colophon

Suggested citation

Shimabukuro PHF, Campbell L, Fouque F, Etang J, Ceccarelli S, Groom Q, Ingenloff K, Svenningsen C, Grosjean M, Martínez JG & Schigel D (2025) Publishing data on disease vectors, hosts and pathogens through biodiversity data platforms. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/doc-mjj8-ng28>

Authors

Paloma Shimabukuro, Lindsay Campbell, Florence Fouque, Josiane Etang, Soledad Ceccarelli, Quentin Groom, Kate Ingenloff, Cecilie Svenningsen, Marie Grosjean, Javier Gamboa Martínez & Dmitry Schigel

Licence

The document *Publishing data on disease vectors, hosts and pathogens through biodiversity data platforms* is licensed under [Creative Commons Attribution-ShareAlike 4.0 Unported License](#).

Persistent URI

<https://doi.org/10.35035/doc-mjj8-ng28>

Document control

Version 1.0.1, September 2025

Cover image credit

GBIF Secretariat 2025, licensed under [CC BY 4.0](#)

1. Introduction

1.1. Rationale

Vector-borne diseases (VBDs) are responsible for approximately 17% of the global burden of all infectious diseases ([WHO 2017a](#)) posing a significant threat to public health and economies, particularly in tropical regions where many of these diseases are endemic. While some VBDs like malaria remain endemic in many countries. In 2023, there were an estimated 263 million new malaria cases in 83 countries worldwide. Malaria case incidence, which accounts for population growth, rose in the period 2015-2023 from 58 to 60.4 cases per 1000 population at risk ([WHO 2024a](#)), other VBDs like dengue are re-emerging in previously endemic areas and are continuing to expand in geographic range. In 2020, GBIF established an **expert task group** to help its network improve the discovery, access and use of biodiversity data of species linked to human diseases, with a strong focus on arthropod vector data.

In 2023, a GBIF-commissioned review published by [Astorga et al \(2023\)](#) compared studies related to human health that used GBIF-mediated data on biodiversity-deemed "positives"—to "negatives" that did not. The authors found distinct differences as the positive list came from biological and ecological

sciences and used data on **host** and **vector** species. The negative list, on the other hand, focused on medicine, public health and veterinary science, suggesting that data shared through GBIF is contributing to more broad scale ecological analyses and less to health-related studies (Astorga et al. 2023).

With this guide, we hope to encourage the publication of vector data under the FAIR Principles of *findability*, *accessibility*, *interoperability* and *reproducibility* (Wilkinson et al. 2016), not only to contribute to a better understanding of disease biology, ecology, and transmission, but also to inform epidemic preparedness and response as well as VBD control and elimination strategies.

Publishing vector, host and pathogen data as occurrences under the FAIR principles has many benefits:

- Recognition for work carried out at the forefront, such as laboratory and field activities, with attribution and credit
- Increased awareness of the importance of producing good quality data by learning about the steps involved in producing and reusing data
- Increased visibility of institutions and compliance with regional, national and international standards/guidelines on open data
- Contribution to global knowledge of biodiversity
- Expanded possibilities for collaboration through exposure in an international repository
- Tracking of data use that can contribute to metrics and impact indicators of the work carried out
- Increased citation since datasets published in GBIF are assigned a DOI

1.2. Target audiences

This guide has been developed for researchers, students and health workers involved in the collection of data related to arthropods vectors of pathogens.

We aim to provide guidance on publishing data on **vectors**, **hosts** and pathogens that cause human and animal diseases through GBIF—the Global Biodiversity Information Facility.

In this guide, we provide two ways of data publication:

- A simplified way to publish all available data in a single file
- A complex way using extensions to accommodate the diversity of variables recorded along with vector records.

1.3. Introduction to vector data

This guide is focused on data from vectors of diseases and associated data content related but not limited to <host,hosts>, <reservoir,reservoirs> and pathogens, often assessed directly from fieldwork or field observations or from laboratory detection assays. For the purpose of this guide, we will refer to both host and reservoir as simply "hosts." For a detailed discussion on the concept of host and reservoir, see Ashford 2003 and Haydon DT et al. 2002.

Vector data can be a mixture of the currently supported **occurrence** and **sampling-event** classes that can also include interaction data between vectors, hosts and pathogens. These data can be layered with information on environmental variables and/or host/vector morphological measurements, molecular data such as genomics or host/pathogen detection assays, combined vector/host/pathogen data and surveillance design data that would require the use of extensions,

such as those for [extendedMeasurementOrFact](#), [MeasurementOrFact](#), [ResourceRelationship](#), and [Humboldt](#), respectively.

Vectors are most commonly sampled for research purposes or for disease surveillance and monitoring, as well as for early-warning systems. Trapping most often occurs in a systematic way, and it is possible to infer species diversity, density, abundance, infection rates, [biting rates](#), resting/biting habits, etc, from such collections. In most countries, both researchers and health programmes follow WHO standardized sampling protocols and guidelines ([WHO 2011](#), [WHO 2018](#), [WHO 2019](#)) used for day-night/indoor-outdoor/domestic-sylvatic collection of vectors. Thus the data output is appropriate for publication as a [sampling event](#) dataset. However, opportunistic, sporadic sampling is not uncommon and the output of such sampling is more appropriate to be published as [occurrence](#) datasets.

The detection of human pathogens in arthropod vectors is generally called xenomonitoring and it is used to estimate the risk of human exposure to transmission of different vector-borne pathogens. Therefore, pathogen data are most commonly obtained from laboratory screening/detection assays, e.g. qPCR or traditional PCR from midgut/tissue/blood samples from both vectors and hosts, serological tests, or direct observation after dissection (microscopy).

While data on hosts is often obtained from traps for blood seeking stage of the vectors, blood meals are analysed either through serology, e.g. precipitin tests, Enzyme-Linked Immunosorbent Assay (ELISA) or genomic techniques, e.g. PCR, high throughput sequencing. Host data might include body measurements, like weight, length, body mass and the axis of measurement for the organism, e.g. snout-vent length, wing length.

A number of variables related to bionomics (species, [human biting rate](#), [anthropo-/zoophily](#), [endo-/exophily](#), [endo-/exophagy](#)), physiological ([parity](#)), epidemiological ([sporozoite rate](#), proportion of blood meals on humans, entomological inoculation rate), insecticide resistance status and genetic (molecular forms, haplotypes, resistance alleles and genes) data may be available; these variables are important for the understanding of disease transmission dynamics and for implementing evidence-based vector control approaches.

And finally, environmental parameters are also recorded, e.g. air temperature and air humidity, water variables such as pH, dissolved oxygen concentration (DOC), conductivity, turbidity.

GBIF's current data model does not support the publication of all the variables described, i.e. interaction data, bioecological and environmental data. Darwin Core terms do not include an appropriate description for interaction data. However, it is possible to publish these types of data by using different extensions. Those for [ResourceRelationship](#), [MeasurementOrFact](#), and the newly ratified [Humboldt Extension for Ecological Inventories](#) extensions can be used to accommodate interactions and different types of measurements, but they can be challenging to use and can generate highly complex formats even for individual sampling events.

For more information on the terms used in this guide, please refer to the [Glossary](#).

1.4. Steps to publish a dataset in GBIF

Anyone interested in publishing a dataset in GBIF can do so by registering an organization - GBIF only publishes datasets from individuals affiliated to an organization. Please check the page on how to [become a publisher](#) and the [quick guide to publishing data through GBIF](#).

The next step is to prepare the data into files that will be published as Darwin Core Archives (see [data standards](#)), prior to publication is important to perform a data cleaning step and prepare the metadata, once the data is uploaded in an IPT, the data holder must select one of the three licences used by GBIF and then, the dataset can be published and registered as a resource in GBIF. [Figure 1](#)

summarizes the basic steps to data publication in GBIF.

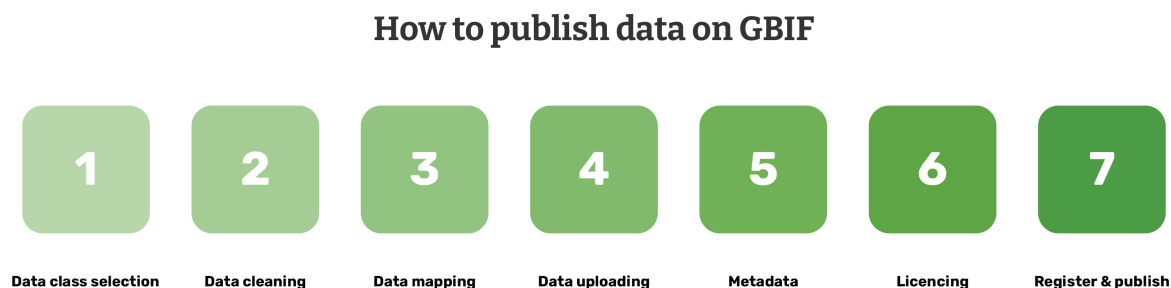


Figure 1. A simplified flow for data publishing in GBIF. Steps will be referred to in the text below with a link to the appropriate sections in this guide.

1.4.1. Choose a dataset class

There are four [dataset classes](#) that are supported for publication in GBIF:

- Metadata-only
- Checklist
- Occurrence
- Sampling-event datasets

For more information, see [how to choose a dataset class on GBIF](#).

For more details on data categorization, read [§2](#) below.

1.4.2. Data cleaning and mapping the data to the DwC standard

Once a dataset class is selected, the data must be transformed to comply with a [data standard](#) that is accepted in GBIF, the most widely used standard in biodiversity is the [Darwin Core standard](#), there are [more than 180 Darwin Core terms](#) that are stable and provide vocabularies related to organisms and their related data. Most terms cover the wide range of variables present in vector collection, but some variables, such as environmental, genomic data, among others are best represented by [extensions](#), which are sets of terms corresponding to a specific category of data.

It is considered best practice to make a copy of the original data file to be transformed into a structured dataset, which is a table with headers in the first row and with no colours, comments, no merged cells or extra formatting, no macros.



Unknown, absent and/or missing data must be left as blank cells.

IMPORTANT Do not use [N/A](#), [zero](#), [?](#) or any other value.

Data transformation follows a set of steps necessary to format some terms, i.e. dates must follow [ISO 8601](#) and geographical coordinates must be in decimal degrees, then we map the original column headers in the spreadsheet to Darwin Core terms. It is during this step that data cleaning and data quality checks can be performed, this step can be performed in any spreadsheet software such as Excel, LibreOffice or Modern CSV or with dedicated software such as [OpenRefine](#).

	A	B	C	D	E	
1	scientificName	lifeStage	sex	individualCount	locality	samplingProtocol
2	VectorSpeciesName	VectorLifeStage	VectorSex	NumberOfVectorsCaught	LocationName	
3	Phlebotomus neglectus	Adults	female	1	Constanța	Light trap
4	Culicoides imicola	Adults	female	56	Pelagoniski	Onderstepoort \
5	Hyalomma marginatum	Adults	male	3	Kukës	Collection from

Figure 2. Example of DwC mapping to the original headers in a table. Suggested Darwin Core terms are mapped in the first row, original headers in the second row and examples are given in the following three rows.

Table 1. Date formatting under the ISO 8601 standard. Left column: original data in different formats; middle column: standardized date formats.

CollectionDate	dwc:eventDate	ISO 8601
3 jan 2002	2002-01-03	YYYY-MM-DD
23/10/04	2004-10-04	YYYY-MM-DD
14-08-95	1995-08-14	YYYY-MM-DD
Dec 2012	2012-12	YYYY-MM
2015	2015	YYYY
23-24/Oct/2004	2004-10-23/2004-10-24	YYYY-MM-DD/YYYY-MM-DD
11/2009 to 12/2009	2009-11/2009-12	YYYY-MM/YYYY-MM



Figure 3. Mapping species names. In this example, three different publishers refer to species names using different headers (lines 2). When mapping to DwC, species names become the `dwc:scientificName` term.

For detailed instructions on mapping vector data to the DwC standard please refer to §2.1, §2.2 and §2.3.

1.4.3. Publishing the dataset with the IPT

A dataset is ready to be uploaded in the IPT, the Integrated GBIF's data publishing tool, when it is structured and formatted to Darwin Core. Please check the IPT user manual for more information. In the IPT, the user will upload the file, it is possible to map the terms to DwC if this step was not done previously. Next the user will fill out the metadata (see metadata requirements in §1.4.4), and it is in this section that the user will be requested to select a Creative Commons licence. There are three types of licences available for the resources:

- **CC0 1.0**: data is available for any use without any restrictions
- **CC BY 4.0**: data is available for any use with appropriate attribution
- **CC BY-NC 4.0**: data is available for any non-commercial use with appropriate attribution

After filling out the metadata information, the user will need to make the dataset public, publish it, and register it in GBIF.

After publication, a Darwin Core Archive (DwC-A) will be generated, this is a zipped archive consisting of one or more files of data, an XML file (`meta.xml`) describing the contents of the text files and how they relate to each other, and an XML file (`eml.xml`) containing the metadata in EML about the dataset.

Once the dataset is published, GBIF provides selected metrics about the datasets, including user download activity and cited reuses in published research and policy.

1.4.4. Metadata

Metadata is "data that provides information about other data" that contains descriptive information about a dataset and helps users discover relevant information and resources. Metadata should inform users on how to access the data, understand its fitness-for-use, and it will provide information about the creator(s), permissions, public licensing, and when and how it was created.

Different metadata standards exist, and the GBIF community uses the [Ecological Metadata Language standard](#) (EML) to record information about datasets using XML document types.

Entering the required information in the metadata section of the IPT metadata editor generates a metadata file that is included in the DwC-A file. This is an [example of a XML metadata file](#).

The IPT presents 12 different metadata forms, but some—such as associated parties, collection data, external links, additional metadata—are not required. See the terms required in the IPT metadata editor below along with examples.

Table 2. Metadata terms based in the GBIF [EML schema](#) with vector-related examples and status based on the [metadata fields required by GBIF](#), with examples from the following dataset: [Mosquitoes \(Diptera: Culicidae\) Distribution in Thailand](#)

Section in the IPT	EML term	Definition	Example	Status
Basic Metadata	title	Title of the dataset	Mosquitoes (Diptera: Culicidae) Distribution in Thailand	Required
	metadataLanguage	-	English	
	type	Please select dataset type from drop-down menu	Sampling Event	
	organizationName	Organization name responsible for the vector collection. The organization must be previously registered with GBIF. Corresponds to dwc:publishingOrganization in the IPT.	Walailak University	Required
	dataLanguage	-	English	
	maintenanceUpdateFrequency	Choose from the menu or leave unknown. Corresponds to dwc:updateFrequency in the IPT.	Unknown, Continually, Irregular etc.	
	licensed	Choose from three types: recommendation is to choose a licence that is as open as possible and only as closed as necessary. Corresponds to dwc:dataLicense in the IPT. See GBIF Terms of use: Data licensing	Creative Commons Attribution (CC BY 4.0)	Required
	abstract	A brief overview of the resource that is being documented. Corresponds to dwc:description in the IPT.	This dataset combines the occurrence records of mosquitoes from various provinces across Thailand.	Required

Section in the IPT	EML term	Definition	Example	Status
Resource Contacts	contact	The list of contacts represents the people and organizations that should be contacted to get more information about the resource, that curate the resource or to whom putative problems with the resource or its data should be addressed. Corresponds to "resourceContact(s)" in the IPT. Required fields in the IPT: last name, position, organization	Last name Sukkanon Position: Assistant Professor Organization: Walailak University	Required
Resource creators	creator	The list of creators represents the people and organizations who created the resource, in priority order. The list will be used to auto-generate the citation (if auto-generation is turned on). Corresponds to resourceCreator(s) in the IPT. Required fields in the IPT: Last name, Position, Organization, Email	Last name Chareonviriyaphap Position: Professor Organization: Kasetsart University	Required
Metadata providers	-	The list of metadata providers represents the people and organizations responsible for producing the resource metadata.+ It is the metadataProvider(s) in the IPT.+ Required fields in the IPT: Last name, Position, Organization, Email	Last name Sukkanon Position: Assistant Professor Organization: Walailak University	Recommended
Geographic Coverage	coverage	A brief description of geographical coverage. Corresponds to geographicCoverage in the IPT.	Mosquitoes from Thailand.	Required

Section in the IPT	EML term	Definition	Example	Status
Keywords (3-5)	-	In the IPT: Keyword list[3]. To improve discoverability we suggest making use of an extensive list. Sources for metadata descriptors for health include DeCS/MeSH, Global Index Medicus and Unified Medical Language System (UMLS)	infectious diseases, One Health, vectors, disease names (e.g. West Nile Virus, dengue, cutaneous leishmaniasis)	Recommended
Project Data	project	Metadata about the project that generated the dataset. Corresponds to projectData in the IPT. Please provide at least the title of the project. Additional fields for identifier, description, funding, study area description or design description are available.	Mosquitoes (Diptera: Culicidae) Distribution in Thailand	Required
Sampling Methods	samplingDescription	Description of the sampling procedures used in the research project. The content of this element would be similar to a description of sampling procedures found in the method section of a journal article. It includes study extent, sampling description and step description. Corresponds to samplingMethods in the IPT.	Mosquito collection has been carried out across Thailand from 2007 to 2023... Immature stages of mosquito were collected...using dipping technique...light-emitting diodes (LEDs) (blue, green, yellow, and red) and 2 fluorescent (ultraviolet [UV] and white) lights were used in the light traps for collecting mosquitoes in urban Bangkok... Traps were operated simultaneously (18:00 to 06:00 h)...over 36 collection nights, 6 replications were conducted for each location with a total of 216 trap-nights. ...the collected mosquitoes were initially identified using well-established morphological keys.[4]	Required

2. Data categorization

GBIF supports the publication of four classes of datasets: resource metadata, checklist, **occurrence** and **[sampling event]**. Figure 4 provides a decision tree to help find the most suitable dataset class based on a minimum of attributes associated with the data.

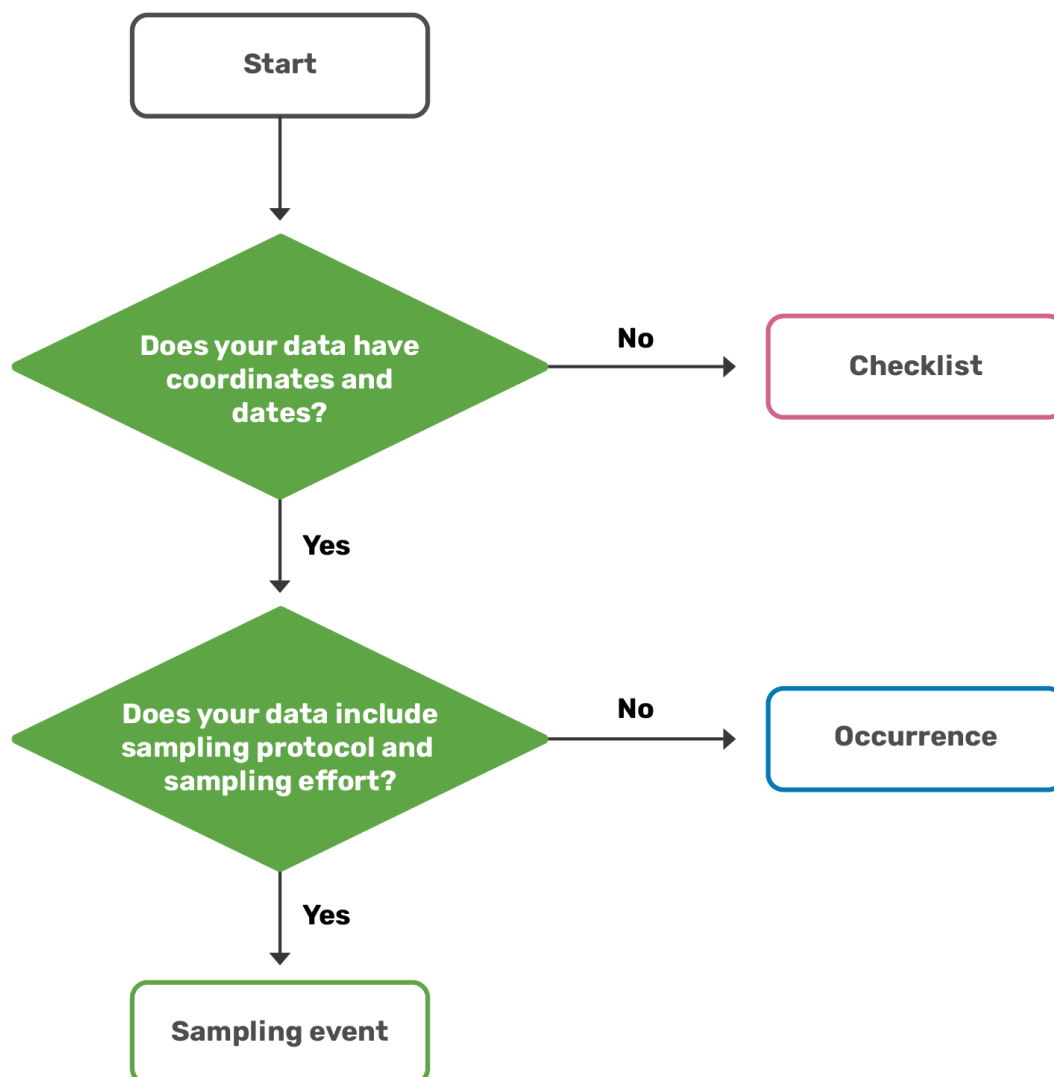


Figure 4. A decision tree for vector data categorization. Datasets can be published as checklists, lists of names that may have occurrences attached to them, better expressed in extensions such as the species profile and species distribution. Occurrence and sampling-event datasets may have additional environmental, biological or epidemiological data that can be published in extensions such as the measurement or fact or the extended measurement or fact or data on pathogens and **hosts** best captured by the **Resource Relationship extension**. And the Humboldt Extension can be used to convey more detailed information on ecological inventories, much similar to vector surveillance studies.

Before publishing a dataset in GBIF, the data has to be arranged in structured tables, these tables can be cores only or cores plus the use of extensions, which means that there are different options for sharing data through GBIF. Extensions are designed to accommodate types of data that do not fit a

particular core.

There has to be always a core table (Occurrence Core or Event Core) that can be published on its own or it can have several extensions. The decision on how to publish the data lies with the data holder to best choose how to represent their data best. The **Occurrence Core** was the first to be created, but there was a need to better represent data from surveys, and the **Event Core** was created. But also, there was an increasing need to represent data associated with occurrences and **extensions** started to be developed.

The simplest way to share data to GBIF is to use the **Occurrence Core** with no extensions, and the terms and examples are shown in section **Table 4**. An Occurrence Core will have observations and/or specimen records without information on sampling methods. However, vector data mostly fall into the sampling-event category, as it is often obtained in the context of vector surveillance and/or monitoring for epidemiological purposes or vector control activities. In these cases, field collection consists of planned sampling events (trapping events) that are focused on capturing a particular vector group. Example: mosquito sampling event that focuses on the collection of *Anopheles gambiae* s.l. or sand-fly sampling event carried out in rural properties targeting possible **vectors** of *Leishmania* (**Figure 5**).

The Occurrence Core can be used with the following extensions: **DNA derived**, **Measurement or Fact**, and **Resource Relationship** extensions. For example, a dataset of *Aedes* mosquitoes will have information on the taxonomy, spatial data, identification, etc., and would have data from molecular assays for mosquito DNA barcode and identify Dengue virus that can be displayed in the **DNA-derived extension**.

The **Taxon Core** (checklist) can be used with the Species Profile and Species Distribution extensions.

And the Event Core can be used with the following extensions: **Occurrence**, **Extended Measurement or Fact**, **Humboldt**, and the **Resource Relationship**. For example, a dataset with monitoring data on *Anopheles* mosquitoes with additional environmental data might have been collected, i.e. air temperature, air humidity, satellite data on vegetation cover and these data can be shown by the **eMOF extension**.

The Extended Measurement or Fact (eMoF) extension can be used for an Event Core with an Occurrence extension. The eMoF extension will allow measurements for the **events** (temperature, air humidity, vegetation cover, etc.) and the measurements associated with the occurrences (length of appendices, ratio between body parts, etc.) to be published together.

The Measurement or Facts extension allows only measurements of the occurrence.

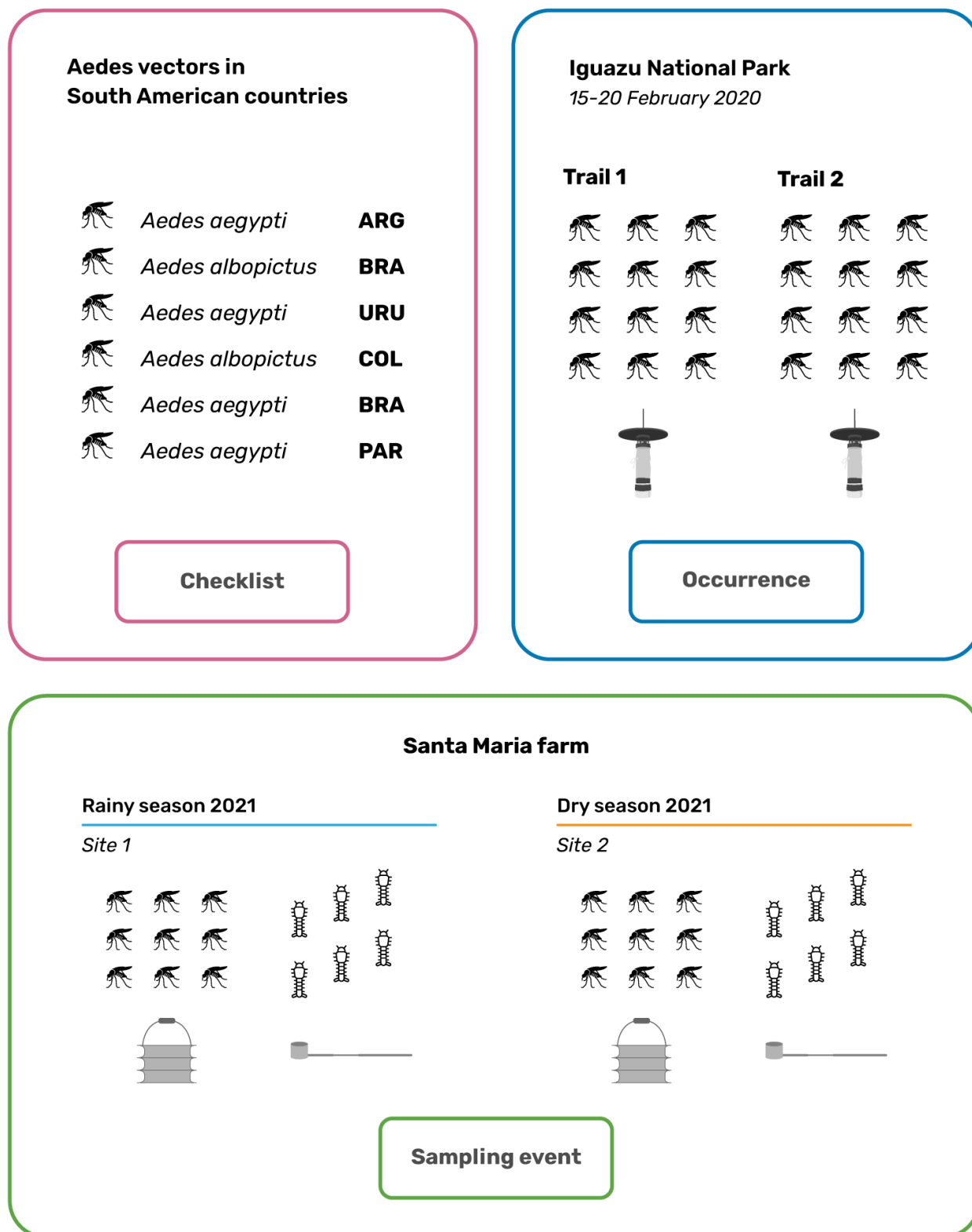


Figure 5. Examples of three dataset classes: Checklist (upper left): a list of *Aedes* species from the countries of South America; Occurrence (upper right): *Aedes* mosquitoes collected at the Iguazu National Park, Brazil, between 15-20 February 2020 at two different sites (trails 1 and 2); **Sampling event bottom**): *Aedes* mosquitoes collected in the Santa Maria Farm monthly during the rainy and dry season of 2021 using two different sampling methods for adults (BG Sentinel trap) and larvae (standard dipper).

For the purpose of this guide, we present two ways of data publication in which all available data can be published in a single file as well as a more elaborate way with the use of extensions to accommodate the diversity of variables recorded along with vector records. We provide a comprehensive list of potentially relevant terms for vector data and suggest using these terms to

improve consistency among datasets. We also provide an [Excel spreadsheet](#) whose sheets provide templates for the occurrence datasets, sampling-event datasets and the extensions discussed in this guide.

The next section provide mapping for both [sampling-event datasets](#) and an [occurrence datasets](#).



Most terms are very similar for both types of datasets, so consider repeating the same terms from the sampling-event mapping as we present only the terms that are applicable to an occurrence dataset in the appropriate section.

2.1. Vector data mapping: single file

2.1.1. Mapping sampling events

This section provides mapping recommendations for sampling-event datasets in which all available data can be published as a single file.

If geographic coordinates are not provided in decimal lat-long, the following terms can be used `dwc:verbatimLatitude`, `dwc:verbatimLongitude`, and `dwc:verbatimCoordinateSystem`.

Table 3. Recommended terms for [Event Core](#) for vector data, with examples from the following datasets: [Registros de los dípteros causantes de la transmisión del agente etiológico del dengue en el departamento del Cauca, Colombia](#); [Species composition and distribution of anopheles gambiae complex circulating in Kinshasa](#); [Data from: Rodent trapping studies as an overlooked information source for understanding endemic and novel zoonotic spillover](#); [Datos de ocurrencia de triatominos americanos del Laboratorio de Triatominos del CEPAVE \(CONICET-UNLP\)](#)

Term name	Definition	Examples	Status
<code>dwc:eventID</code>	An identifier for the set of information associated with an event (something that occurs at a place and time). May be a global unique identifier or an identifier specific to the dataset.	E-AL-Tr-1-200518	Required
<code>dwc:eventDate</code>	It is the date-time when the <code>dwc:Event</code> was recorded. Recommended best practice is to use a date that conforms to ISO 8601-1:2019.	2016-10-02	Required
<code>dwc:samplingProtocol</code>	Sampling method	Human landing catches	Required
<code>dwc:samplingSizeValue</code>	Sample size, N, #, No.	32	Required
<code>dwc:samplingSizeUnit</code>	The unit of measurement of the size (time duration, length, area, or volume) of a sample in a sampling <code>dwc:Event</code> .	hour	Required

Term name	Definition	Examples	Status
dwc:samplingEffort	The amount of effort expended during a dwc:Event.	1 Night of Human Landing Catch indoors and outdoors (2 x 16 observer-hours)	Strongly recommended
dwc:scientificName	The full scientific name, with authorship and date information if known. This term should not contain identification qualifications, which should instead be supplied in the dwc:identificationQualifier term. It also should not contain the scope of a taxon when it has been used to define more than one set of lower-level taxons, such as species complexes, sibling species, i.e. sensu lato, s.l., etc.	Anopheles (Cellia) funestus Giles, 1900	Required
dwc:higherClassification	A list (concatenated and separated) of taxa names terminating at the rank immediately superior to the referenced dwc:Taxon. Recommended best practice is to separate the values in a list with space-pipe-space), with terms in order from the highest taxonomic rank to the lowest.	Animalia Insecta Hemiptera Reduviidae Triatoma	Share if available
dwc:parentEventID	An identifier for the broader dwc:event that groups this and potentially other dwc:Events. Use a globally unique identifier for a dwc:event or an identifier for a dwc:event that is specific to the dataset.	E-AL-Tr-1	Strongly recommended
dwc:recordedBy	Person or people who recorded the original occurrence.	Erika Santamaría Harold Fernández	Share if available

Term name	Definition	Examples	Status
dwc:recordedByID	ORCID iD of person/people that recorded the original occurrence	https://orcid.org/0000-0002-7232-718X	Share if available
dwc:individualCount	The number of individuals present at the time of the dwc:occurrence.	1	Strongly recommended
dwc:organismQuantity	A number or enumeration value for the quantity of dwc:organism.	10	Share if available
dwc:organismQuantityType	The type of quantification system used for the quantity of dwc:organism.	individuals	Share if available
dwc:sex	The sex of the organism. Recommended best practice is to use a controlled vocabulary: male, female.	female	Share if available
dwc:lifeStage	The age class or life stage of the dwc:Organism(s) at the time the dwc:Occurrence was recorded. Recommended best practice is to use a controlled vocabulary: egg, larva, nymph, pupa, adult.	adult	Share if available
dwc:reproductiveCondition	The reproductive condition of the biological individual(s) represented in the occurrence. Comments or notes about the organism. Data on parity (nulliparous, parous females); stages of the gonotrophic cycle (unfed, fully fed, semi-gravid, gravid); fecundity (number of eggs laid per batch).	unfed	Share if available

Term name	Definition	Examples	Status
dwc:degreeOfEstablishment	The degree to which a organism survives, reproduces, and expands its range at the given place and time. Recommended best practice is to use controlled value strings from the controlled vocabulary designated for use with this term, listed at http://rs.tdwg.org/dwc/doc/doe/ For details, refer to https://doi.org/10.3897/biss.3.38084	invasive	Share if available
dwc:associatedTaxa	A list of identifiers or names of the record and the associations of this occurrence to each of them. The ResourceRelationship extension can alternatively be used. This term should not be used to establish relationships between records, only between the specific occurrences with other taxon.	host:Homo sapiens	Share if available
dwc:occurrenceRemarks	Comments or notes about the dwc:occurrence.	Abdomen missing Female taking in blood meal Vouchered: Registered Collection found dead	Share if available
dwc:habitat	A category or description of the habitat in which the dwc:Event occurred. Can include outdoor/indoor collection, urban/rural environments. Recommended practice is to use ENVO environmental ontology.	rural environment outdoor plastic container	Strongly recommended
dwc:eventRemarks	Comments or notes about the dwc:event.	Rain TrapStatus:Valid	Share if available

Term name	Definition	Examples	Status
dwc:countryCode	The standard code for the country in which the dcterms:Location occurs. Recommended best practice is to use an ISO 3166-1-alpha-2 country code.	TH	Strongly recommended
dwc:stateProvince	The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the dcterms:Location occurs.	Kinshasa	Share if available
dwc:locality	The specific description of the place.	Vallée de la Funa	Strongly recommended
dwc:locationID	An identifier for the set of dcterms:Location information. May be a global unique identifier or an identifier specific to the dataset.	BG323	Strongly recommended
dwc:verbatimLocality	The original textual description of the place.	Farm located in 265 km of Transnational Road	Share if available
dwc:identifiedBy	Person or people who identified the organism	Victoire Nsabatien Nsongtsa	Share if available
dwc:identifiedByID	ORCID iD of person or people who identified the organism	https://orcid.org/0000-0002-0750-6106	Share if available
dwc:IdentificationRemarks	Comments or notes about the dwc:Identification.	It was identified by the presence of a pair of white longitudinal submedial lyres on the scutum, the clypeus with white scales (females), the separate scales on the mesanepimeron and the tarsomeres of the third leg or hind leg with white basal scales.	Share if available

Term name	Definition	Examples	Status
dwc:identificationReferences	A list (concatenated and separated) of references (publication, global unique identifier, URI) used in the dwc:Identification.	Gillies M, Meillon D. 1968. The Anophelinae of Africa south of the Sahara. Publication of the South African Institute for Medical Research, Johannesburg, 54, 1–343. Gillies MT. 1987. A supplement to the Anophelinae of Africa south of the Sahara (Afrotropical Region). Publications of the South African Institute for Medical Research, 55, 1–143	Share if available
dwc:taxonRank	The taxonomic rank of the most specific name in the dwc:scientificName. Recommended best practice is to use a controlled vocabulary. The taxon ranks of algae, fungi and plants are defined in the International Code of Nomenclature for algae, fungi, and plants (Schenzhen Code Articles H3.2, H4.4 and H.3.1).	species	Share if available
dwc:taxonRemarks	Comments or notes about the taxon or name.	Previously mentioned as <i>Triatoma lecticularia</i>	Share if available
dwc:bibliographicCitation	A bibliographic reference for the resource.	Diatta, G., Duplantier, JM, Granjon, L., Ba, K., Chauvancy, G., Ndiaye, M., Trape, JF. Borrelia infection in small mammals in West Africa and its relationship with tick occurrence inside burrows. Acta Trop 152 131–140. (2015). DOI/ISSN/ISBN: 10.1016/j.actatropica.2015.08.016	Share if available

Term name	Definition	Examples	Status
dwc:decimalLatitude	The geographic latitude (in decimal degrees, using the spatial reference system given in dwc:geodeticDatum) of the geographic center of a dcterms:Location . Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.	9.236691273	Strongly recommended
dwc:decimalLongitude	The geographic longitude (in decimal degrees, using the spatial reference system given in dwc:geodeticDatum) of the geographic center of a dcterms:Location . Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.	-5.668395403	Strongly recommended
dwc:geodeticDatum	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in dwc:decimalLatitude and dwc:decimalLongitude are based. Recommended best practice is to use the EPSG code of the SRS, if known. Otherwise use a controlled vocabulary for the name or code of the geodetic datum, if known. If none of these is known, use the value unknown.	WGS84	Strongly recommended

Term name	Definition	Examples	Status
dwc:coordinateUncertaintyInMeters	The horizontal distance (in meters) from the given dwc:decimalLatitude and dwc:decimalLongitude describing the smallest circle containing the whole of the dterms:Location. Leave the value empty if the uncertainty is unknown, cannot be estimated, or is not applicable (because there are no coordinates). Zero is not a valid value for this term.	100	Strongly recommended
dwc:occurrenceStatus	Presence, absence information	present	Strongly recommended
dwc:verbatimIdentification	A string representing the taxonomic identification as it appeared in the original record.	Anopheles gambiae complex	Strongly recommended
dwc:verbatimTaxonRank	The taxonomic rank of the most specific name in the dwc:scientificName as it appears in the original record.	s.l.	Share if available
dwc:dynamicProperties	A list of additional measurements, facts, characteristics, or assertions about the record. Meant to provide a mechanism for structured content. Recommended best practice is to use a key:value encoding schema for a data interchange format such as JSON.	"TemperatureInCelsius": "26.5", "relativeHumidity": "64.2"	Share if available

2.1.2. Mapping occurrences

This section provides mapping recommendations for occurrence datasets in which all available data can be published as a single file.

We recommend using the following terms presented in Table 3: dwc:eventDate, dwc:scientificName, dwc:recordedBy#, hdwc:recordedByID, dwc:individualCount, dwc:organismQuantity, dwc:organismQuantityType, dwc:sex, dwc:lifeStage#, dwc:reproductiveCondition, dwc:degreeOfEstablishment, dwc:occurrenceStatus, dwc:associatedTaxa, dwc:occurrenceRemarks, dwc:eventID, dwc:habitat, dwc:samplingProtocol, dwc:samplingSizeValue, dwc:samplingSizeUnit, dwc:samplingEffort, dwc:eventRemarks, dwc:countryCode, dwc:stateProvince, dwc:locality, dwc:verbatimLocality, dwc:decimalLatitude, dwc:decimalLongitude, dwc:geodeticDatum, dwc:identifiedBy, dwc:identifiedByID, dwc:identificationReferences, dwc:IdentificationRemarks, dwc:taxonRank, dwc:taxonRemarks, dwc:verbatimIdentification, dwc:verbatimTaxonRank, dwc:bibliographicCitation, dwc:dynamicProperties.

For detailed explanation and examples for the above terms, please refer to Table 3 for sampling event.

Table 4. Recommended term for Occurrence Core for vector data, with examples from the following datasets: Registros de los dípteros causantes de la transmisión del agente etiológico del dengue en el departamento del Cauca, Colombia; Species composition and distribution of anopheles gambiae complex circulating in Kinshasa; Data from: Rodent trapping studies as an overlooked information source for understanding endemic and novel zoonotic spillover; Datos de ocurrencia de triatominos americanos del Laboratorio de Triatominos del CEPAVE (CONICET-UNLP)

Term name	Definition	Examples	Status
dwc:basisOfRecord	The specific nature of the data record. For field collected organisms use HumanObservation, for specimens deposited in biological collections/museums use PreservedSpecimen. For data abstracted from the literature use MaterialCitation. For DNA-derived occurrences, tissue/blood samples use MaterialSample. And for organisms from laboratory colonies use LivingSpecimen.	HumanObservation	Required
dwc:occurrenceID	An identifier for the dwc:Occurrence (as opposed to a particular digital record of the dwc:Occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the dwc:occurrenceID globally unique.	308-Anopheles_Collection_308-CD	Required
dwc:preparations	A preparation or preservation method for a specimen.	Placed in RNA Later Pinned	Share if available
dwc:associatedSequences	A list (concatenated and separated) of identifiers (publication, global unique identifier, URI) of genetic sequence information associated with the dwc:materialEntity.	https://www.ncbi.nlm.nih.gov/nuccore/MK918898	Share if available
dwc:kingdom	The full scientific name of the kingdom in which the dwc:Taxon is classified.	Animalia	Strongly recommended

Term name	Definition	Examples	Status
dwc:genus	The full scientific name of the genus in which the dwc:Taxon is classified.	<i>Triatoma</i>	Share if available
dwc:specificEpithet	The name of the first or species epithet of the dwc:scientificName.	<i>gerstaeckeri</i>	Share if available

2.2. Vector data mapping using extensions

This section provides mapping recommendations for the use of extensions with either the Occurrence Core, the Event Core or the Taxon Core (checklist).

GBIF has a list of registered [extensions](#) and [vocabulary](#) that can be useful in the standardization of terms, but we also suggest checking the [controlled vocabulary & ontologies](#) section for more specific information.

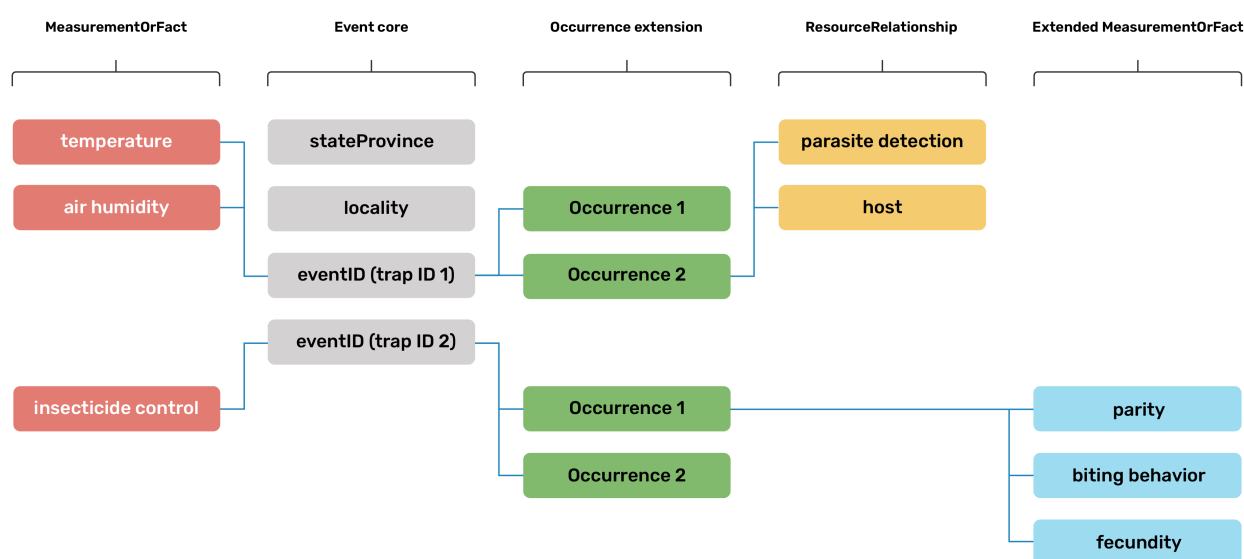


Figure 6. Event Core diagram with suggested use of extensions for vector data. Abiotic measurements collected during the sampling event, with occurrences linked to sampling events using the eventID (full lines). Biotic measurements are linked to occurrences using the occurrenceID term of the ExtendedMeasurementOrFact Extension (dashed lines). Figure adapted from the OBIS Manual.

2.2.1. Extended Measurement or Facts Extension

The [Extended Measurement or Fact extension](#) (eMoF) supports the publication of generic measurements or facts linking to occurrences. This extension was developed to be used in combination with the Event Core, but is also compatible with other cores.

Table 5. Recommended terms from the Extended Measurement or Fact extension, with examples from the following dataset: [A minimum data standard for wildlife disease research and surveillance](#)

Term name	Definition	Examples	Status
dwc:eventID	An identifier for the set of information associated with an event (something that occurs at a place and time). May be a global unique identifier or an identifier specific to the dataset.	OS BZ1995	Required
occurrenceID	A unique identifier for the occurrence, it is recommended to construct one from a combination of identifiers in the record that will most closely make the dwc:occurrenceID globally unique. This term allows that the same occurrence can be recognized in different versions of a dataset.	BZ19114	Required
dwc:measurementAccuracy	The description of the potential error associated with the measurement Value. See also terms:dwc[dwc:measurementAccuracy]	0.01	Share if available

Term name	Definition	Examples	Status
dwc:measurementDeterminedBy	A list (concatenated and separated) of names of people, groups, or organizations who determined the value of the MeasurementOrFact. See also dwc:measurementDeterminedBy	Javier de la Torre	Share if available
dwc:measurementDeterminedDate	The date on which the MeasurementOrFact was made. Recommended best practice is to use an encoding scheme, such as ISO 8601:2004(E). See also dwc:measurementDeterminedDate	2020-07-15	Share if available
dwc:measurementID	An identifier for the MeasurementOrFact (information pertaining to measurements, facts, characteristics, or assertions). May be a global unique identifier or an identifier specific to the dataset. See also dwc:measurementID	TP205-A	Share if available

Term name	Definition	Examples	Status
dwc:measurementMethod	A description of or reference to (publication, URI) the method or protocol used to determine the measurement, fact, characteristic, or assertion.	qPCR	Share if available
dwc:measurementRemarks	Comments or notes accompanying the MeasurementOrFact	tip of tail missing	Share if available
dwc:measurementType	The nature of the measurement, fact, characteristic, or assertion. Recommended best practice is to use a controlled vocabulary. See also dwc:measurementType	wing length	Share if available
measurementTypeID	An identifier for the measurementType (global unique identifier, URI). The identifier should reference the measurementType in a vocabulary.	http://vocab.nerc.ac.uk/collection/P01/current/ODRYBM01	Share if available

Term name	Definition	Examples	Status
dwc:measurementUnit	The units associated with the measurement Value. Recommended best practice is to use the International System of Units (SI). See also dwc:measurementUnit	Ct	Share if available
measurementUnitID	An identifier for the measurement Unit (global unique identifier, URI). The identifier should reference the measurement Unit in a vocabulary.	http://vocab.nerc.ac.uk/collection/P06/current/ULCM	Share if available
dwc:measurementValue	The value of the measurement, fact, characteristic, or assertion. See also dwc:measurementValue	12	Share if available

Term name	Definition	Examples	Status
measurementValueID	An identifier for facts stored in the column measurementValue (global unique identifier, URI). This identifier can reference a controlled vocabulary (e.g. for sampling instrument names, methodologies, life stages) or reference a methodology paper with a DOI. When the measurementValue refers to a value and not to a fact, the measurementValueID has no meaning and should remain empty.	Standard Sherman Live Trap	Share if available

2.2.2. Measurement or Facts Extension

The **Measurement or Fact extension** (MoF) provides extended support for multiple measurements or facts associated with a Darwin Core Occurrence, Event, or Taxon Core dataset. *Note:* The recommendation for each of the terms in **Table 6** is **Share if available**.

*Table 6. Recommended terms from the Measurement or Fact extension, with examplee from the following dataset: **Anopheles collections in the health districts of Korhogo (Côte d'Ivoire) and Diébougou (Burkina Faso) (2016–2018)***

Term name	Definition	Examples
dwc:measurementID	An identifier for the dwc:MeasurementOrFact (information pertaining to measurements, facts, characteristics, or assertions). May be a global unique identifier or an identifier specific to the dataset.	1BAP1_16

Term name	Definition	Examples
dwc:measurementType	The nature of the measurement, fact, characteristic, or assertion. Recommended best practice is to use a controlled vocabulary. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	Daytime land surface temperature 1 weeks preceding the event
dwc:measurementValue	The value of the measurement, fact, characteristic, or assertion. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	29.35904105
dwc:measurementAccuracy	The description of the potential error associated with the dwc:measurementValue.	1-km resolution
dwc:measurementUnit	The units associated with the dwc:measurementValue. Recommended best practice is to use the International System of Units (SI). This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	degree Celsius
dwc:measurementDeterminedBy	A list (concatenated and separated) of names of people, groups, or organizations who determined the value of the dwc:MeasurementOrFact. Recommended best practice is to separate the values in a list with the post character, . This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	Paul Taconet
dwc:measurementDeterminedDate	The date on which the dwc:measurementOrFact was made. Recommended best practice is to use a date that conforms to ISO 8601-1:2019.	2019-02-01
dwc:measurementMethod	A description of or reference to (publication, URI) the method or protocol used to determine the measurement, fact, characteristic, or assertion. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	Derived from satellite products (detailed method: https://doi.org/10.1186/s13071-021-04851-x)
dwc:measurementRemarks	Comments or notes accompanying the dwc:MeasurementOrFact.	Meteorology preceding the event

2.2.3. Occurrence Extension

This extension uses the same terms as the Occurrence Core (see [Table 4](#)).

2.2.4. Humboldt Extension for Ecological Inventories

The [Humboldt Extension for Ecological Inventories](#) provides support for `dwc:Events` related to ecological inventories. *Note:* The recommendation for each of the terms in [Table 7](#) is **Share if available**.

NOTE: For guidance on how to include the Humboldt Extension in a sampling-event dataset, see [Ingenloff \(2025\)](#).

Table 7. Recommended terms from the Humboldt Extension, drawn from [Ingenloff \(2025\)](#).

Term name	Definition	Examples
<code>eco:inventoryTypes</code>	The type(s) of search processes used to conduct the inventory.	<code>restrictedSearch</code>
<code>eco:protocolNames</code>	Categorical descriptive names for the methods used during the <code>dwc:Event</code> . Recommended best practice is to use a controlled vocabulary and separate multiple values in a list with <code> </code> . This term has an equivalent in the <code>dwciri:</code> namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	<code>vectorSurveillance</code>
<code>eco:protocolDescriptions</code>	Detailed description of methods used during the <code>dwc:event</code>	1 light trap installed in forest and animal shelter each. They operated from 6:00pm-6:00am.
<code>eco:samplingPerformedBy</code>	A person, group, or organization responsible for recording the <code>dwc:Event</code> .	Instituto de Salud Carlos III - ISCIII
<code>eco:isSamplingEffortReported</code>	The sampling effort associated with the <code>dwc:event</code> was reported. Typically values of effort would be captured under the terms <code>eco:samplingEffortValue</code> and <code>eco:samplingEffortUnit</code> .	TRUE
<code>eco:samplingEffortValue</code>	The sampling effort associated with the <code>dwc:Event</code> was reported. Typically values of effort would be captured under the terms <code>eco:samplingEffortValue</code> and <code>eco:samplingEffortUnit</code> .	30

Term name	Definition	Examples
eco:samplingEffortUnit	The units associated with the eco:samplingEffortValue. Recommended best practice is to use a controlled vocabulary. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	trapNight
eco:samplingEffortProtocol	A description of or reference (publication or URL) to the methods used to determine the sampling effort. This description should be associated with the values reported in eco:samplingEffortValue and eco:samplingEffortUnit. This is a specialization of eco:protocolDescription focused on effort, distinct from the survey method. The effort relates to the intensity of sampling and therefore can assist in interpreting estimates of completeness. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	12 hours/night
eco:totalAreaSampledValue	The numeric value for the total area surveyed during the dwc:Event. This area is always less than or equal to the term:eco[eco:geospatialScopeAreaValue. An eco:totalAreaSampledValue must have a corresponding eco:totalAreaSampledUnit.	250
eco:totalAreaSampledUnit	The units associated with eco:totalAreaSampledValue. Recommended best practice is to use an IRI from a controlled vocabulary of SI units, derived units, or other non-SI units accepted for use within the SI.	square metre
eco:isAbundanceReported	The number of dwc:organism collected or observed was reported. Typically the abundance values would be reported in the dwc:organismQuantity and dwc:organismQuantityType terms for the child dwc:occurrence records for this dwc:Event.	TRUE

Term name	Definition	Examples
eco:isLeastSpecificTargetCategoryQuantityInclusive	The total detected quantity for a <code>dwc:taxon</code> (including subcategories thereof) in a <code>dwc:Event</code> is given explicitly in a single record (<code>dwc:organismQuantity</code> value) for that <code>dwc:taxon</code> . Recommended values are 'true' and 'false'. This term is only relevant if <code>dwc:organismQuantity</code> is a number. For a detailed explanation, see http://rs.tdwg.org/eco/docs/inclusive/ .	TRUE
eco:isAbundanceCapReported	A maximum number of <code>dwc:organisms</code> was reported, as specified or restricted by the protocol used. Values of abundance cap should be captured under the term <code>eco:abundanceCap</code> .	TRUE
eco:abundanceCap	A maximum number of <code>dwc:organisms</code> was reported, as specified or restricted by the protocol used. Values of abundance cap should be captured under the term <code>eco:abundanceCap</code> .	85
eco:eventDurationValue	The numeric value for the duration of the DwC Event. An <code>eco:eventDurationValue</code> must have a corresponding <code>eco:eventDurationUnit</code> .	3
eco:eventDurationUnit	The units associated with the <code>eco:eventDurationValue</code> . Recommended best practice is to use an IRI from a controlled vocabulary of SI units, derived units, or other non-SI units accepted for use within the SI.	nights
eco:siteNestingDescription	Textual description of the hierarchical sampling design, e.g. Study consists of a series of sampling events at 10 different sites. Each Event site is identified using X coding system.	2 sampling sites per habitat
eco:verbatimSiteNames	All location codes or site names included in the study area	code(s) for specific event
eco:verbatimSiteDescriptions	Original textual description of the site(s). Site refers to the location at which observations are made or samples/measurements are taken. The site can be at any level of hierarchy. Recommended best practice is to separate multiple values in a list with the post character: .	Patch of forest Barn where sheep spend nights
eco:verbatimTargetScope	The verbatim original description of the <code>dwc:event</code> scope.	Ex. All adult mosquitoes of...

Term name	Definition	Examples
eco:targetTaxonomicScope	The taxonomic group(s) targeted for sampling during the dwc:event.	<i>Ixodidae</i>
eco:targetLifeStageScope	The age classes or life stages of the dwc:organisms targeted for sampling during the dwc:event.	Nymph adult

2.2.5. Resource Relationship

The **Resource Relationship extension** provides support for relationships between resources in a Darwin Core Occurrence, Event, or Taxon Core to resources in an extension or external to the dataset. The identifiers for subject (resourceID) and object (relatedResourceID) may exist in the dataset or be accessible via an externally resolvable identifier.

Relationships can be one-way as of between a **vector** and the pathogens detected by a molecular assay, or it can be two-way, that is, between a vertebrate **host**, a vector and the pathogen detected in the vector. We provide examples below of a one-way relationship between ticks and pathogens and a two-way relationship between ticks, vertebrate hosts and pathogens. *Note:* The recommendation for each of the terms in **Table 8** and **Table 9** is **Share if available**.

Table 8. Recommended terms from the ResourceRelationship Extension for one-way relationships involving ticks and pathogens, with examples from an as-yet unpublished dataset from the One Health VBD Hub, "Tick Occurrences in the UK—Review of Literature."

Field name	Definition	Examples
dwc:ResourceRelationship	A relationship of one rdfs:Resource to another. Resources can be thought of as identifiable records or instances of classes and may include, but need not be limited to instances of dwc:occurrence, dwc:organism, dwc:materialEntity, dwc:event, dcterms:Location , dwc:geologicalContext, dwc:identification, or dwc:taxon.	an instance of a dwc:Organism is the vector of another instance of a term:dwc
dwc:resourceRelationshipID	An identifier for an instance of relationship between one resource (the subject) and another (dwc:relatedResource, the object).	ICL:UK:831:ICL:UK:1:1075
dwc:resourceID	An identifier for the resource that is the subject of the relationship.	ICL:UK:831
dwc:relationshipOfResourceID	An identifier for the relationship type (predicate) that connects the subject identified by dwc:resourceID to its object identified by dwc:relatedResourceID. Recommended best practice is to use the identifiers of the terms in a controlled vocabulary, such as the OBO Relation Ontology.	[.break-all]# http://purl.obolibrary.org/obo/RO_0002459

Field name	Definition	Examples
dwc:relatedResourceID	An identifier for a related resource (the object, rather than the subject of the relationship).	ICL:UK:1:1075
dwc:relationshipOfResource	The relationship of the subject (identified by dwc:resourceID) to the object (identified by dwc:relatedResourceID). Recommended best practice is to use a controlled vocabulary.	is vector for
dwc:relationshipRemarks	Comments or notes about the relationship between the two resources.	pathogen 1

Table 9. Recommended terms from the ResourceRelationship extension for two-way relationships involving ticks and pathogens, with examples from the following published dataset: [VectorNet](#)

Terms name	Definition	Examples
dwc:resourceRelationshipID	An identifier for an instance of relationship between one resource (the subject) and another (dwc:relatedResource, the object).	2A
dwc:resourceID	An identifier for the resource that is the subject of the relationship.	1A
dwc:relationshipOfResourceID	An identifier for the relationship type (predicate) that connects the subject identified by dwc:resourceID to its object identified by dwc:relatedResourceID. Recommended best practice is to use the identifiers of the terms in a controlled vocabulary, such as the OBO Relation Ontology.	http://purl.obolibrary.org/obo/RO_0002459
dwc:relatedResourceID	An identifier for a related resource (the object, rather than the subject of the relationship).	1A:2A

Terms name	Definition	Examples
dwc:relationshipOfResource	The relationship of the subject (identified by dwc:resourceID) to the object (identified by dwc:relatedResourceID). Recommended best practice is to use a controlled vocabulary.	is vector for reservoir host of pathogen of
dwc:relationshipRemarks	Comments or notes about the relationship between the two resources.	pathogen 1

Figure 7a and 7b below presents some examples of how resourceRelationship tables from the Occurrence Core may look.

id	basisOfRecord	occurrenceID	individualCount	lifeStage	associatedTaxa	eventDate
21-VECT-000475-A	HumanObservation	21-VECT-000475-A	10	Adults	host: "Canis canis"	23/12/2003
2A	HumanObservation	2A	3		host: "Hyalomma marginatum"	12/02/2004
3A	HumanObservation	3A	2		host: "Hyalomma marginatum"	12/02/2004

samplingProtocol	sampleSizeValue	locationID	decimalLatitude	decimalLongitude	verbatimIdentification	scientificName	family
detection on animals		CH011	46.674	6.570672		Hyalomma marginatum	Ixodidae
PCR	8				CCHFV	Crimean-Congo hemorrhagic fever orthonairovirus	Nairoviridae
PCR	6				TBEV	Tick-borne encephalitis virus	Flaviviridae

Figure 7. Example of DwC mapping to the original headers in an Occurrence Core table (above) with a two-way relationship between vertebrate hosts, a tick and the pathogen detected in the vector (below).

resourceRelationshipID	resourceID	relatedResourceID	relationshipOfResource	relationshipOfResourceID
21-VECT-000475-A	21-VECT-000475-A	2A:3A	vector hosting	http://purl.obolibrary.org/obo/RO_0002459
2A	2A	2-VECT-000475-A	pathogen found on	http://purl.obolibrary.org/obo/APOLLO_SV_00000237
3A	3A	2-VECT-000475-A	pathogen found on	http://purl.obolibrary.org/obo/APOLLO_SV_00000237

Figure 8. Example of DwC mapping to the original headers in an ResourceRelationship extension table with a two-way relationship between vertebrate hosts, a tick and the pathogen detected in the vector.

2.2.6. Species Profile

The Species Profile extension provides a basic taxonomic profile with characteristics in addition to written descriptions, which are covered by the description extension and can be used in addition to the Taxon Core checklist.

Table 10. Recommended terms from the Species Profile extension, with examples from the following datasets: NEON ticks sampled using drag cloths and tick pathogen status, Fiocruz/COLFLEB - Coleção de Flebotomíneos, Catálogo Taxonômico da Fauna do Brasil.

Term name	Definition	Examples	Status
isFreshwater	A Boolean flag indicating whether the taxon occurs in freshwater habitats, i.e. can be found in/above rivers or lakes	FALSE	Share if available

Term name	Definition	Examples	Status
isTerrestrial	A Boolean flag indicating the taxon is a terrestrial organism, i.e. occurs on land as opposed to the sea	TRUE	Share if available
isInvasive	Flag indicating a species known to be invasive/alien in some are of the world. Detailed native and introduced distribution areas can be published with the distribution extension.	FALSE	Recommended
isExtinct	Flag indicating an extinct organism. Details about the time period the organism has lived in can be supplied below	TRUE	Share if available
livingPeriod	The (geological) time a currently extinct organism is known to have lived. For geological times of fossils ideally based on a vocabulary like http://en.wikipedia.org/wiki/Geologic_column	Miocene	Share if available
ageInDays	Maximum observed age of an organism given as number of days	30	Share if available
sizeInMillimeters	Maximum observed size of an organism in millimeter. Can be either height, length or width, whichever is greater.	26	Share if available
dwc:sex	The sex of the organism. Recommended best practice is to use a controlled vocabulary: male, female.	female	Strongly recommended

Term name	Definition	Examples	Status
dwc:habitat	A category or description of the habitat in which the dwc:event occurred. Can include outdoor/indoor collection, urban/rural environments. Recommended practice is to use ENVO environmental ontology.	rural environment outdoor chicken coop	Strongly recommended
source	Source reference for this distribution record. Can be proper publication citation, a web page URL, etc.	Catálogo Taxonômico da Fauna do Brasil. Published on the Internet http://fauna.jbrj.gov.br/fauna/f aunadobrasil/55443	Share if available
dwc:datasetID	An identifier for a subset of data. See also datasetID	https://doi.org/10.48443/nygx-dm71	Strongly recommended

2.2.7. Species Distribution

The Species Distribution extension is a geographic distribution of a taxon and can be used with the Taxon Core checklist.

In addition to the terms in Table 11, we recommend using the terms: **source** and dwc:datasetID, which is described above in Table 9.

Table 11. Recommended terms from the Species profile Extension, with examples from the dataset: Anopheles collections in the health districts of Korhogo (Côte d'Ivoire) and Diébougou (Burkina Faso) (2016–2018).

Term name	Definition	Examples	Status
dwc:locationID	A code for the named area this distribution record is about. Use a prefix for each code to indicate the source of the code, see http://rs.gbif.org/areas/ for list of coding schemes and their recommended prefix. See also http://rs.gbif.org/areas/	1B0H1	Strongly recommended
dwc:locality[locality]	The verbatim name of the area this distribution record is about.	Bohéro	Strongly recommended
dwc:countryCode	ISO 3166 alpha 2 or alpha 3 country codes the area belongs to or as an alternative for a locationID if the area is a country. For multiple countries separate values with a comma ", ".	BF	Strongly recommended

Term name	Definition	Examples	Status
dwc:lifeStage	The distribution information pertains solely to a specific life stage of the taxon	adult	Share if available
dwc:occurrenceStatus	Statement about the presence or absence of the taxon in the given area.	present	Share if available

Term name	Definition	Examples	Status
dwc:establishmentMeans	Statement about whether the taxon has been introduced to the given area and time through the direct or indirect activity of modern humans. Recommended best practice is to use controlled value strings from the controlled vocabulary designated for use with this term, listed at http://rs.tdwg.org/dwc/doc/em/ . For details, refer to https://doi.org/10.3897/biss.3.38084	native	Share if available

Term name	Definition	Examples	Status
dwc:degreeOfEstablishment	<p>The degree to which the taxon survives, reproduces, and expands its range at the given area and time.</p> <p>Recommended best practice is to use controlled value strings from the controlled vocabulary designated for use with this term, listed at http://rs.tdwg.org/dwc/doc/doi/. For details, refer to https://doi.org/10.3897/biss.3.38084</p>	native	Share if available

Term name	Definition	Examples	Status
dwc:pathway	<p>The process by which the taxon came to be in the given area at the given time.</p> <p>Recommended best practice is to use controlled value strings from the controlled vocabulary designated for use with this term, listed at http://rs.tdwg.org/dwc/doc/pw/. For details, refer to https://doi.org/10.3897/biss.3.38084</p>	Ship/boat ballast water	Share if available

Term name	Definition	Examples	Status
dwc:eventDate	Relevant temporal context for this entire distribution record including all properties preferably given as a year range or single year on which the distribution record is valid. For the same area and taxon there could therefore be several records with different temporal context, e.g. in 5 year intervals for invasive species.	2017-01-27	Strongly recommended

Term name	Definition	Examples	Status
dwc:startDayOfYear	Seasonal temporal subcontext within the eventDate context. Useful for migratory species. The earliest ordinal day of the year on which the distribution record is valid. Numbering starts with 1 for 1 January and ends with 365 or 366 for 31 December.	27	Share if available
dwc:endDayOfYear	Seasonal temporal subcontext within the eventDate context. The latest ordinal day of the year on which the distribution record is valid.	29	Share if available
dwc:occurrenceRemarks	Comments or notes about the distribution.	Collected near a lake	Share if available

2.3. Specific requirements for publishing vector data

2.3.1. How to better describe species complexes/assemblages or sibling species with the DwC standard

Vector data presents some specific demands with regard to taxonomy, because in many groups,

the occurrence of species complexes and assemblages or sibling species is well-documented (WHO 2007, Garros et al. 2005, Motoki et al. 2009, Harbach 2012, Gutierrez et al. 2021, Aguilar-Vega et al. 2021, Cotes-Perdomo et al. 2023).

The DwC standard can only handle subspecies with the `dwc:infraspecificEpithet` term, but there is no appropriate term to accommodate well species complexes/assemblages or sibling species.

One way to sort out this specific issue with vector data is to leave the `dwc:scientificName` at the lowest level of identification possible, in this case, at genus level, and then display the species complexes/assemblages or sibling species status in the `dwc:verbatimIdentification` term, or even include any qualifier (such as `s.l.`, `sp.`, `cf.` or `aff.` in the `dwc:verbatimTaxonRank` term to improve the alignment with the taxonomic backbone. **Identification qualifiers** {Ed.: too loose? Isn't there a specific DwC term to borrow phrasing from here?} should not be included in the `dwc:scientificName` term.

It is important to remember that when the dataset is uploaded in GBIF, the taxon names are matched to **GBIF Taxonomic Backbone**, which is an updated list of names. However, to improve the IPT's ability to handle the names in a unambiguous way, it's important to add higher taxonomy to the data, even if just at kingdom level. This way, similar species names that are found in different kingdoms, can be classified correctly and it prevents the IPT from assigning *Incertae sedis* to the names.

2.3.2. How to best describe the native status of vectors

Understanding how vectors spread and where they come from is an important aspect of disease surveillance. The global distribution of **VBDs** is already affected by climate change, land use changes, deforestation, global trade, among other factors, causing varying degrees of impacts across regions.

There are several examples of how human activity is reshaping the distribution of vector species, as with *Aedes aegypti*, which has spread to more than 300 cities since its introduction to California in the U.S. in 1913 (Kelly et al. 2021); the introduction and spread of *Anopheles stephensi* in Africa (Sinka et al. 2020); and the reappearance of *Anopheles sacharovi* in Italy after more than 50 years, due to an increase of natural areas with favourable climatic and environmental conditions (Raele et al. 2024). These introductions, reintroductions, and spread of vector species require constant updates in their status as invasive, established or native to better inform both decision-making and the design of control strategies.

Data sources originate from a wide range of data holders, e.g. governmental surveillance and control programmes, and routine activities, research, and private businesses which provide data without any standardization. Since standardized, good quality data are necessary to inform control strategies, species distribution models, risk assessments and early warning systems, it is best practice to use the appropriate terms and controlled vocabulary. These are provided by the DwC standard (e.g. introduction `dwc:pathway` and the `dwc:degreeOfEstablishment` and a suggested controlled vocabulary has been proposed by Groom et al. (2019), see **Appendix: Table 1** for the full list of controlled vocabulary for the `dwc:degreeOf Establishment` term.

2.3.3. How to handle data that does not fit any DwC term available

The DwC term `dwc:dynamicProperties` provides a way to list additional measurements, facts, characteristics, or assertions about a record; it can be a way to provide a mechanism for structured content. And recommended best practice is to use a key:value encoding schema for a data interchange format such as JSON.

Table 12. Examples key:value statements for use as `dwc:dynamicProperties` with vector data.

Examples	Explanation
LocationCode:DEC05	For a code for location, such as the NUTS Code
LocationCode:DEC05,Host Bodypart:midgut	For displaying location code + host body parts
TemperatureInCelsius:29.8, relativeHumidity:64.2	For environmental data

2.3.4. Controlled vocabulary

Controlled vocabularies are standardized words and phrases and they provide a consistent way to organize knowledge for subsequent retrieval. In addition to the links in §1.5.3], we provide an [appendix] with DwC terms, such as [samplingProtocol](#), [dwc:lifeStage](#), [dwc:sex](#), [degreeOfEstablishment](#), [habitat](#).

2.3.5. Unique identifiers within datasets

An identifier consists of a unique identification code assigned to an object for unambiguous retrieval. Three key DwC terms that must be either global unique identifiers or a unique to within a specific are:

- [dwc:occurrenceID](#)
- [dwc:eventID](#)
- [dwc:parentEventID](#)

It is important to consider the level of granularity of the data, that is how much detail about the data the identifier will cover. If we have, for example, a surveillance program running across different states or even different countries, we will need unique identifiers for occurrences and events, so they can unambiguously be retrieved.

Another thing to consider is opacity, or how much the identifier allows us to learn anything from the format of the identifier itself.

For a general introduction to identifiers, see [White et al. 2011](#) and this [GBIF Community Forum post](#) about UUIDs (Universally Unique Identifiers) and opacity.

2.3.6. Additional recommendations on relevant terms

Here we propose additional terms that are useful in vector datasets, either by providing more detailed metadata, publication of verbatim data, more details on the classification, etc.

Table 13. Recommended additional terms of relevance to vector data, with examples from the datasets: [Registros de los dípteros causantes de la transmisión del agente etiológico del dengue en el departamento del Cauca, Colombia](#), [Species composition and distribution of anopheles gambiae complex circulating in Kinshasa](#), [Fiocruz/COLFLEB - Coleção de Flebotomíneos](#)

Term name	Definition	Examples*	Status
dwc:rightsHolder	A person or organization owning or managing rights over the resource.	University of Kinshasa	Share if available

Term name	Definition	Examples*	Status
dwc:institutionCode	The name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record.	UNIKIN	Share if available
dwc:informationWithheld	Additional information that exists, but that has not been shared in the given record.	This dataset is a simplified version of the original one. We have removed non-biological information, information that could compromise the privacy of citizen scientists and some redundant information.	Share if available
dwc:organismRemarks	Comments or notes about the dwc:Organism instance.	Nulliparous	Share if available
dwc:pathway	The process by which a dwc:Organism came to be in a given place at a given time.+ Recommended best practice is to use controlled value strings from the controlled vocabulary designated for use with this term, listed at http://rs.tdwg.org/dwc/doc/pw/ . For details, refer to https://doi.org/10.3897/biss.3.38084 . This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	container/bulk [ship/boat ballast water	Share if available

Term name	Definition	Examples*	Status
dwc:associatedOccurrences	A list (concatenated and separated) of identifiers of other dwc:Occurrence records and their associations to this dwc:Occurrence. This term can be used to provide a list of associations to other dwc:Occurrences. Note that the dwc:ResourceRelationship class is an alternative means of representing associations, and with more detail. Recommended best practice is to separate the values in a list with space vertical bar space ().	pathogen test of:2a85f793-09aa-4e49-b5fe-523b2440440f-00006
Share if available	dwc:associatedSequences	A list (concatenated and separated) of identifiers (publication, global unique identifier, URI) of genetic sequence information associated with the dwc:MaterialEntity.	http://www.ncbi.nlm.nih.gov/nuccore/U34853.1
Share if available	dwc:year	The four-digit year in which the dwc:Event occurred, according to the Common Era Calendar.	2020
Share if available	dwc:month	The integer month in which the dwc:Event occurred.	12
Share if available	dwc:day	The integer day of the month on which the dwc:Event occurred.	21
Share if available	dwc:verbatimEventDate	The verbatim original representation of the date and time information for a dwc:Event.	X-1936

Term name	Definition	Examples*	Status
Share if available	dwc:county	The full, unabbreviated name of the next smaller administrative region than stateProvince (county, shire, department, etc.) in which the dcterms:Location occurs. Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names. Recommended best practice is to leave this term blank if the dcterms:Location spans multiple entities at this administrative level or if the dcterms:Location might be in one or another of multiple possible entities at this level. Multiplicity and uncertainty of the geographic entity can be captured either in the term dwc:higherGeography or in the term dwc:locality, or both.	Miranda
Share if available	dwc:municipality	The full, unabbreviated name of the next smaller administrative region than county (city, municipality, etc.) in which the dcterms:Location occurs. Do not use this term for a nearby named place that does not contain the actual dcterms:Location . Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names. Recommended best practice is to leave this term blank if the dcterms:Location spans multiple entities at this administrative level or if the dcterms:Location might be in one or another of multiple possible entities at this level. Multiplicity and uncertainty of the geographic entity can be captured either in the term dwc:higherGeography or in the term dwc:locality, or both.	Divinópolis

Term name	Definition	Examples*	Status
Share if available	dwc:locationRemarks	Comments or notes about the dcterms:Location .	pigeon loft at 50m
Share if available	dwc:verbatimLatitude	The verbatim original latitude of the dcterms:Location . The coordinate ellipsoid, geodeticDatum, or full Spatial Reference System (SRS) for these coordinates should be stored in dwc:verbatimSRS and the coordinate system should be stored in dwc:verbatimCoordinateSystem.	25° 33' 0" S
Share if available	dwc:verbatimLongitude	The verbatim original longitude of the dcterms:Location . The coordinate ellipsoid, geodeticDatum, or full Spatial Reference System (SRS) for these coordinates should be stored in dwc:verbatimSRS and the coordinate system should be stored in dwc:verbatimCoordinateSystem.	54° 34' 60" W
Share if available	dwc:footprintWKT	A Well-Known Text (WKT) representation of the shape (footprint, geometry) that defines the dcterms:Location . A dcterms:Location may have both a point-radius representation (see dwc:decimalLatitude) and a footprint representation, and they may differ from each other. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	POLYGON ((10 20, 11 20, 11 21, 10 21, 10 20))
Share if available	dwc:phylum	The full scientific name of the phylum or division in which the dwc:taxon is classified.	Arthropoda

Term name	Definition	Examples*	Status
Strongly recommended	dwc:class	The full scientific name of the class in which the dwc:Taxon is classified.	Insecta
Share if available	dwc:order	The full scientific name of the order in which the dwc:taxon is classified.	Diptera
Share if available	dwc:family	The full scientific name of the family in which the dwc:taxon is classified.	Culicidae
Share if available	dwc:identificationQualifier	A brief phrase or a standard term ("cf.", "aff.") to express the determiner's doubts about the dwc:Identification. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	Sciopemyia' aff. 'microps
Strongly recommended	dwc:identificationVerificationStatus	A categorical indicator of the extent to which the taxonomic identification has been verified to be correct.+ Recommended best practice is to use a controlled vocabulary such as that used in HISPID and ABCD. This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.	1 a

Term name	Definition	Examples*	Status
Share if available	dwc:nameAccordingTo	Harbach, R.E. 2013. Mosquito Taxonomic Inventory, https://mosquito-taxonomic-inventory.myspecies.info/ accessed on 25 april 2023	<p>The reference to the source in which the specific taxon concept circumscription is defined or implied - traditionally signified by the Latin "sensu" or "sec." (from secundum, meaning "according to"). For taxa that result from identifications, a reference to the keys, monographs, experts and other sources should be given. This term provides context to the dwc:scientificName.</p> <p>Together with the dwc:scientificName, separated by sensu or sec., it forms the taxon concept label, which may be seen as having the same relationship to dwc:taxonConceptID as, for example, dwc:acceptedNameUsage has to dwc:acceptedNameUsageID.</p>

Term name	Definition	Examples*	Status
Strongly recommended	dwc:behavior	A description of the behavior shown by the subject at the time the dwc:occurrence was recorded.+ Recommended best practice is to use a controlled vocabulary. Terms in the dwciri namespace are intended to be used in RDF with non-literal objects.	Endophagic exophytic zoophytic
Share if available	dwc:dataGeneralizations	Actions taken to make the shared data less specific or complete than in its original form. Suggests that alternative data of higher quality may be available on request. Terms in the dwciri namespace are intended to be used in RDF with non-literal objects.	Coordinates reflect the location of the study site where the sample was collected
Share if available	dwc:vitality	An indication of whether a dwc:organism was alive or dead at the time of collection or observation. Recommended best practice is to use a controlled vocabulary. Intended to be used with records having a dwc:basisOfRecord of preservedSpecimen, materialEntity, materialSample, or humanObservation. Terms in the dwciri namespace are intended to be used in RDF with non-literal objects.	dead

2.3.7. How to handle sensitive data

There might be cases in which a dataset might contain sensitive information (e.g. human subjects, location of people's homes, location data on endangered species), or perhaps one's institution has specific guidelines or policies regarding personal data or they might not want certain details publicly accessible. Types of sensitive data include:

- Personal data such as ethnic origin, genetic or biometric data, health/personal data relating to a natural person's routine and habits, socio-economic background data etc.
- Location of people's homes where entomological traps are placed
- Location data on endangered or protected species

To better address these sensitive data but still be able to contribute with data in GBIF it is possible to use some strategies to blur location data:

- Chapman AD (2020) Current Best Practices for Generalizing Sensitive Species Occurrence Data. Copenhagen: GBIF Secretariat. <https://doi.org/10.15468/doc-5jp4-5g10>.
- Astorga F, Rodrigues A & Waller J (2024) Sensitive species data generalization: Exploring risks, benefits and best practices. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/doc-yr9g-nx02>
- **Wicket** is a Javascript library that reads and writes **well-known text** (WKT) strings and generates a polygonal representation of the geometry to go into the `dwc:footprintWKT` term.
- **SimpleMappr**, a free tool to create point maps but also provides polygons' coordinates (Shorthouse 2010).
- Generalizing location information by obtaining wider, regional coordinates using <http://www.getty.edu/research/tools/vocabularies/tgn/Getty> Thesaurus of Geographic Names^] or **Google Maps**

With regards to health-related data, WHO has specific principles (WHO 2016, WHO 2017b, WHO 2024b) and **policies** on how to handle health-related data, including **WHO's policy on sharing and reuse of research data**, all of which include instructions on how to handle sensitive data with anonymization and de-identification information.

3. Future prospects

As **VBDs** continue to affect endemic countries or emerge and re-emerge throughout different regions of the world, continued efforts to provide open and accessible data about the biodiversity of these systems is crucial to developing effective prevention, control and elimination strategies. The value of data sharing has been evident during multiple public health emergencies of international concern (PHEICs), including the 2003 SARS outbreak, the Zika epidemic in 2015, and the COVID-19 pandemic, resulting in multiple manuscripts and frameworks published (WHO 2022). One way to improve data sharing is to focus on efforts in long-term endemic systems, as well as during interepidemic periods by facilitating learning and consolidating data sharing practices, which will strengthen concerted, focused public health actions against endemic and epidemic prone VBDs. With this guide, we hope to provide a set of practical recommendations to help publish open vector data in GBIF.org and other biodiversity data platforms to contribute to the understanding of vector distributions and disease dynamics.

Currently, searching for GBIF extensions is possible by using the GBIF API which enables users to make advanced queries that are not supported by the website. Please check the **API reference** and the **API beginners guide**. The **Resource Relationship extension** remains the best choice for handling interaction data without some degree of data loss, and GBIF and the **Darwin Core Maintenance Group** are working towards developing a new data model that will encompass the range and complexity of additional data types (Wieczorek & Robertson 2023). Overall, sharing data on vectors in GBIF will contribute to better preparedness, prevention and control of VBDs to improve human population health.

Acknowledgements

This guide is based on the work and discussions of the **task group** on mobilization and use of biodiversity data for research and policy on human diseases that was active between 2020-2025.

We would like to thank the **Special Programme for Research and Training in Tropical Diseases(TDR/WHO)** and **Scott Edmunds** from Gigabyte Journal **special issue** and **paper** for sponsoring and supporting, respectively, the series on data papers describing datasets on vectors of human diseases.

Special thanks to [Olivier Briet](#) for valuable discussion on mapping terms and developing the Resource Relationship extension for vector data and [Sofie Dhollander](#) and [Cedric Marsboom](#) for supporting open vector data. We are thankful to [Clara Baringo](#) for discussion on the Darwin core standard, and to [Victoire Nsabatien](#) and [Catalina Marcelo Diaz](#) for revising the draft version of this guide. We acknowledge [Theeraphap Chareonviriyaphap](#), [Sylvie Manguin](#) and [Marianne Sinka](#) for their support as members of the GBIF Task-group on mobilization and use of biodiversity data for research and policy on human diseases. And we are also very grateful to [Andrea Hahn](#) and [Kyle Copas](#) for support and valuable discussions on the GBIF data publishing infrastructure.

Glossary

anthropophily

Description of vectors that show a preference for feeding on humans, even when non-human hosts are available.

biting rate

Average number of vector bites a host receives in a unit of time, specified according to host and vector species (usually measured by human landing catch).

Darwin Core Archive (DwC-A)

Compressed (ZIP) file format for exchange of biodiversity data compiled in accordance with the [Darwin Core standard](#) (DwC). This self-contained set of interconnected CSV files and an XML document includes files and data columns and describes their mutual relationships.

Darwin Core (DwC) standard

Exchange standard for sharing and publishing biodiversity data comprising a set of identifiers, labels, and definitions to describe biodiversity data, originating from the [Biodiversity Information Standards](#) (TDWG) community. See the [Quick Reference Guide](#) for more information.

endophagy

Tendency of vectors to blood-feed indoors.

endophily

Tendency of vectors to rest indoors; usually quantified as the proportion of vectors resting indoors; important when assessing indoor residual spraying effectiveness.

entomological surveillance

The regular, systematic collection, analysis and interpretation of entomological data for risk assessment, planning, implementation, monitoring and evaluation of vector control interventions.

event

In GBIF context, species occurrences in time and space together with details of sampling effort.

exophagy

Tendency of vectors to blood feed outdoors.

exophily

Tendency of vectors to rest outdoors; usually quantified as the proportion of mosquitoes resting outdoors versus indoors; important when estimating outdoor transmission risks.

host

An ecologic system in which an infectious agent survives indefinitely (after [Ashford 2003](#)).

human biting rate

The number of adult female vectors that attempt to feed or are freshly blood-fed, per person per unit time.

IPT

Integrated Publishing Toolkit software developed and maintained by GBIF for managing and publishing open biodiversity data.

One Health

Integrated approach that considers that the health of humans, non-humans animals, plants, and the environment are closely linked and interdependent.

occurrence

In GBIF context, it is the occurrence of a species at a particular place and a specified date.

parity

The number of offspring a female has borne. In medical entomology, parity works as a proxy for the survival time of adult female vectors, mainly mosquitoes, and establishes whether a parasite has sufficient time to complete its life cycle within the vector, assisting in determining if the insect will serve as an effective vector.

parasite

Invertebrate organisms that live on or in another organism (the host), and benefit at the expense of the other. Traditionally excluded from definition of parasites are pathogenic bacteria, fungi, viruses and plants, which, though they may live parasitically, are termed pathogens.

pathogen

An organism causing disease to its host. Pathogens are found in a wide range of taxonomic groups and comprise viruses and bacteria as well as unicellular and multicellular eukaryotes.

reservoir

Sources which harbor disease-causing organisms and thus serve as potential sources of disease outbreaks.

resource

In the GBIF context, resources are datasets. **sampling event**: Investigating the presence/absence of an organism in a particular time and place, the investigation is well-documented by protocols and documentation of the sampling effort. Sampling events produce quantitative, calibrated data. The data can be very simple—a single event with a single occurrence (or no occurrences)—to highly hierarchical, with multiple parent-child event relationships.

species complex

A group of closely related organisms that are morphologically indistinguishable, and often other identification methods are employed to allow identification at species level.

sporozoite rate

Proportion of adult female vectors with sporozoites (motile stage of the malaria parasite) in their salivary glands.

vector

Invertebrates or non-human vertebrates which transmit infective organisms from one host to another.

vector-borne diseases (VBDs)

Infectious diseases transmitted by vectors.

zoophily

Preferring or seeking a non-human host over another animal.

References

- Aguilar-Vega C, Rivera B, Lucientes J, Gutiérrez-Boada I & Sánchez-Vizcaíno JM (2021) A study of the composition of the *Obsoletus* complex and genetic diversity of *Culicoides obsoletus* populations in Spain. *Parasites & Vectors* 14(1): 351. <https://doi.org/10.1186/s13071-021-04841-z>
- Ashford RW (2003) When Is a Reservoir Not a Reservoir? *Emerging Infectious Diseases* 9(11): 1495-1496. <https://doi.org/10.3201/eid0911.030088>
- Astorga F, Groom Q, Shimabukuro PHF, Manguin S, Noesgaard D, Orrell T et al. (2023) Biodiversity data supports research on human infectious diseases: Global trends, challenges, and opportunities. *One Health* 16:100484. <https://doi.org/10.1016/j.onehlt.2023.100484>
- Brazilian Zoology Group (2025). Catálogo Taxonômico da Fauna do Brasil. Version 1.30. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Checklist dataset. <https://doi.org/10.15468/c4cauy>
- Ceccarelli S, Balsalobre A, Cano ME, Vicente ME, Rabinovich JE, Medone P, Rocchi VM, Galliari JG & Marti GA (2022) Datos de ocurrencia de triatominos americanos del Laboratorio de Triatominos del CEPAVE (CONICET-UNLP). Version 1.6. Centro de Estudios Parasitológicos y de Vectores (CEPAVE). Occurrence dataset. <https://doi.org/10.15468/fbywtn>
- Colón-González FJ, Sewe MO, Tompkins AM, Sjödin H, Casallas A, Rocklöv J et al. (2021) Projecting the risk of mosquito-borne diseases in a warmer and more populated world: a multi-model, multi-scenario intercomparison modelling study. *The Lancet Planetary Health* 5(7): e404–e414. [https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196\(21\)00132-7/fulltext](https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(21)00132-7/fulltext)
- Cotes-Perdomo AP, Nava S, Castro LR, Rivera-Paéz FA, Cortés-Vecino JA & Uribe JE (2023) Phylogenetic relationships of the *Amblyomma cajennense* complex (Acari: Ixodidae) at mitogenomic resolution. *Ticks and Tick Borne Diseases* 14(3): 102125. <https://doi.org/10.1016/j.ttbdis.2023.102125>
- Dilermando J (2023) Fiocruz/COLFLEB - Coleção de Flebotomíneos. Version 1.59. FIOCRUZ - Oswaldo Cruz Foundation. Occurrence dataset. <https://doi.org/10.15468/sxcpfp>
- Edmunds SC, Fouque F, Copas KA, Hirsch T, Shimabukuro PHF, Andrade-filho JD et al. (2022). Publishing data to support the fight against human vector-borne diseases. *GigaScience* 11: giac114. <https://doi.org/10.1093/gigascience/giac114>
- Estrada-Franco JG, Fernández-Santos NA, Adebisi AA, López-López MJ, Aguilar-Durán JA, Hernández-Triana LM et al. (2020) Vertebrate-Aedes aegypti and Culex quinquefasciatus (Diptera)-arbovirus transmission networks: Non-human feeding revealed by meta-barcoding and next-generation sequencing. *PLoS Neglected Tropical Diseases* 14(12): e0008867. <https://doi.org/10.1371/journal.pntd.0008867>
- Garros C, Harbach RE & Manguin S (2005) Morphological assessment and molecular phylogenetics of the *Funestus* and *Minimus* groups of *Anopheles* (Cellia). *Journal of Medical Entomology* 42(4):522-36. <https://doi.org/10.1093/jmedent/42.4.522>
- Groom Q, Desmet P, Reyserhove L, Adriaens T, Oldoni D, Vanderhoeven S et al. (2019) Improving Darwin Core for research and management of alien species. *Biodiversity Information Science and Standards* 10(3): e38084. <https://doi.org/10.3897/biss.3.38084>
- Gutierrez MAC, Lopez ROH, Ramos AT, Vélez ID, Gomez RV, Arrivillaga-Henríquez J & Uribe S (2021) DNA barcoding of *Lutzomyia longipalpis* species complex (Diptera: Psychodidae), suggests the existence of 8 candidate species. *Acta Tropica* 221: 105983. <https://doi.org/10.1016/j.actatropica.2021.105983>
- Harbach RE (2012) *Culex pipiens*: species versus species complex taxonomic history and perspective. *Journal of the American Mosquito Control Association* 28(4 Suppl): 10-23.

<https://doi.org/10.2987/8756-971X-28.4.10>

- Ingenloff K (2025) Survey and Monitoring Data Quick-Start Guide: A how-to for updating a Darwin Core dataset using the Humboldt Extension. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/doc-7t3p-ve38>
- IPBES (2020) Workshop Report on Biodiversity and Pandemics of the Intergovernmental Platform on Biodiversity and Ecosystem Services. Daszak P, Amuasi J, das Neves CG, Hayman D, Kuiken T, Roche B et al. Bonn, Germany: IPBES Secretariat. <https://doi.org/10.5281/zenodo.4147317>
- Kelly ET, Mack LK, Campos M, Grippin C, Chen TY, Romero-Weaver AL et al. (2021) Evidence of Local Extinction and Reintroduction of *Aedes aegypti* in Exeter, California. *Frontiers in Tropical Diseases* 2. <https://doi.org/10.3389/fitd.2021.703873>
- Marceló-Díaz C, Morales CA, Fuya OP, Salamanca JA, Lesmes MC, Mendez-Cardona SA et al. (2023) Registros de los dípteros causantes de la transmisión del agente etiológico del dengue en el departamento del Cauca, Colombia. v1.10. Instituto Nacional de Salud. Dataset/Occurrence. <https://doi.org/10.15472/dxbowv>
- Motoki MT, Wilkerson RC & Sallum MA (2009) The *Anopheles albitarsis* complex with the recognition of *Anopheles oryzalimnetes* Wilkerson and Motoki, n. sp. and *Anopheles janconnae* Wilkerson and Sallum, n. sp. (Diptera: Culicidae). *Memórias do Instituto Oswaldo Cruz* 104(6): 823-850. <https://doi.org/10.1590/s0074-02762009000600004>
- Paull S (2022) NEON ticks sampled using drag cloths and tick pathogen status. National Ecological Observatory Network. Sampling event dataset. <https://doi.org/10.15468/b52b9z>
- Raele DA, Severini F, Toma L et al. (2024) *Anopheles sacharovi* in Italy: first record of the historical malaria vector after over 50 years. *Parasites & Vectors* 17: 182. <https://doi.org/10.1186/s13071-024-06252-2>
- Richards K, White R, Nicolson N & Pyle R (2011). A Beginner's Guide to Persistent Identifiers. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/mjgq-d052>
- Schwantes, CJ, Sánchez CA, Stevens T. et al. (2025) A minimum data standard for wildlife disease research and surveillance. *Scientific Data* 12: 1054. <https://doi.org/10.1038/s41597-025-05332-x>
- Shimabukuro P, Groom Q, Fouque F, Campbell L, Chareonviriyaphap T, Etang J et al. (2024) Bridging Biodiversity and Health: The Global Biodiversity Information Facility's initiative on open data on vectors of human diseases. *Gigabyte* 11: 1–11. <https://doi.org/10.46471/gigabyte.117>
- Shorthouse DP (2010) SimpleMappr, an online tool to produce publication-quality point maps. Retrieved from <https://www.simplemappr.net>. Accessed August 02, 2024.
- Shorthouse DP (2020) Slingshot With Four Giants on a Quest to Credit Natural Historians for our Museums and Collections. *Biodiversity Information Science and Standards* 4: e59167. <https://doi.org/10.3897/biss.4.59167>
- Simons D, Attfield L A, Jones K E, Watson-Jones D, Kock R (2023). Data from: Rodent trapping studies as an overlooked information source for understanding endemic and novel zoonotic spillover. *GigaScience Press*. Occurrence dataset. <https://doi.org/10.15468/zr6frj>
- Sinka ME, Pironon S, Massey NC, Longbottom J, Hemingway J, Moyes CL et al. (2020) A new malaria vector in Africa: Predicting the expansion range of *Anopheles stephensi* and identifying the urban populations at risk. *Proceedings of the National Academy of Sciences of the United States of America* 117 (40) 24900–24900. <https://doi.org/10.1073/pnas.2003976117>
- Soma D D, Zogo B, Taconet P, Mouline K, Ahoua Alou LP, Dabiré RK et al. (2024) *Anopheles* collections in the health districts of Korhogo (Côte d'Ivoire) and Diébougou (Burkina Faso) (2016–2018). Version 1.1. IRD - Institute of Research for Development. Sampling event dataset. <https://doi.org/10.15468/v8fvyn>

- Sukkanon C, Suwonkerd W, Thanispong K, Saeung M, Jhaiaun P, Pimnon S et al. (2023). Mosquitoes (Diptera: Culicidae) Distribution in Thailand. Version 1.1. Walailak University, School of Allied Health Sciences. Sampling event dataset. <https://doi.org/10.15468/tbd7fz>
- Stevens T, Zimmerman R, Albery G, Becker DJ, Kading R, Keiser CN et al. (2024) A minimum data standard for wildlife disease studies. EcoEvoRxiv. <https://doi.org/10.32942/X2TW4J>
- VectorNet, European Centre for Disease Prevention and Control, European Food Safety Authority (2025). VectorNet. VectorNet. Occurrence dataset. <https://doi.org/10.15468/f3k8r9>
- Wieczorek J & Robertson T (2023) Diversifying the GBIF Data Model. <https://www.gbif.org/new-data-model>
- World Health Organization (2007) Anopheline species complexes in South and South-East Asia. New Delhi: WHO Regional Office for South-East Asia. <https://iris.who.int/handle/10665/204779>
- World Health Organization, Regional Office for South-East Asia (2011) Comprehensive guidelines for prevention and control of dengue and dengue haemorrhagic fever. New Delhi: WHO Regional Office for South-East Asia. <https://iris.who.int/handle/10665/204894>
- World Health Organization (2016) Policy statement on Data Sharing by WHO in the Context of Public Health Emergencies. Geneva: World Health Organization. https://cdn.who.int/media/docs/default-source/publishing-policies/data-policy/who-policy-on-use-and-sharing-of-data-collected-in-member-states-outside-phe_en.pdf?sfvrsn=713112d4_27 https://cdn.who.int/media/docs/default-source/publishing-policies/data-policy/who-policy-statement-on-data-sharing-by-who-in-context-of-phe.pdf?sfvrsn=a97091f7_5
- World Health Organization (2017) Global vector control response 2017–2030. Geneva: World Health Organization. <https://www.who.int/publications/i/item/9789241512978>
- World Health Organization (2017) Policy on Use and Sharing of Data Collected in Member States by WHO Outside the Context of Public Health Emergencies. Geneva: World Health Organization. https://cdn.who.int/media/docs/default-source/publishing-policies/data-policy/who-policy-on-use-and-sharing-of-data-collected-in-member-states-outside-phe_en.pdf?sfvrsn=713112d4_27
- World Health Organization (2018) Malaria surveillance, monitoring & evaluation: a reference manual. Geneva: World Health Organization. <https://www.who.int/publications/i/item/9789241565578>
- World Health Organization (2019) Guidelines for malaria vector control. Geneva: World Health Organization. <https://iris.who.int/handle/10665/310862>
- World Health Organization (2022) Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance. Geneva: World Health Organization. <https://iris.who.int/handle/10665/352859>
- World Health Organization (2024) World malaria report 2024: Addressing inequity in the global malaria response. Geneva: World Health Organization. <https://www.who.int/publications/i/item/9789240104440>
- World Health Organization (2024) Information Disclosure Policy. Geneva: World Health Organization. https://cdn.who.int/media/docs/default-source/documents/about-us/infodisclosurepolicy.pdf?sfvrsn=c1520275_11
- Wilkinson M, Dumontier M, Aalbersberg I et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. <https://doi.org/10.1038/sdata.2016.18>
- Zanga J, Metelo E, Mvuama N, Nsabatien V, Mvudi V, Banzulu D et al. (2023) Species composition and distribution of anopheles gambiae complex circulating in Kinshasa. Version 1.2. University of Kinshasa. Sampling event dataset. <https://doi.org/10.15468/excax3>

Appendices

Vocabularies

Table A. Suggested controlled vocabulary for degree of establishment (from Groom et al. 2019).

degreeOfEstablishment	
Definition	Controlled value string
Not transported beyond limits of native range	native
Individuals in captivity or quarantine (i.e. individuals provided with conditions suitable for them, but explicit measures of containment are in place)	captive
Individuals in cultivation (i.e. individuals provided with conditions suitable for them, but explicit measures to prevent dispersal are limited at best)	cultivated
Individuals directly released into novel environment	released
Individuals released outside of captivity or cultivation in a location, but incapable of surviving for a significant period	failing
Individuals surviving outside of captivity or cultivation in a location, no reproduction	casual
Individuals surviving outside of captivity or cultivation in a location, reproduction is occurring, but population not self-sustaining	reproducing
Individuals surviving outside of captivity or cultivation in a location, reproduction occurring, and population self-sustaining	established
Self-sustaining population outside of captivity or cultivation, with individuals surviving a significant distance from the original point of introduction	colonising
Self-sustaining population outside of captivity or cultivation, with individuals surviving and reproducing a significant distance from the original point of introduction	invasive
Fully invasive species, with individuals dispersing, surviving and reproducing at multiple sites across a greater or lesser spectrum of habitats and extent of occurrence	widespreadInvasive

Table B. Suggested controlled vocabulary for traps (from VectorNet).

dwc:samplingProtocol
Active collections
Aedes gravid trap
Aspiration from animal bait
Aspirator collection from resting places
BG sentinel (+ UV light) trap
BG Sentinel + CO2
BG Sentinel trap with chemical lure
BG Sentinel trap with lure and CO2

dwc:samplingProtocol
Blacklight
Box animal baited trap
Carbon dioxide trap without light
CDC Light trap
CDC Light trap + CO ₂
CO ₂ baited trap
Collection from live/dead hosts
Cone animal baited trap
Disney animal baited trap
Dragging/flagging by surface
Dragging/flagging by time
Emergence trap
EVS Light trap
EVS Light trap + CO ₂
Gravid trap
Hair dryer by area
Hair dryer by time
Host-baited trap
Human landing catches
IMT trap
Lethal ovitrap
Light CO ₂ sticky
Light sticky trap
Light trap
Light trap + CO ₂
Malaise trap
Mosquito magnet trap
Netting
Non-standard dipping
NS trap + CO ₂
Onderstepoort Veterinary Institute (OVI) black-light trap
Other
Other CO ₂ trap
Other sticky trap
Ovitrap
Resting catch (outdoor/indoor)
RIEB light trap
Rothamsted trap

dwc:samplingProtocol
Soil sampling
Soil screening by surface
Soil screening by weight
Soil screening by weight
Standard dipping
Sticky trap
UK light trap
Unknown
Updraft box trap with light

Table C. Suggested controlled vocabulary for life stage.

lifeStage
egg
larva
nymph
pupa
adult
subadult
juvenile

Table D. Suggested controlled vocabulary for sex.

sex
female
male
hermaphrodite

Table E. Suggested controlled vocabulary for production sites for mosquitoes from the [Comprehensive Guidelines for Prevention and Control of Dengue and Dengue Haemorrhagic Fever](#).

habitat
Water evaporation cooler
Water storage tank/cisternrelation
Drum (40–55 gallons)
Flower vase with water
Potted plants with saucers
Ornamental pool/fountain
Roof gutter/sun shades
Animal water container
Ant-trap
Used tyres
Discarded large appliances

habitat
Discarded buckets
Discarded food and drink containers
Tree holes
Rock holes

Table F. Suggested controlled vocabulary for depicting relationships, based on the [OBO Ontology](#).

resourceRelationshipID
obligate parasite of
is vector for
intracellular endoparasite of
has reservoir host
reservoir_host_of
has vector
has pathogen
obligate parasite of
facultative parasite of
interacts with via parasite-host interaction
parasite of
host of
pathogen of

Useful links to resources

Here we provide a list of useful or relevant links to websites and resources to facilitate data publishing in GBIF. For convenience, links are organized by category.

General information and guides

- [Darwin Core term Quick Reference Guide](#)
- [Darwin Core text guide](#)
- [GBIF Technical Documents](#)
- [GBIF dataset classes](#)
- [GBIF resource metadata](#)
- [Publishing DNA-derived data guide](#)
- [Best practices](#) for publishing sampling event datasets
- Introduction to [event](#) data
- How to [publish](#) datasets
- Consortium E, Body G, de Mousset M, Chevallier E, Scandura M, Pamerlon S, et al. Applying the Darwin core standard to the monitoring of wildlife species, their management and estimated records. EFSA Supporting Publications. 2020;17(4):1841E. <https://onlinelibrary.wiley.com/doi/abs/10.2903/sp.efsa.2020.EN-1841>
- Consortium E, Jaroszynska F, Body G, Pamerlon S, Archambeau AS. Applying the Darwin Core data standard to wildlife disease – advancements toward a new data model. EFSA Supporting Publications. 2022;19(11):7667E. <https://onlinelibrary.wiley.com/doi/abs/10.2903/sp.efsa.2022.EN-7667>
- Groom Q, Desmet P, Reyserhove L, Adriaens T, Oldoni D, Vanderhoeven S, et al. Improving Darwin Core for research and management of alien species. Biodiversity Information Science and Standards. 2019 Oct 10;3:e38084.<https://biss.pensoft.net/article/38084/>[<https://biss.pensoft.net/article/38084/>]

Coordinates, mapping and georeferencing

- [Coordinate Conversion Tool](#): used to convert coordinates from degrees minutes seconds to decimal degrees
- [InfoXY](#) from CRIA for validating geographic data
- [Georeferencing Calculator](#), a point-radius method for georeferencing locality descriptions and calculating associated uncertainty
- Bloom DA, Wieczorek JR & Zermoglio PF (2020) Georeferencing Calculator Manual. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/gdwq-3v93>
- Chapman AD & Wieczorek JR (2020) Georeferencing Best Practices. Copenhagen: GBIF Secretariat. <https://doi.org/10.15468/doc-gg7h-s853>
- Zermoglio PF, Chapman AD, Wieczorek JR, Luna MC & Bloom DA (2020) Georeferencing Quick Reference Guide. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/e09p-h128>
- [Getty Thesaurus of Geographic Names](#): search interface for place names

Controlled vocabulary and ontologies

- **Darwin Core Basis of Record Vocabulary**: definitions for terms used to populate basisOfRecord
- Metadata descriptors for health: **DeCS/MeSH**, **Global Index Medicus**, **Unified Medical Language System (UMLS)**
- Repository for biomedical ontologies: **EMBL-EBI Ontology Lookup Service** , for ontologies on habitat, e.g. **barn**; gonotrophic cycle, e.g. **exophily**; detection assays, e.g. **ELISA**; among others.
- For more general health information, the **HL7® FHIR®** is an interoperability standard designed to enable health data sharing.

Taxonomy and date issues

- **GBIF Name parser**: tool to divide scientific names into their components and to check them against the taxonomic backbone used by GBIF
- GBIF's **species lookup**: to normalize species names against the GBIF backbone
- **ChecklistBank**
- **Canadensys Date parser**: tool to parse dates into component parts
- A **guide** to date issues
- **How to** handle date intervals

Quality control tools and resources

- Data quality **checklist**
- GBIF's **data validator**
- Zermoglio PF, Plata Corredor CA, Wieczorek JR, Ortiz Gallego R & Buitrago L (2021) Guía para la limpieza de datos sobre biodiversidad con OpenRefine. Versión 3. Copenhagen: GBIF Secretariat. <https://doi.org/10.15468/doc-gzjg-af18>
- **OpenRefine** free, open source tool for cleaning data.
- **Hmisc R package** for summarizing data, useful for checking data quality and outliers
- **Bionomia**, is an online platform that links people to specimens specimens (**Shorthouse 2020**), and allows for name disambiguation when one collector has used different names, or a name is written in more than one way. The specimen data associated with the terms **recordedBy**, **recordedByID**, **identifiedBy**, **identifiedByID** are used by Bionomia to generate attribution.
- **Excel to Darwin Core Standard (DwC) Tool**: creates templates for Event, Occurrence, MeasurementsOrFacts, Extended MeasurementsOrFacts (EMoF), and Simple Multimedia tables.
- **Darwin Core Template Generator for Event and Occurrence tables**: user can choose DwC terms to generate template for Event and Occurrence tables, developed by the Nansen Legacy project.
- Template **generator** by GBIF Norway
- Table converter from columns to rows <https://mycena.sibbr.gov.br/> by GBIF Brazil - SiBBR (in Portuguese).

Publishing tools

- **Integrated Publishing Toolkit (IPT) User Manual**
- **Ecological Metadata Language (EML) guide**
- **Become a GBIF publisher**: necessary for dataset publication in GBIF