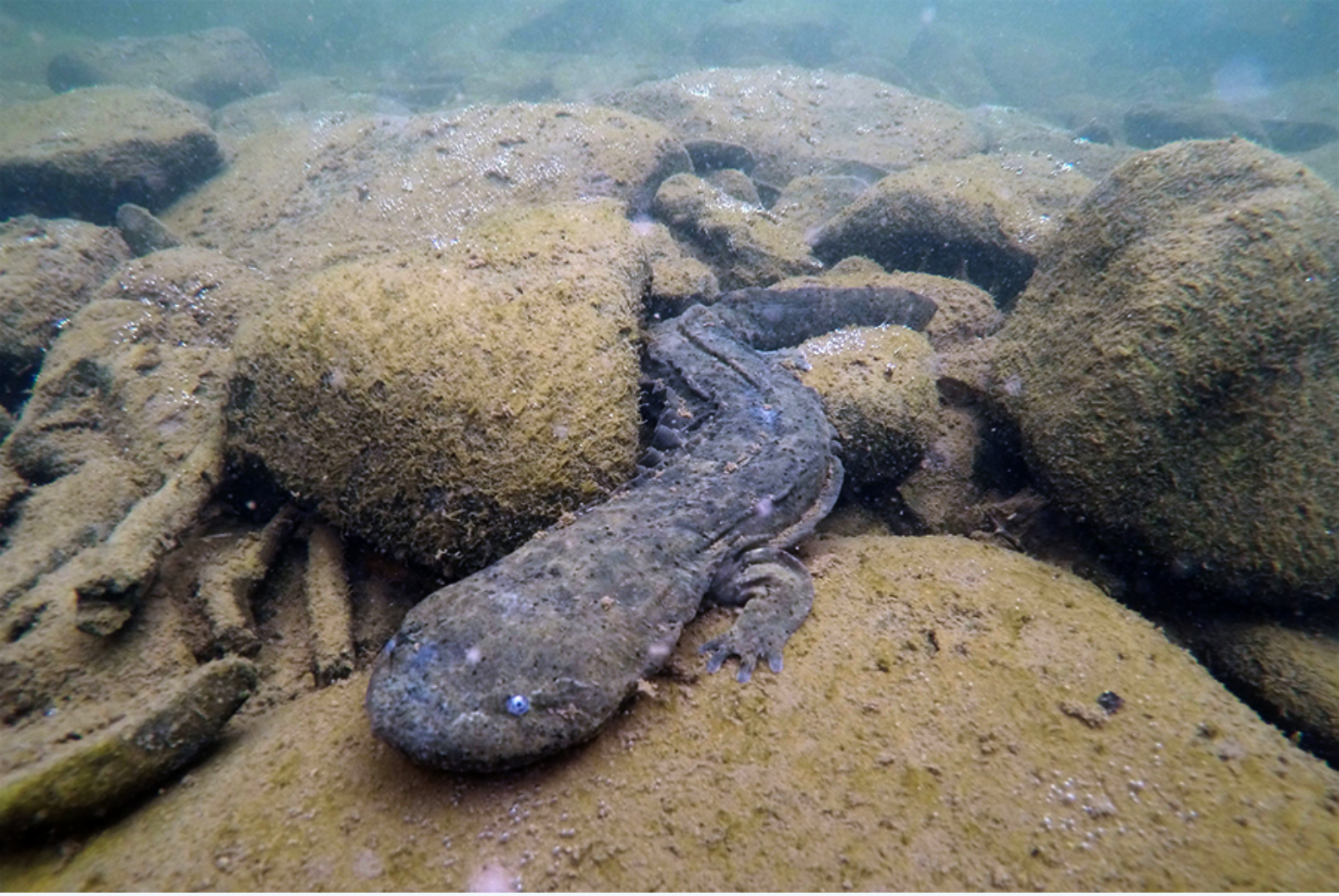

Freshwater Data Publishing Guide

Jennifer Lento • Astrid Schmidt-Kloiber

Version fba46d5, 2025-02-10 12:26:41 UTC

GBIF Global Biodiversity Information Facility

Table of Contents

Colophon

Suggested citation

Lento J & Schmidt-Kloiber A (2025) Freshwater Data Publishing Guide. Copenhagen: GBIF Secretariat. https://doi.org/10.35035/doc-sw3k-w725

Authors

Jennifer Lento & Astrid Schmidt-Kloiber

Licence

Persistent URI

https://doi.org/10.35035/doc-sw3k-w725

Document control

v.1.0.1 Initial release, February 2025

Introduction

Freshwater data publishing

Freshwaters are vital to human well-being, yet they are among the most threatened ecosystems on the planet (Dudgeon et al. 2006; Reid et al. 2019). Despite increasing efforts over the past decades to protect freshwaters and improve their state, our ability to monitor, detect and manage changes to freshwater ecosystems and their biodiversity is limited by the availability of supporting data (GEO BON & FWBON 2022). Recent research highlights the need for better data availability to support scientific analyses (Darwall et al. 2018; Tickner et al. 2020; Harper et al. 2021; Van Rees et al. 2021). Data that follow the FAIR principles of *Findability*, *Accessibility*, *Interoperability* and *Reusability* (Wilkinson et al. 2016) in particular can support advancements in establishing freshwater biodiversity baselines and detecting changes in response to stressors. For example, improved data availability has the potential to facilitate large-scale analyses of spatial patterns and temporal trends in freshwater biodiversity, such as the state of Arctic freshwater biodiversity assessment conducted by the Circumpolar Biodiversity Monitoring Program (Lento et al. 2019). Furthermore, improved spatial and temporal coverage of freshwater biodiversity data, particularly from biodiversity hotspots and remote or unique habitats, can support the inclusion of freshwater species in global indices, such as WWF's freshwater Living Planet Index (Ledger et al. 2023).

Despite the growing calls for open data and increasing requirements that publicly funded data be made open access (Beno et al. 2017; Sholler et al. 2019), a number of impediments remain to data mobilization, including

- a lack of capacity (time, funding, expertise) to set up data in the required formats for data

publishing

- concerns over intellectual property rights (e.g. ownership, attribution) of shared data
- lack of incentives to publish data
- a reluctance to publish data prior to publication (Beno et al. 2017; Schmidt-Kloiber & De Wever 2018)

A recent global survey on data mobilization among freshwater practitioners (unpublished) noted that financial support for publishing and long-term data management of data was a major concern for most respondents. While some of the identified barriers to data mobilization reflect institutional obstacles, such as the lack of funding and data publishing incentives, there is an opportunity to improve the ease with which data providers can share their data.

Potential users of open data may encounter other obstacles, such as a lack of interoperability of data from different data portals, a lack of supporting metadata (e.g. information about sites and methods), ambiguities or inaccuracies in the (meta)data, and difficulties in searching for and selecting relevant data (e.g. finding diversity data only for freshwater species) (Beno et al. 2017). A number of these issues can possibly be controlled by improving the quality of published data and metadata, for example, by ensuring a standard set of metadata is always provided with the data, ensuring data have relevant tags that will allow them to be located by potential users, and following a standard template for data publishing. For freshwater data, these improvements could facilitate greater and more widespread use of open data from portals like GBIF.org feeding into global biodiversity assessments.

Scope of the guide

This guide describes best practices for freshwater data publishing, including a list of the (meta)data fields that are necessary to describe sample locations and sampling methods in sufficient detail to allow meta-analysis across datasets. One of the aims of the guide is to provide the user with sufficient background information to understand why particular fields are necessary for freshwater data, for example, fields describing the type of freshwater ecosystem in which the organism was observed, or indicating the freshwater organism group to which an individual belongs. This information is intended to improve the quality of published freshwater data.

Another goal of this guide is to make the process of formatting data for publication more user-friendly by describing what information should be included in each field and providing freshwater-relevant examples. One of the reasons why scientists are often still reluctant to publish data is that preparing, curating and formatting it requires a lot of time that is usually not budgeted for in research projects (Schmidt-Kloiber & De Wever 2018). Therefore, it is important that technically straightforward procedures are available to support general freshwater data mobilization. This guidance manual will help to support data publishing by providing instruction on best practices for formatting data, with a focus on the GBIF infrastructures.

Target audience

The target audience for this guide is researchers in government, academia and NGOs, as well as all freshwater practitioners (e.g. people engaged in designing and carrying out monitoring), as the best practices described herein offer insights for anyone involved in collecting, recording and managing freshwater data. But in particular, the guide is intended to assist those who are or will be formatting freshwater data with the intent of publishing those data. The goal is to give those individuals the tools needed to ensure the data they publish are consistent with global standards and contain sufficient supporting information to ensure their usefulness in meta-analyses of biodiversity.

1. Specific freshwater data considerations

1.1. Characteristic features of freshwater data

Freshwater ecosystems are unique because they are contained within terrestrial landscapes and are influenced by activities within the terrestrial environments that they drain (the catchment). Standing waters (lakes, ponds, wetlands) are often connected by running waters (rivers, streams), but dispersal of organisms within these hydrologic networks is limited by the organism's mobility and dispersal traits (Comte and Olden 2018; Sarremejane et al. 2020), as well as by the influence of the terrain, flow conditions, and incidence of flooding/drought (Gido et al. 2016; Carvajal-Quintero et al. 2019). The dispersal limitations imposed by the degree of connectedness between freshwater systems help to define spatial patterns of biodiversity.

Biodiversity of freshwater organisms is also affected by local water quality conditions and upstream influences. Rivers, for example, are longitudinal systems that are highly dependent on conditions and changes occurring upstream. Biodiversity in these systems and throughout the hydrologic network is influenced by the entire upstream catchment, with impacts to terrestrial and freshwater ecosystems leading to biological responses downstream (Ward 1998).

As a result of their unique habitat, freshwater organisms have certain characteristics that set them apart from organisms in other realms, and these characteristics must be either reflected in the descriptive metadata that accompany observation data or in the occurrence datasets themselves. Details regarding the observed organism's life cycle stage, the water body type they live in, or the way they are sampled are important for understanding the degree of comparability among different datasets. This guidance document highlights the ways in which such information should be included in datasets to support a harmonized approach to data publishing. In many cases, the information that should be provided differs depending on the organism group (e.g. fish, benthic macroinvertebrates, phytoplankton) (see §1.2.3). Here, we introduce the characteristics of freshwater data that should be considered when preparing the datasets to be published on GBIF and/or elsewhere.

1.1.1. Habitat use by freshwater organisms

There are several spatial scales at which habitat information should be reported with observation data to improve dataset utility and usability. Here, we define these based on definitions in the IUCN Global Ecosystem Typology, which is a hierarchical classification system that groups the world's ecosystems by realm, biome, and ecosystem functional groups. At the largest scale, **realms** differentiate between terrestrial, freshwater, marine, subterranean, and atmospheric components of the biosphere, as well as transitional zones between realms. **Biomes** are ecologically similar subcomponents of realms with broadly similar features. Within biomes, the typology classifies ecosystems by joining those with similar ecological conditions into **ecosystem functional groups**. Different classification levels in the hierarchy of this typology offer relevant information to better understand observations of freshwater data.

One of the challenges of freshwater data is the prevalence of organisms that make use of multiple realms. For example, a large number of freshwater macroinvertebrates are insects, many of which live in freshwater only during particular life stages (e.g. as immature larvae or nymphs) and live in the terrestrial realm as adults. Some fish species are anadromous or catadromous, which means that part of their life is spent in freshwater and part is spent in marine habitats. Many bird species make use of freshwater for feeding, while breeding and nesting in terrestrial habitats. Some plant species are capable of growing in freshwater systems but also in high-moisture terrestrial habitats. In these and other cases, it is not sufficient to simply know that the species was observed. It is highly relevant to know whether observations of the species were made in and around freshwater or in another realm, as this provides important information about life stage, spatial distribution, and habitat use.

Information about the biome in which an observation was made is also important for understanding the data and for grouping comparable datasets. The IUCN Global Ecosystem Typology separates the freshwater realm into three biomes—rivers and streams; lakes; and artificial wetlands—while groundwater, brackish water, palustrine wetlands, and coastal systems are grouped within transitional realms.

This classification, which describes the type of water body in which the organism was observed, can provide important ecological information to support observations. Within biomes, ecosystem functional groups, which describe ecological conditions (e.g. permanent, seasonal or episodic/ephemeral; freeze-thaw; upland or lowland; large or small), provide further information that is relevant to understanding how comparable ecosystems (and therefore observations) might be. Classification of sampled ecosystems within these groups is necessary to understand the data and facilitate broad-scale assessments of biodiversity.

At a smaller spatial scale, beyond the scope of the IUCN Global Ecosystem Typology, the habitat zones sampled within a water body contribute to biodiversity differences between datasets. Within freshwater bodies, there are natural differences in taxonomic composition among habitat zones that highlight the importance of indicating this information in the supporting metadata. In lakes, these differences are evident among lake zones, which define habitats based on depth and characteristics related to light penetration, oxygen levels, substrates and temperature. For example, the taxonomic composition of macroinvertebrate and algae samples collected from the littoral (shallow shoreline) zone in lakes differs greatly from that of samples collected in the profundal (deep) zone and these samples are generally not comparable. Similarly, natural differences in taxonomic composition may be expected in plankton and fish samples collected from littoral and pelagic (open water) zones of lakes. River mesohabitats differentiate between types of flow, with riffles being fast-flowing shallow rocky areas, runs representing deeper fast-flowing areas, and pools indicating areas of slow-flowing or standing water, all of which might be expected to house different taxa and biomass. Furthermore, benthic samples collected along the margins of a larger river may differ naturally in composition from deeper samples collected from the center of the channel due to differences in substrate and flow conditions. Information about the sampled lake zone or river mesohabitat is therefore necessary to assess comparability of datasets. At the finest scale, information on sampled microhabitats within lakes or rivers (e.g. samples collected from a particular substrate type such as sandy or rocky), when available, can provide a further indication of expected biodiversity patterns. It can also be informative to differentiate between samples collected across multiple microhabitats from those that were specific to a particular microhabitat type.

1.1.2. Organism life cycle stage and sample timing

For some freshwater species, the life cycle stage of the observed organism is important for understanding community and population dynamics, as well as providing information about life history timing, or phenology. For example, the life cycle stage of insect species provides an indication of whether the observation was of a freshwater or terrestrial habitat for those species that have freshwater juvenile life stages and terrestrial adult life stages. For insects that have multiple freshwater life stages (e.g. beetles that live in freshwater as larvae, pupae, and adults), this information contributes to a greater understanding of population dynamics by indicating the relative proportion of adults and juveniles at the time of sampling. The relative proportion of larvae and pupae for insects with complete metamorphosis, along with the timing of sampling, also provides information about the timing of adult emergence, which is important to track in relation to changes in water temperature.

Life stage details provide similarly valuable information about population dynamics for other organism groups, such as zooplankton and fish. Within zooplankton assemblages, it may be useful to know how many juvenile copepodites and adult copepods are present, and labeling of individuals as nauplii (the first larval life stage for copepods) might be necessary if individuals are too young for species-level

identification. For fish, tagging of individuals as young-of-the-year (fish age 0, born within the last year) is important for tracking population dynamics, particularly of threatened or at-risk species. Life stage information is also helpful to understand the timing of important life history events, such as fish migration, spawning, and hatching.

The timing of sampling is highly relevant to understanding life stage information for observed taxa. Full details about the date of sampling (including day, month, and year) are critical for tracking changes in the timing of life history events. Furthermore, information about the season of sampling (recognizing differences between the northern and southern hemisphere) provides context for sampling and allows datasets to be grouped by similar seasonal conditions.

1.1.3. Sampling methods

Sampling methods play a major role in determining how comparable different freshwater datasets are and whether freshwater data from different sources can be combined in a meaningful meta-analysis of biodiversity measures and community composition (Lento et al. 2019; Jarvis et al. 2023). For example, different freshwater fish sampling gear types are only effective on a portion of the fish assemblage, and some net mesh sizes fail to capture small-bodied fish species. Combining fish datasets with diverse sampling methods may thus introduce methodological bias into the meta-analysis.

Similarly, the type of sampling equipment used to collect benthic macroinvertebrates has an impact on the collected data. For example, grab samplers (e.g. Ekman and Ponar grabs) and dredges would not be expected to collect comparable samples to fixed area sampling equipment such as Surber and Hess samplers or deployed equipment such as Hester-Dendy samplers. Furthermore, fixed-area samplers may not collect as much diversity as multi-habitat samplers such as kick nets. Mesh size of samplers also plays a role in determining comparability of benthic macroinvertebrate samples, as smaller net mesh sizes capture smaller animals which might increase the number of species at a given site.

For small-bodied planktonic organisms, the net mesh size or the filter pore size is critical for understanding the degree of comparability of different samples. For example, zooplankton nets vary in mesh size, with the larger sizes excluding rotifers and other small-bodied zooplankton and thus underestimating diversity and potentially excluding an entire phylum (Mack et al. 2012; Pansera et al. 2014). Phytoplankton samples are often taken by collecting and filtering a water sample, but the filter pore size will impact whether picoplankton and nanoplankton (among the smallest size classes of phytoplankton) are retained.

Differences in the amount of effort spent sampling, as measured by time, area, or number replicates, ultimately impacts the abundance of collected taxa as well as the probability of collecting more taxa. The greater the effort, the higher the diversity of collected samples (up to a point where additional effort does not increase the number of taxa collected; Gotelli and Colwell 2001). However, it is important to recognize that some compromise is necessary when combining datasets for analysis. While differences in sampling equipment and mesh size can have dramatic effects on the comparability of different datasets, differences in effort may be accounted for in analysis and interpretation.

1.2. Dataset categorization and terminology

1.2.1. GBIF dataset classes

GBIF defines and supports four classes of datasets: resources metadata (metadata-only datasets), checklist datasets, occurrence datasets, and sampling-event datasets (for detailed definitions and metadata requirements, see Dataset classes and How to choose a dataset class on GBIF?).

Differences between dataset classes are defined in terms of the amount of information provided by the data holder. In brief:

- **Resources metadata** is the most simple class, providing information about datasets that are not digitized or that are housed elsewhere and cannot be uploaded to GBIF. They do not provide taxon observation data, but they indicate the existence of such information, and may provide some details about the datasets as well as information on how to access such datasets (if at all possible).
- **Checklist datasets** provide summary taxa lists without dates or locations for individual observations. They include lists of taxa that are found within a region or country, regional lists of threatened species, and similar summaries.
- **Occurrence datasets** record observations of the occurrence of a taxon, including the taxon name and information about where and when the taxon was observed. Occurrence datasets may be provided with or without counts for each taxon. Location and date information may be coarse for these datasets (e.g. providing only country and year), though recommended best practice is to be as specific as possible (e.g. always providing coordinates).
- **Sampling-event datasets** represent the most detailed dataset class, and have to consist of two files: one occurrence dataset file (taxon presence or counts) with detailed information on location and date, as well as a separate file with information about sampling methods that were used.

Each dataset class allows for different usage of the data. The simpler classes allow for more basic descriptions of the geographic range of available records, observed geographic ranges of taxa, or summaries of expected taxa within a region. In contrast, the most detailed classes (e.g. the sampling-event dataset) allow for the assessment of community composition and biodiversity measures.

1.2.2. Freshwater data categories

To support the effective use of GBIF datasets, whether in simple summaries or more in-depth assessments, there are additional ways to categorize freshwater datasets beyond the four defined GBIF classes. While the GBIF classes largely reflect the amount of available data or metadata, it is important to categorize occurrence and sampling-event datasets based on the type of observation that was made. Based on the type of observation, freshwater datasets can be:

- **Opportunistic observation data**: unplanned observations that are not part of a systematic sampling event, but that occur as circumstances allow. Specific effort is not made to observe or collect particular species or an assemblage of species, and no sampling protocol is used.
  Example: data originating from bird watching or records from iNaturalist or similar apps.
- **Targeted sampling data**: planned sampling events that are focused on capturing a particular species or a subset of an assemblage of species. Observations of other (non-target) species in the assemblage are not recorded.
  Example: fish sampling event that is focused only on collecting Atlantic salmon, or zooplankton sampling event that is focused on cladoceran zooplankton only.
- **Assemblage sampling data**: planned sampling events in which the goal is to sample the full assemblage. Observations are recorded for all species in the assemblage that are collected.
  Example: benthic macroinvertebrate sampling of the entire assemblage at a site, or fish assemblage sampling at a site, as part of a biomonitoring program.

The importance of categorizing freshwater datasets based on the type of observation relates to how the data can be used in further analyses. If data represent opportunistic observations, they can only be used to indicate species presence. Opportunistic observations cannot be used to indicate where a species is not found (e.g. to draw conclusions about its conservation status) nor can they describe the abundance of a species, because no systematic effort has been made to detect the species or

quantify its abundance. Caution is therefore advised when combining opportunistic observation data with data from targeted or assemblage sampling, as the conclusions that can be drawn from opportunistic observations are more limited than what might be possible with data that resulted from structured sampling efforts.

Caution is also necessary when combining datasets from organized sampling efforts. Targeted sampling data and assemblage sampling data cannot be compared in terms of diversity or community composition because targeted sampling does not represent an attempt to record all observed taxa and thus does not describe the assemblage as a whole. While the absence of a particular taxon from assemblage sampling data suggests that the taxon was not found in a particular location during the sampling event, its absence from targeted sampling data may simply reflect the fact that it was not the species of interest during sampling and was therefore not recorded.

Freshwater datasets should also be categorized based on the type of data contribution, which we define as:

- **Professional data**: data that were collected by researchers, scientists, or taxonomic experts, that result from samples processed by a professional laboratory, or that have undergone quality assurance/quality control, thus indicating high confidence in the accuracy of the data.
- **Community-based research data**: data that were collected through organized public participation in sampling events or public-led sampling events, designed and/or operated through collaboration with professionals. Expert training by professionals instills confidence in the accuracy of the data, but the potential for error is higher than for professional data.
- **Citizen science data**: data collected through observations by members of the public without formal training/expertise or professional support (see Citizen Science for an overview). This includes individual observations recorded through platforms that share their data with GBIF, such as iNaturalist or observation.org.

The type of data contribution has implications for the types of quality checks that may be necessary for datasets retrieved from GBIF. For example, citizen science data may require different quality checks than professional data provided by taxonomic experts or observations from lab-processed samples (Jarvis et al. 2023), particularly for taxonomic groups that must be identified with a microscope. The distinction between community-based research data and citizen science data in our definitions is based on the degree to which there has been training and/or collaboration with professionals, increasing the probability of accurate sampling results. Under these definitions, citizen science data are those collected without training or support from professionals, which are therefore most likely to require quality checks before further data use.

1.2.3. Organism groups

Users who search for data on GBIF may be interested in the general biodiversity of all organisms in a region, but many have an interest in the diversity of a particular organism group. Organism groups are collections of biologically and ecologically similar organisms that are generally grouped together and described as an assemblage. For example, phytoplankton is an organism group that refers to microscopic and planktonic (passive floaters/drifters and weak swimmers that are carried by current) autotrophic (self-feeding) organisms, including algae and bacteria. Benthic macroinvertebrates refers to a group of organisms that can be seen with the naked eye (not microscopic), that have no backbone and that live on the bottom of lakes, rivers, and wetlands, including worms, snails, clams, and aquatic life stages of insects. Generally, freshwater organism groups often comprise more than one order/class/phylum (e.g. benthic macroinvertebrates consist of Trichoptera, Plecoptera, Gastropoda, etc.). The groupings offer a way to refer to particular components of freshwater communities generally studied together.

Adding the organism group to which an observation belongs is a way to make data easier to find and

select within GBIF. For example, someone who is interested in phytoplankton diversity would find it useful to be able to select data by the organism group name (phytoplankton) rather than having to search separately for the taxonomic classes that are part of this assemblage. Furthermore, someone who is interested in identifying the spatial distribution of benthic macroinvertebrate sampling data globally would have more success in finding data if each of the taxa of interest (reaching from class to orders) were annotated with the organism group name. Table 1 outlines the organism groups that we recommend adding to freshwater records in GBIF.

Table 1. Freshwater organism groups, their status as aquatic and/or semi-aquatic, and a description of each group with examples of taxa that are part of the group.

Organism group	**Aquatic status**	**Description**
Fungi	Aquatic	Freshwater fungi
Microbes	Aquatic	Freshwater microbial species, such as bacteria, fungi, protozoa, viruses, and other microorganisms
Benthic algae	Aquatic	Microscopic plants (algae) and autotrophs collected from bottom habitats, such as diatoms, green algae, red algae, golden algae, cyanobacteria, and others
Phytoplankton	Aquatic	Microscopic plants (algae) and autotrophs collected from the water column, such as diatoms, green algae, red algae, golden algae, cyanobacteria, and others
Macrophytes	Aquatic, semi-aquatic	Aquatic and semi-aquatic macroscopic plants and mosses, such as emergent, submergent, or floating types, found in or near freshwater
Zooplankton	Aquatic	Microscopic planktonic invertebrates, generally collected from the water column, such as cladocerans, copepods, or rotifers
Benthic macroinvertebrates	Aquatic, semi-aquatic	Macroscopic invertebrates collected from benthic habitats, such as segmented and unsegmented worms, molluscs, and freshwater insects; may also include crustaceans
Decapods *may be grouped with benthic macroinvertebrates*	Aquatic	Macroscopic crustaceans with 10 legs that may require specialized sampling approaches, separate from those of macroinvertebrates, such as crayfish, shrimp, and crabs
Fish	Aquatic	Fish that live all or part of their lives in freshwater (including anadromous and catadromous species)
Amphibians	Aquatic, semi-aquatic	Amphibians living in and around freshwater, such as frogs, newts, and mudpuppies
Reptiles	Aquatic, semi-aquatic	Reptiles living in and around freshwater, such as turtles, snakes, and crocodiles
Birds	Aquatic, semi-aquatic	Birds that live in or around freshwater for at least part of the year, such as wading and diving birds
Mammals	Aquatic, semi-aquatic	Mammals that live in or around freshwater, such as otters, beavers, and muskrats

Many of the details about sampling methods recommended for inclusion in published freshwater datasets vary depending on the organism group, and applying the labels in Table 1 would facilitate the use of conditional or recommended fields during dataset upload. For example, life stage is a relevant

field for benthic macroinvertebrate or fish samples, but not for benthic algae samples. Below, we provide information about relevant fields and sampling details for freshwater organism groups.

1.3. Metadata requirements

1.3.1. Publishing specific data categories on GBIF

An important part of publishing datasets on GBIF is ensuring that sufficient metadata are provided to allow future use of the published dataset. Some metadata are registered at the resource (dataset) level (i.e., the dataset description, version, citation, rights, keywords, contacts, taxonomic and geographic scope) while other metadata can be captured in the records themselves in either occurrence or sampling-event tables.

Freshwater datasets published on GBIF should include the GBIF dataset class (listed as type of dataset: resources metadata, checklist, occurrence, or sampling-event) in the metadata. We recommend adding the type of observation (opportunistic observation data, targeted sampling data, or assemblage sampling data (see §1.2.2)) and the type of data contribution (professional data, community-based monitoring data, or citizen science data) to the occurrence dataset (see <<§2.2>> and <<§3.1.1>>). These categories reflect the opportunities and limitations of each dataset for large-scale data compilation and biodiversity assessment more accurately than the GBIF dataset classes. Table 2 indicates which of these categories can be applied to occurrence or sampling-event datasets. Note that the freshwater data categories may apply to different GBIF dataset classes depending on the amount of information available in the dataset, as indicated below.

Table 2. GBIF dataset classes and the freshwater observation and contribution types that may be applied to each class. The "X" indicates which types of observations and contributions can be submitted to GBIF as either occurrence data or sampling-event data. GBIF dataset classes and freshwater data categories are defined in §1.1 and §1.2.

Freshwater data categories	GBIF dataset class	
	Occurrence data	Sampling-event data
Type of observation		
Opportunistic observation	X	
Targeted sampling data	X	X
Assemblage sampling data	X	X
Type of data contribution		
Professional data	X	X
Community-based research	X	X
Citizen science	X	

Opportunistic observation data are not collected as part of a planned sampling event, e.g. they are not collected through a structured effort to describe the assemblage composition or estimate the geographic distribution or population size of a particular species. Instead, these data may represent secondary observations of non-target species or casual observations of species. Opportunistic observations are grouped as occurrence datasets under GBIF's dataset classification system because there are no specific sampling methods to report (Table 2). Opportunistic observation data include presence-only records or counts, but the latter is not particularly meaningful without information about the planned effort that can quantify abundance.

Targeted species sampling occurs as part of a planned sampling event but is focused on the

collection of a particular species or a subset of species. Assemblage sampling is similarly part of a planned sampling event, but effort is made to record all species observed during the event. Both targeted sampling data and assemblage sampling data are likely to be grouped as sampling-event datasets in GBIF (Table 2), as the sampling effort is documented following a protocol. However, whether these data are grouped as occurrence datasets or sampling-event datasets depends on whether the details and methods of sampling are available.

Under the definition provided in §1.1, most citizen science data are categorized as opportunistic observations. These observations are generally not made as part of an organized sampling effort following specific protocols (such an organized effort would generally constitute community-based monitoring), and there are no sampling methods to report. In contrast, professional data and community-based research data are generally collected as part of an organized sampling effort with a sampling protocol and can be grouped as either occurrence datasets or sampling-event datasets, depending on whether or not event data are published (Table 2).

1.3.2. GBIF-required metadata

GBIF requires metadata in XML format corresponding to the GBIF Metadata Profile, which is based on the Ecological Metadata Language (EML). All GBIF dataset classes require the same set of metadata for each dataset (Table 3).

It is useful to know that when datasets are downloaded individually from GBIF, the XML metadata file is included and metadata fields from this table are automatically added to the occurrence file. When data are selected for download from within a polygon (thereby choosing datasets from multiple studies over a given geographic area), less of the metadata is provided in the occurrence table, but the permanent link to the data selection (provided by GBIF with the data download) allows the user to explore metadata for each individual project.

Table 3. Terms with freshwater-specific definitions, examples and comments for the metadata fields required by GBIF

Term	Definition	Example(s)	Status	Comment
`title`	A descriptive title of the dataset	`Amazon Fish Database`	Required	
`abstract`	Short description of the dataset	`The Amazon Fish Database contains all fish occurrence records in the Amazon Basin.⋯`	Required	Corresponds to "description in the IPT.
`metadataLanguage`	Language in which the metadata is provided	`English`, `German`, etc.	Recommended	Not required for EML, but provides useful information.
`dataLanguage`	Language in which the data is provided	`English`, `German`, etc.	Recommended	Not required for EML, but provides useful information.
`organizationName`	Name of the organization that will be listed as the dataset publisher at gbif.org; the publishing organization is the institution which holds or owns the dataset and is in charge of its contents and maintenance	`UMR EDB`	Required	Corresponds to "publishingOrganization" in the IPT. Can be left empty if you plan to publish your dataset through the FIP/BioFresh IPT

Term	Definition	Example(s)	Status	Comment
`type`	Type of dataset, using one of GBIF's dataset classes	One of `resources metadata`, `checklist`, `occurrence`, `sampling event`	Recommended	Not required for EML, but provides useful information.
`maintenanceUpdateFrequency`	The frequency with which changes are made to the dataset after its first publication	One of `daily`, `weekly`, `monthly`, `biannually`, `annually`, `as needed`, `continually`, `irregular`, `not planned`, `unknown`, `other maintenance period`	Recommended	Corresponds to "updateFrequency" in the IPT.
`licensed`	Licence under which the dataset can be used; GBIF encourages publishers to adopt the least restrictive possible from the three machine readable options; datsets with other licences cannot be registered with GBIF.	`Public Domain (CC0 1.0)` `Creative Commons Attribution (CC-BY 4.0)` `Creative Commons Attribution Non Commercial (CC-BY-NC 4.0)`	Required	Correponds to "dataLicense" in the IPT. More information can be found here: https://www.gbif.org/terms
`contact`	People and organizations that should be contacted to get more information about the dataset	first name: `Max` last name: `Fisher` position: `professor` organization: `Amazon Research Center`	Required	Corresponds to "resourceContact(s)" in the IPT. Please provide first name, last name, position and organization in seperate fields
`creator`	People and organizations who created the dataset	first name: `Moritz` last name: `King` position: `senior scientist` organization: `Amazon Research Center`	Required	Corresponds to "resourceCreator(s)" in the IPT. List creators in priority order. The list will be used to auto-generate the citation of the dataset. Please provide first name, last name, position and organization in separate fields.
`metadataProvider(s)`	People and organizations responsible for producing the metadata of the dataset	first name: `Max` last name `Fisher` position: `professor` organization: `Amazon Research Center`	Recommended	Please provide first name, last name, position and organization in separate fields.

Term	Definition	Example(s)	Status	Comment
`coverage`	Location (bounding box) of the dataset	E.g. a bounding box: `West -72.949; East -49.746; South -9.449; North 2.636`, or description: `Amazon Basin`	Required	Corresponds to "geographicCoverage" in the IPT. Please provide the coordinates for the bounding box in four separate fields. Additonally a description is needed.
`project`	Metadata about the project under which the dataset was produced	`Amazonas Fish Project`	Required	Correponds to "projectData" in the IPT. Please provide at least the title of the project. Add separte fields for identifier, description, funding, study area description or design description, if wanted. More information on the additional fields can be found here: https://ipt.gbif.org/manual/en/ipt/latest/manage-resources#metadata

Term	Definition	Example(s)	Status	Comment
samplingDescription	Metadata about the sampling methods used for data collection, including study extent, sampling description and step description	For example, study extent: Sampling of 24 rivers in the area during the years 2020 to 2022, sampling description: Samples were taken according to the Amazonas Standard Fish Protocol, step description: Fishes were identified to species level according to Ama & Zon 2023; analyses were undertaken with the R package 'zn pack'.	Required	Corresponds to "samplingMethods" in the IPT. Mandatory in situations where data come from a sampling event. Please use separate fields for study extent, sampling description and step description. More information on the additional fields can be found here: https://ipt.gbif.org/manual/en/ipt/latest/manage-resources#metadata
citation	Suggestion for how your dataset should be cited	Fisher, M. & King, M., 2023: Amazon Fish Project 2020-2022. Project Deliverable.	Recommended	Not required for EMP, but provides useful information. When data from a single project are downloaded from GBIF, reference will be provided in a file with the data download. When data from multiple projects are selected via polygon, a DOI will be generated for the full data selection and provided to the user (dataset-specific references available at the DOI).

1.3.3. Metadata required for freshwater data

As outlined in §1.1, there are additional metadata fields that are necessary to describe details about the dataset, including where, when and how the data were collected. Some of this information can be reported within the resource metadata, while other fields may be better associated with the occurrence or sampling-event datasets.

Habitat descriptions should at minimum include the realm and biome to indicate whether observations were made in freshwater and in what water body type. For example, these fields may indicate that a semi-aquatic plant was found adjacent to a pond rather than in the pond. The habitat zone is also required to indicate comparability of data, as for organism groups such as benthic macroinvertebrates and zooplankton, assemblage composition will differ naturally in different lake zones and river mesohabitats.

The amount of sampling method information that is required to make informed decisions about data comparability and data selection also differs among organism groups. In some cases, minimal sampling method information is required for datasets to retain usability and broad compatibility. Additional information is particularly needed for organism groups in which methods or equipment may selectively sample only a subset of size classes or taxa. For example, mesh size of sampling nets is important for zooplankton, benthic macroinvertebrates, and fish, as taxa and age classes may be excluded from larger mesh sizes. For phytoplankton, filter pore size is similarly important to ensure different sets of data are focused on a similar portion of the phytoplankton assemblage. Sampling equipment type is highly relevant for benthic macroinvertebrates and fish and can have an impact on the degree of comparability among samples. For microscopic organism groups, it might also be necessary to report the microscope magnification used when processing samples. For some other organisms groups such as macrophytes, amphibians, reptiles, birds, and mammals, the method itself may provide the most relevant information about sample comparability. Across all organism groups, sampling effort, measured as sampled area, time, catch per unit effort, or other similar measures, can be used to standardize estimates of abundance of taxa, even if sampling methods differ. All of these details improve the utility of dataset published on GBIF and can facilitate large-scale analyses of data from different sources.

2. Specific freshwater dataset publishing considerations

2.1. Dataset preparation and standards

GBIF makes use of the Darwin Core Standard (DwC) to provide a standardized framework for formatting datasets for publication (see What is Darwin Core, and why does it matter?). DwC is a global standard that allows for integration of datasets from different sources through a common format and a number of required and recommended fields.

Though GBIF suggests the required and recommended fields for its four dataset classes, freshwater datasets need a few more specifications to form useful and reusable data for global analyses (see §2). In the following section we provide guidance on all mandatory and recommended terms of the different GBIF dataset classes and how they are best used when preparing freshwater datasets by giving freshwater-specific examples and recommendations. For some terms we suggest using selected terms only to improve consistency among datasets. To make freshwater dataset publication as easy as possible, we have also provided a Excel template for all dataset classes.

What is important for checklist and occurrence datasets is the inclusion of well-structured scientific names for taxa. When datasets are uploaded, the GBIF Backbone Taxonomy is used to check taxon names and update nomenclature as needed for consistency with current naming conventions. However, this process relies on a lack of ambiguity in the provided nomenclature. For example, taxonomic names at the genus or species level that are provided without any higher classifications (e.g. kingdom, class) may end up being misclassified by the taxonomic backbone if the same or similar species names are found in different kingdoms. This would result in inaccurate data. GBIF therefore recommends that users provide as much information as possible about higher classifications. Identification qualifiers (such as "cf."), working names, and common names should not be included in

the `scientificName` field, as these will not align with the taxonomic backbone and should instead be captured in the `identificationQualifier` field. Author names are also an important component of scientific names to avoid misclassification, particularly for species-rich groups such as diatoms. The taxonomic authority (author who first published the species name following international rules) as well as the year of naming should be provided where possible, following the rules of author citation (learn about proper citation rules).

GBIF provides a species matching tool that allows users to normalize their species lists and ensure they match the taxonomic backbone in GBIF for data compatibility.

2.1.1. Checklist datasets

Checklist datasets are not necessarily specific to a location and do not always represent individual observations. However, they can be location specific (like the country-specific Global Register of Introduced and Invasive Species (GRIIS) checklists) and may contain occurrences (example). Checklist datasets may provide detailed information about species geographic distributions or species profiles that describe species characteristics.

Checklist datasets can be used to make it easier to select all freshwater assemblage data from a particular region, if a checklist of the freshwater species in the region is available. For more information on the possible types of checklists and application of checklists, see GBIF Checklist Datasets and Data Gaps.

There are fewer required fields for checklist datasets compared to other dataset classes (Table 4). As noted, higher classifications (e.g. kingdom) are helpful to include in the data to ensure taxa are not misassigned to the wrong name in the taxonomic backbone.

Table 4. Terms with freshwater-specific definitions, examples, and comments for fields GBIF requires or recommends for checklist datasets*.*

Term	Definition	Example(s)	Status	Comment
`taxonID`	A unique identifier for the taxon; may be a global unique identifier or an identifier specific to the dataset.	`8fa58e08-08de-4ac1-b69c-1235340b7001`, `32567`, `ID-fwe-32567`	Required	Ideally, the taxonID is a persistent global unique identifier. As a minimum requirement, it has to be unique within the published dataset.

Term	Definition	Example(s)	Status	Comment
scientificName	The full scientific name of the organism, to the most precise taxonomic rank that is possible to supply, and including authorship and year of the name where applicable/known	Adicella cremisa Malicky, 1972 (genus + specificEpithet + scientificNameAuthorship)	Required	Names should be compliant to the most recent nomenclatural code. This term should not contain identification qualifications (e.g. cf.), which should instead be supplied in the identificationQualifier term. Ideally, the name supplied is at species level or below. Not permitted are, e.g. working names ("Adicella sp.4"), common names ("creamy caddisfly"), or names containing identification qualifiers ("Adicella cf. cremisa").
taxonRank	Taxonomic rank of the most most precise taxonomic level provided in scientificName	One of kingdom, phylum, class, order, family, subfamily, genus, species, subspecies, varietas, forma	Required	Recommended best practice is to use a controlled vocabulary.
kingdom	Full scientific name of the kingdom in which the taxon is classified.	One of Animalia, Plantae, Fungi, Protista, Monera	Recommended	Inclusion of kingdom helps to ensure taxa are not misassigned to the wrong taxonomic name in GBIF.
parentNameUsageID	The taxonID of the next available higher-ranked (parent) entry within the checklist dataset, if higher taxon names are supplied as separate entries in the list	2704173 (GBIF), tsn:41074 (ITIS) etc.	Recommended	This supports the representation of the dataset as a hierarchy, e.g. for the publication of a taxonomy. For Darwin Core Archives, the related record should be present locally in the same archive.

Term	Definition	Example(s)	Status	Comment
acceptedNameUsageID	Within the record of a synonym, the taxonID of the accepted taxon name entry within the checklist dataset, if both synonyms and accepted names are supplied.	2704179 (GBIF), tsn:41107 (ITIS), etc.	Recommended	This supports the representation of synonymy for a taxonomic dataset. For Darwin Core Archives the related record should be present locally in the same archive.

2.1.2. Occurrence datasets

Occurrence datasets provide information about observations of taxa and the locations where they were found (Table 5). Although only coarse location information is required, the recommended best practice is to always provide coordinates (decimalLatitude and decimalLongitude in decimal degrees), a geodeticDatum which will automatically be interpreted to WGS84 when data are published to GBIF, and a measure of the uncertainty around the coordinates (coordinateUncertaintyInMeters).

Occurrence datasets can be provided as presence data (e.g. a "1" for a site where the taxon was observed) or as counts in the field individualCount (Table 5). Counts in this case refer to situations where there is not an effort to estimate the total abundance of the taxon (e.g. by collecting a sample), but instead, numbers of individuals are recorded (tallied) as individuals are encountered. This could include point counts (e.g. in bird surveys, when an observer counts the number of individuals of each species that is viewed or heard) or opportunistic observations. When an effort is made to estimate, for example, abundance, density or biomass as part of targeted or assemblage sampling, these measures should be recorded in the field organismQuantity with units recorded in organismQuantityType (Table 5). Ideally, such occurrence datasets should also be accompanied by sampling-event datasets to provide details on sampling methods. Finally, if effort has been put into recording true absences (e.g. through systematic and/or extensive sampling procedures), then presence or absence can be recorded in the field occurrenceStatus (Table 5). These distinctions will facilitate meta-analysis of data collected in a similar manner or will allow for data to be adjusted as needed for analysis (e.g. all data converted to presence data).

Table 5. Terms with freshwater-specific definitions, examples, and comments for fields GBIF requires or recommends for occurrence datasets

Term	Definition	Example(s)	Status	Comment
occurrenceID	Identifier for the 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 occurrenceID globally unique	AT:BOKU:DAN_0003:8755 (country:institutionCode: sampleCode:speciesID)	Required	This should be a unique identifier for the occurrence, allowing the same occurrence to be recognised across dataset versions as well as through data downloads and use. At the very least the identifier should be unique to the dataset, and ideally a globally unique identifier.

Term	Definition	Example(s)	Status	Comment
basisOfRecord	The specific nature (type) of the individual data record	One of PreservedSpecimen, FossilSpecimen, HumanObservation	Required	Use "PreservedSpecimen", if the species is preserved somewhere, so that checking back is possible. "FossilSpecimen" refers to fossil samples from, e.g. sediment cores. "HumanObservation" refers to observations of living organisms that were not collected (e.g. catch and release or point count).
scientificName	The full scientific name of the organism, to the most precise taxonomic rank that is possible to supply, and including authorship and year of the name where applicable/known	Adicella cremisa Malicky, 1972 (genus + specificEpithet + scientificNameAuthorship)	Required	Names should be compliant to the most recent nomenclatural code (see ICZN Code). This term should not contain identification qualifications (e.g. cf.), which should instead be supplied in the identificationQualifier term. Ideally, the name supplied is at species level or below. Not permitted are, e.g. working names ("Adicella sp.4"), common names ("creamy caddisfly"), or names containing identification qualifiers ("Adicella cf. cremisa").
eventDate	The date or interval during which an event occurred/the occurrence record was collected; not suitable for a time in a geological context (e.g. 5000 BP)	1809-02-12 (12 February 1809)	Required, though year, month, day, or other terms could be used instead	Use the following format: yyyy-mm-dd four-digit year-month-day. Please make sure to provide separate columns for year, month and day as well. Note that the time should not be included as part of this element, please use eventTime instead where required.

Term	Definition	Example(s)	Status	Comment
`eventID` (*if linked to an event*)	Identifier for the set of information associated with an event (something that occurs at a place and time) allowing to link individual occurrences to a specific event; may be a global unique identifier or an identifier specific to the dataset.	`AT:BOKU:DAN_0003:MHS` (country:institutionCode: sampleCode:method)	Required, if event dataset is available	If occurrence has an event dataset (e.g. methods metadata describing the sampling event during which the occurrence was recorded), provide the identifier for the information associated with the event. This can e.g. be entered as the `occurrenceID` without the species code and with the method added.
`taxonRank`	Taxonomic rank of the most most precise taxonomic level provided in `scientificName`.	One of `kingdom`, `phylum`, `class`, `order`, `family`, `subfamily`, `genus`, `species`, `subspecies`, `varietas`, `forma`	Recommended	Recommended best practice is to use a controlled vocabulary.
`kingdom`	Full scientific name of the kingdom in which the taxon is classified.	One of `Animalia`, `Plantae`, `Fungi`, `Protista`, `Monera`	Recommended	Inclusion of kingdom helps to ensure taxa are not misassigned to the wrong taxonomic name in GBIF.
`decimalLatitude`	Geographic latitude (in decimal degrees, using the spatial reference system given in `geodeticDatum`) of the geographic center of a location.	`-41.0983423`	Recommended	Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive. For freshwater datasets, best practice is that coordinates are mandatory, although the GBIF data description indicates that this can be coarse (e.g. country).

Term	Definition	Example(s)	Status	Comment
decimalLongitude	Geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a location.	-121.1761111	Recommended	Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive. For freshwater datasets, best practice is that coordinates are mandatory, although the GBIF data description indicates that this can be coarse (e.g. country).
geodeticDatum	The coordinate system and set of reference points upon which the geographic coordinates given in decimalLatitude and decimalLongitude are based.	EPSG:4326, WGS84, unknown	Recommended	Recommended best practice is to use the EPSG code of the spatial reference system, if known. If no geodetic datum is specified, GBIF's indexing process assumes "WGS84".
coordinateUncertaintyInMeters	The horizontal distance (in meters) from the given decimalLatitude and decimalLongitude describing the smallest circle containing the whole of the location.	30 (reasonable lower limit on or after 2000-05-01 of a GPS reading under good conditions if the actual precision was not recorded at the time) 100 (reasonable lower limit before 2000-05-01 of a GPS reading under good conditions if the actual precision was not recorded at the time)	Recommended	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. Uncertainty can be used to specify the radius of a sampling area around a central point provided in decimalLatitude and decimalLongitude.
countryCode	Standard code for the country in which the location occurs.	AR (Argentina) SV (El Salvador)	Recommended	Recommended best practice is to use ISO 3166-1-alpha-2 country codes. Recommended best practice is to leave this field blank if the location spans multiple entities at this administrative level.

Term	Definition	Example(s)	Status	Comment
individualCount	Number of individuals at the time of the occurrence, indicated as presence or as a count.	1	Recommended	If you have presence data, please indicate "1" here. If a dataset derives from observed counts (e.g. point counts or opportunistic observations of individuals as encountered), enter the counts here. As these are only counts (not density or biomass), there are no units. If the dataset derives from efforts to estimate abundance of particular taxa (targeted sampling) or composition/abundance of different taxa in the assemblage (assemblage sampling), please enter abundance under organismQuantity with "individuals" entered under organismQuantityType. If the dataset derives from standard protocols for measuring and monitoring biodiversity or abundance, please consider to use the sampling-event dataset.

Term	Definition	Example(s)	Status	Comment
organismQuantity	Number or enumeration value for the quantity of organisms as abundance, density, or biomass.	27 (organismQuantity) with "individuals per m²" (organismQuantityType) 12.5 (organismQuantity) with "% biomass" (organismQuantityType) 150 (organismQuantitiy) with "mg dry mass" (organismQuantityType) 800 (organismQuantity) with "individuals" (organismQuantityType)	Recommended	An entry for organismQuantity must have a corresponding organismQuantityType. If you have abundance data, fill in the number individuals and add unit for it in organismQuantityType. If the dataset derives from efforts to estimate abundance of particular taxa (targeted sampling) or composition/abundance of different taxa in the assemblage (assemblage sampling), please enter abundance here with "individuals" entered under organismQuantityType. If the dataset derives from standard protocols for measuring and monitoring biodiversity or abundance, please consider to use the sampling-event dataset.

Term	Definition	Example(s)	Status	Comment
organismQuantityType	Type of quantification system used for the quantity of organisms	"27" (organismQuantity) with `individuals per m²` (organismQuantityType) "12.5" (organismQuantity) with `% biomass` (organismQuantityType) "150" (organismQuantitiy) with `mg dry mass` (organismQuantityType) "800" (organismQuantity) with `individuals` (organismQuantityType)	Recommended	A organismQuantityType must have a corresponding organismQuantity. If you have abundance data, fill in the number individuals in organismQuantity and add unit for it here.
occurrenceStatus	Statement about the presence or absence of a taxon at a location	One of `present` or `absent`	Share if available	For occurrences, the default vocabulary is recommended to consist of present and absent, but the value 'absent' should be used here to record that the sampling did not detect the species, i.e. effort was put into trying to detect the species and it was not detected. For example, if using targeted sampling to estimate species range, non-detections can be identifed here and used to estimate species range using a chosen model for inference, or if a species was previously noted at this location but was not there at the time of the sampling (potentially indicating species loss), then please indicate "absent" here.

2.1.3. Sampling-event datasets

Sampling-event data are structured and systematic surveys (i.e. periodical or singular surveys, routine or one-time environmental monitoring) that must include metadata describing sampling methods (Table 6). Please note that each event dataset consists of two files: the sampling-event dataset and the associated occurrence dataset. The associated occurrence dataset looks like the one in §2.1.2. but needs to be amended with the `eventID` (mandatory; identifying the event and linking the two datasets) and the `occurrenceStatus` (recommended to indicate whether a taxon was present or not detected at a site).

Sampling methods are described in the sampling-event dataset with the field `samplingProtocol`, which provides a name/link to a specific protocol and/or description of the protocol (Table 6). The recommended best practice is to have a separate event for each sampling method used. In addition to describing the protocol, the field `sampleSizeValue` and `sampleSizeUnit` can be used to indicate the spatial or temporal extent of sampling for the described sampling event, as a measure of sampling effort for each event. In addition, the field `samplingEffort` can be used to record the total effort spent on the event, for example, when there were multiple nets, multiple microhabitats sampled, or multiple periods of time over which sampling occurred. Additional details about sampling methods are recommended to be included in the freshwater DwC extensions described in §3.1.

Table 6. Terms with freshwater-specific definitions, examples, and comments for fields GBIF requires or recommends for sampling-event datasets

Term	Definition	Example(s)	Status	Comment
`eventID`	Identifier for the set of information associated with an event (something that occurs at a place and time) allowing to link individual occurrences to a specific event; may be a global unique identifier or an identifier specific to the dataset	`AT:BOKU:DAN_0003:MHS1` (country:institutionCode: sampleCode:method)	Required	If occurrence has an event dataset (e.g. methods metadata describing the sampling event during which the occurrence was recorded), provide the identifier for the information associated with the event. This can e.g. be entered as the `occurrenceID` without the species code and with the method added.

Term	Definition	Example(s)	Status	Comment
eventDate	The date or interval during which an event occurred/the occurrence record was collected; not suitable for a time in a geological context	1809-02-12 (12 February 1809)	Required, though year, month, day, or other terms could be used instead	Use the following format: yyyy-mm-dd four-digit year-month-day. Please make sure to provide separate columns for year, month and day as well. Note that the time should not be included as part of this element, please use eventTime instead where required.
samplingProtocol	Names of, references to, or descriptions of the methods or protocols used during an event	Environment Canada. (2012). Canadian Aquatic Biomonitoring Network Field Manual - Wadeable Streams. Available at http://publications.gc.ca/pub?id=9.696248&sl=0 SS-EN 27 828, Water quality - Methods for biological sampling - Guidance on the handnet sampling of benthic macroinvertebrates net fishing and full/partly following NS-EN 14757	Required	Recommended best practice is describe an event with no more than one sampling protocol/method, and have a separate event for each method used, with occurrences separated by method. If a more detailed description of the method or protocol exists, providing a reference is strongly encouraged.
sampleSizeValue	Numeric value for a measurement of the size (time duration, length, area, or volume) of an individual sample in the sampling event	5 (sampleSizeValue with "metre" as sampleSizeUnit)	Required	A sampleSizeValue must have a corresponding sampleSizeUnit. The sample size can relate to time duration, a spatial length (e.g. of a trawl), an area or a volume.

Term	Definition	Example(s)	Status	Comment
sampleSizeUnit	The unit of measurement of the size (time duration, length, area, or volume) of a sample in a sampling event	minute, metre, square metre	Required	A sampleSizeUnit must have a corresponding sampleSizeValue. Recommended best practice is to use a controlled vocabulary for the sampleSizeUnit.
parentEventID	Identifier for the broader event that groups this and potentially other events; may be a global unique identifier or an identifier specific to the dataset	A1 (parentEventID to identify a transect of samples with its own eventIDs: "A1:1", "A1:2"), AT:BOKU:DAN (country:institutionCode:projectCode)	Recommended	Used in situations where the event is part of an event series. In order to be able to reference a parent event, this event needs to be specified as a separate entry, typically within the same dataset, carrying its own eventID. Refer to the eventID of the parent event in the sample event record to specify the relationship between the two entries.

Term	Definition	Example(s)	Status	Comment
samplingEffort	Measure for the amount of effort expended during an event	40 trap-nights, 10 observer-hours	Recommended	Used to provide evidence of the rigour of the sampling event, e.g. the number of people involved, total area sampled (summed across different sampled microhabitats), or the total number of hours spent on the event (e.g. net set time summed across multiple nets). There is no controlled vocabulary, but the recommendation is to keep this information brief and factual, giving users enough information to compare between sampling events.
locationID	Identifier that links to a set of data describing the sample event location, if available; may be a global unique identifier or an identifier specific to the dataset	http://www.geonames.org/10793757/dnb-6.html	Recommended	If such a reference cannot be meaningfully supplied, consider supplying more location details, e.g. through use of the data elements locality, minimumElevationInMeters, minimumDepthInMeters, stateProvince, locationRemarks etc.

Term	Definition	Example(s)	Status	Comment
decimalLatitude	Geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a location	-41.0983423	Recommended	Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive. Note that a sample event that spans an area rather than a point location should additionally supply the coordinateUncertaintyInMeters to specify the approximate extension of the area.
decimalLongitude	Geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a location	-121.1761111	Recommended	Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive. Note that a sample event that spans an area rather than a point location should additionally supply the coordinateUncertaintyInMeters to specify the approximate extension of the area.
geodeticDatum	The coordinate system and set of reference points upon which the geographic coordinates given in decimalLatitude and decimalLongitude are based	EPSG:4326, WGS84, unknown	Recommended	Recommended best practice is to use the EPSG code of the spatial reference system, if known. If no geodetic datum is specified, GBIF's indexing process assumes "WGS84".

Term	Definition	Example(s)	Status	Comment
coordinateUncertaintyInMeters	The horizontal distance (in meters) from the given decimalLatitude and decimalLongitude describing the smallest circle containing the whole of the location	30 (reasonable lower limit on or after 2000-05-01 of a GPS reading under good conditions if the actual precision was not recorded at the time) 100 (reasonable lower limit before 2000-05-01 of a GPS reading under good conditions if the actual precision was not recorded at the time)	Share if available	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. Uncertainty can be used to specify the radius of a sampling area around a central point provided in decimalLatitude and decimalLongitude.
footprintWKT	An area description, specifying the location of the sample event in well-known text (WKT) markup language	POLYGON ((10 20, 11 20, 11 21, 10 21, 10 20)) (a one-degree bounding box with opposite corners at longitude=10, latitude=20 and longitude=11, latitude=21)	Recommended	A WKT representation of the shape (footprint, geometry) that defines the location. This differs from the point-radius representation that is combined from the elements decimalLatitude, decimalLongitude and coordinateUncertaintyInMeters in that it can define shapes that are not circles. Note that it is possible to supply both a point-radius and a footprintWKT location for the same sample event.

Term	Definition	Example(s)	Status	Comment
footprintSRS	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geometry given in footprintWKT is based	EPSG:4326, unknown	Recommended	Recommended best practice is to use the EPSG code of the SRS, if known. If none of these is known, use the value "unknown". It is also permitted to provide the SRS in Well-Known-Text, especially if no EPSG code provides the necessary values for the attributes of the SRS. Do not use this term to describe the SRS of the decimalLatitude and decimalLongitude, nor of any verbatim coordinates - use the geodeticDatum and verbatimSRS instead.
countryCode	Standard code for the country in which the location occurs	AR (Argentina) SV (El Salvador)	Recommended	Recommended best practice is to use ISO 3166-1-alpha-2 country codes. Recommended best practice is to leave this field blank if the location spans multiple entities at this administrative level.

2.2. Specific requirements for publishing freshwater datasets (freshwater amendments)

Table 7 lists the DwC fields that are useful to include in freshwater datasets to support large-scale data compilation and analysis. In some cases, fields should be included in an extension. Extensions offer a way to include additional information and to provide multiple measurements (e.g., different habitat variables) to link to a single event. Freshwater amendment fields are tagged as:

- **Freshwater required**: as an addition to the GBIF required fields, we recommend required fields for freshwater samples
- **Freshwater recommended**: data that are useful to be reported
- **Freshwater share if relevant**: data that should be reported, but that are only relevant to particular organism groups or habitats (as indicated)

We provide examples for the content of the fields, and in some cases, the full range of controlled values to choose from.

The freshwater amendments include general fields describing the site where the observation was made, such as the water body name, a description of the location, and the elevation (Table 7). The organism group should be included for all GBIF dataset classes, and this information can be captured in the Humboldt Ecological Inventory extension in the field targetTaxonomicScope, which is designed to indicate the taxonomic group that was targeted during sampling. Organism groups should be listed in this field using the assemblage categories described in Table 1.

Current extensions do not have fields that correspond to the freshwater data categories (§1.2.2), either for the type of observation or type of contribution. This information can be captured in the dynamic properties field until specific fields are created.

There are fields recommended for freshwater data that describe the sampled environment, such as the depth of sampling and any abiotic measurements taken in the field, including temperature, pH, and dissolved oxygen (Table 7). Physical and chemical measurements from the habitat should be included in the Extended Measurement or Facts extension to allow multiple measurements to be linked to a single Event dataset.

Other freshwater-specific habitat descriptions, including the biome, ecosystem functional group, lake zone, river mesohabitat, and microhabitat (e.g. sand, gravel, cobble), can be entered in the `habitat` field. This is a multivalue, free-text field, and a proposed structure for this information is provided in Table 7. Although the ultimate goal is to create specific fields for these terms, this field represents the currently available option for recording the information.

Further details about the event time and date are also recommended for inclusion (Table 7). For example, it is recommended that data providers include year, month and day as separate columns in their dataset. This avoids ambiguities that might occur due to regional differences in how year, month, and day are combined into a single field (e.g. confusion of month and day). Furthermore, it is important that all years be entered as four-digit numbers, as historical data (e.g. early 1900s) might be present in GBIF because of digitizing of old records, and full four-digit years ensure that dates are not mishandled.

Additional fields for observation data include the sex and life stage, both of which are conditional based on the organism group (for example, sex can be determined and is relevant for fish, mammals, birds, and decapods; life stage can be determined and is relevant for copepods, benthic macroinvertebrates, fish and birds). The DwC term lifeStage has a controlled vocabulary (see GBIF vocabulary - LifeStage for full list), and this vocabulary does not include all terms that are relevant for freshwater. For example, young of year is not part of the controlled vocabulary, and it is recommended that juvenile be used instead. Similarly, juvenile can be used instead of copepodite, and immature can be used instead of early instar. Additional fields provide details on the identification of the observed taxon, such as references and verification status.

Table 7. Terms, definitions, examples, and comments recommended for inclusion with freshwater datasets. The dataset in which each field should be included (metadata, occurrence or event) is indicated, as is whether fields are required, recommended, or share if relevant on particular organism groups (more information on the specific fields).

Term	Definition	Example(s)	Status	Comment	Inclusion
rightsHolder	A person or organisation owning or managing the rights over the resource	BOKU University (University of Natural Resources and Life Sciences, BOKU Vienna)	Recommended		Metadata
institutionCode	Name or acronym of the institution having custody of the dataset or record	BOKU (University of Natural Resources and Life Sciences, BOKU Vienna) UNB (University New Brunswick)	Recommended		Metadata
collectionID	Identifier for the collection or dataset from which the record was derived	urn:lsid:biocol.org:col:34818, https://www.gbif.org/grscicoll/collection/fbd3ed74-5a21-4e01-b86a-33d36f032d9c	Recommended	For physical specimens, the recommended best practice is to use a globally unique and resolvable identifier from a collections registry such as the Global Registry of Scientific Collections.	Occurrence
informationWithheld	Additional information that exists, but that has not been shared in the given record	location information not given for endangered species	Recommended	A note on possible information that was intentionally not included into the dataset.	Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
dynamicProperties	List of additional measurements, facts, characteristics, or assertions about the record, meant to provide a mechanism for structure content	type of observation:opportunistic observation, type of contribution:community-based research data	Recommended	Recommended best practice is to use a "key:value" encoding schema for a data interchange format (such as JSON). Please use this field for indicating the type of observation (opportunistic observation, targeted sampling, or assemblage sampling data) and type of contribution (professional, community-based research, or citizen science data).	Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
habitat	A description of the habitat in which the event occurred	biome:river, ecosystem functional group:lowland river, lake zone:littoral, river mesohabitat:riffle, microhabitat:sand; for biome and ecosystem functional group, use classifications in the IUCN Global Ecosystem Typology (https://global-ecosystems.org/page/typology); for lake zone use one of: "littoral", "sub-littoral", "pelagic", "profundal"; for river mesohabitat use one of: "riffle", "run", "pool"	Required	Recommended best practice is to use a "key:value" encoding schema for a data interchange format (such as JSON). Please use this field for adding information on e.g. biome, ecosystem functional group, lake zone, river mesohabitat, or microhabitat until specific fields have been created for these categories.	Occurrence, Event

Term	Definition	Example(s)	Status	Comment	Inclusion
targetTaxonomicScope	The taxonomic group(s) targeted for sampling during the dwc:Event.	One of fungi,microbes,benthic algae,phytoplankton,macrophytes,zooplankton,benthic macroinvertebrates,decapods,fish,amphibians,reptiles,birds,mammals	Required	For freshwater, the targeted taxonomic group should be at the level of biologically- and ecologically-similar organisms that are generally grouped together and described as an assemblage. Use broader groups as listed here even if a subset of the group was the focus. For example, use zooplankton even if only cladocerans were sampled. Benthic algae refers to benthic samples of diatoms and other algae, which may include planktonic individuals that have settled.	Checklist, Occurrence (map to Humboldt Ecological Inventory extension)

Term	Definition	Example(s)	Status	Comment	Inclusion
recordedBy	A list (concatenated and separated) of names of people, groups or organizations responsible for recording the original occurrence; the primary collector or observer should be listed first	Jen Lento | Astrid Schmidt-Kloiber	Recommended	Recommended best practice is to separate the values in a list with space vertical bar space, or post (|).	Occurrence
recordedByID	A list (concatenated and separated) of the globally unique identifier for the person, people, groups, or organizations responsible for recording the original Occurrence.	https://orcid.org/0000-0002-8098-4825 | https://orcid.org/0000-0001-8839-5913	Recommended	Recommended best practice is to separate the values in a list with space vertical bar space, or post (|). The primary collector or observer should be listed first.	Occurrence
sex	The sex of the individual(s) represented in the occurrence.	One of female or male or indeterminate	Share if relevant (based on the organism group (Decapoda, fish, mammals, birds))		Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
lifeStage	The age class or life stage of the organism(s) at the time the occurrence was recorded	One of egg, larva, pupa, adult, subadult, juvenile, nymph, immature, nauplius	Share if relevant (based on the organism group (benthic inverte brates, zooplankton - Copepoda, fish, birds))	Controlled vocabulary does not include all terms that are relevant for freshwater. For example, use juvenile instead of young of year or copepodite, and use immature instead of early instar.	Occurrence
occurrenceRemarks	Comments or notes about the occurrence	found dead outside of the water	Recommended		Occurrence
eventType	The nature of the event	sample, observation, bioblitz, expedition, survey, "project`, site visit, biotic interaction	Recommended		Event
eventTime	The time or interval during which an event occurred	14:07-0600 (2:07pm in the time zone six hours earlier than UTC) 13:00:00Z/15:30:00Z (the interval between 1pm UTC and 3:30pm UTC)	Share if available	Recommended best practice is to use a time of day that conforms to ISO 8601-1:2019. Please also add the time zone in relation to UTC.	Event
year	Four-digit year in which the event occurred	2008	Share if available	Please fill this column additionally to the eventDate.	Occurrence Event
month	Month in which the event occurred	01 (January), 10 (October)	Share if available	Please fill this column additionally to the eventDate.	Occurrence Event

Term	Definition	Example(s)	Status	Comment	Inclusion
day	Day of the month on which the event occurred	09, 28	Share if available	Please fill this column additionally to the eventDate.	Occurrence Event
verbatimEventDate	The verbatim original representation of the date and time information for an event	spring 1900, Marzo 2002	Share if available	Please keep your original date/time stamp here (if applicable).	Occurrence Event
eventRemarks	Comments or notes about the event	After the recent rains the river is nearly at flood stage.	Share if available		Event
waterBody	Name of the water body in which the location occurs	River Danube, Lake Constance	Required	Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names.	Occurrence
locality	The specific description of the place, providing regional context to the observation	25 km downstream Vienna	Recommended	Less specific geographic information can be provided in other geographic terms (higherGeography, continent, country, stateProvince, county, municipality, waterBody, island, islandGroup.	Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
minimumElevationInMeters	The lower limit of the range of elevation (altitude, usually above sea level), in metres	100	Recommended	If sampling was done at one altitude only (e.g. no range), enter the actual altitude at which your sample was taken in both this field and in maximumElevationInMeters.	Occurrence
maximumElevationInMeters	The upper limit of the range of elevation (altitude, usually above sea level), in metres	200	Share if available	If sampling was done at one altitude only (e.g. no range), enter the actual altitude at which your sample was taken in this field and in minimumElevationInMeters.	Occurrence
verbatimElevation	The original description of the elevation (altitude, usually above sea level) of the location	100-200 m	Share if available		Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
minimumDepthInMeters	The lesser depth of a range of depth below the local surface, in metres	0.5	Recommended	If sampling took place over a range of depths (e.g. depth-integrated sample or composite sample from water column), enter the minimum depth here and the maximum depth of the range in maxiumDepthInMeters. If sampling was depth-specific (e.g. at one single depth), enter the actual depth in which your sample was taken in this field and in maximumDepthInMeters.	Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
maximumDepthInMeters	The greater depth of a range of depth below the local surface, in metres	1	Share if available	If sampling took place over a range of depths (e.g. depth-integrated sample or composite sample from water column), enter the minimum depth here and the maximum depth of the range in maxiumDepthInMeters. If sampling was depth-specific (e.g. at one single depth), enter the actual depth in which your sample was taken in this field and in minimumDepthInMeters.	Occurrence
verbatimDepth	The original description of the depth below the local surface	0.5 - 1 m	Share if available		Occurrence
identificationQualifier	A brief phrase or a standard term ("cf.", "aff.") to express the determiner's doubts about the identification	cf.	Recommended	Can be used to add doubts, but it is recommend to only report "safe" records	Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
identifiedBy	A name or a list (concatenated and separated) of names of people, groups, or organizations who assigned the taxon to the subject	Hans Malicky, Jen Lento | Astrid Schmidt-Kloiber	Recommended	Recommended best practice is to separate the values in a list with space vertical bar space, or post (|).	Occurrence
identifiedByID	A list (concatenated and separated) of the globally unique identifier for the person, people, groups, or organizations responsible for assigning the Taxon to the subject.	https://orcid.org/0000-0002-8098-4825 | https://orcid.org/0000-0001-8839-5913	Recommended	Recommended best practice is to separate the values in a list with space vertical bar space, or post (|).	Occurrence
identificationReferences	A reference or a list (concatenated and separated) of references (publication, global unique identifier, URI) used in the identification	Malicky, H. 2004 (2nd edtion): Atlas of European Trichoptera. Springer. 1-341.	Recommended	Recommended best practice is to separate the values in a list with space vertical bar space, or post (|). Add a DOI if available.	Occurrence
identificationVerificationStatus	A categorical indicator of the extent to which the taxonomic identification has been verified to be correct	one of verified, unverified, requires verification	Recommended		Occurrence
identificationRemarks	Comments or notes about the identification	Verified by H. Malicky	Share if available (based on identificationVerificationStatus)	Use this field to indicate the person who has verified the identification. You can also use it for describing difficulties with the identification.	Occurrence

Term	Definition	Example(s)	Status	Comment	Inclusion
class	The full scientific name of the class in which the taxon is classified	Mammalia, Insecta	Share if available	Inclusion of class helps to ensure taxa are not misassigned to the wrong taxonomic name in GBIF.	Occurrence
vernacularName	Common or vernacular name	Wassergeistchen, yellow-bellied toad	Recommended		Occurrence
measurementType	The nature of the measurement, fact, characteristic, or assertion	temperature, pH	Share if available	This field is for additional measurements in the field, e.g. abiotic data. A measurementType must have a corresponding measurementValue and measurementUnit.	Event (map to Extended Measurement or Facts extension)
measurementValue	The value of the measurement, fact, characteristic, or assertion	-1, 7.1	Share if available	This field is for additional measurements in the field, e.g. abiotic data. A measurementType must have a corresponding measurementValue and measurementUnit.	Event (map to Extended Measurement or Facts extension)

Term	Definition	Example(s)	Status	Comment	Inclusion
measurementUnit	The unit associated with the measurementValue	°C, g, %	Share if available	This field is for additional measurements in the field, e.g. abiotic data. A measurementType must have a corresponding measurementValue and measurementUnit.	Event (map to Extended Measurement or Facts extension)
measurementMethod	A description of or reference to (publication, URI) the method or protocol used to determine the measurement, fact, characteristic, or assertion	water thermometer, pH meter	Share if available		Event (map to Extended Measurement or Facts extension)
measurementRemarks	Comments or notes accompanying the measurementType	water partly frozen	Share if available		Event (map to Extended Measurement or Facts extension)

3. Future prospects

3.1. Improving freshwater data publishing

The goal of this guide was to provide information on how to set up freshwater data for publishing on GBIF and to offer guidance on fields that have particular importance for freshwater data. While we have recommended that occurrence and sampling-event datasets be amended with other specific DwC fields when publishing freshwater datasets (see §2.2 and Table 7), several of the fields that should be included in the (meta)data as best practices do not currently have appropriate equivalents in DwC. This is a shortcoming that makes it difficult to ensure the required information for freshwater is provided in a consistent manner and in relevant, searchable fields. While we have suggested options for publishing this information in currently existing fields in §2.2 and Table 7, this section lists improvements that could be made in the future.

The type of observation (opportunistic, targeted, or assemblage sampling; see §1.2.2) and type of

contribution (professional, community-based research, or citizen science data; Table 8) are currently recommended to be included in `dynamicProperties`, which is not a searchable field. Specific fields for these data categories would support improved usage of the data in meta-analyses, as they would provide context for the data. While a similar field currently exists in the Humboldt Ecological Inventory extension as inventoryTypes, it is only relevant for data that represent an inventory (see controlled vocabulary that lists relevant types of inventories at term link), and freshwater data collection may not always match the definition of an inventory. A broader term that captures the categories described for freshwater would be beneficial.

Habitat descriptors such as the biome, ecosystem functional group, microhabitat, and freshwater lake zone or river mesohabitat (conditional on the biome) are currently recommended to be included in the field `habitat`, but the creation of specific fields for each of these descriptors would support improved data classification. In the case of biome, ecosystem functional group, and microhabitat, these fields would more broadly apply to data from all realms.

In terms of organism groups addressed, many freshwater researchers work at the assemblage level (e.g., looking for or sampling a combination of taxa within an organism group rather than only one species/genus/family/order) and would benefit from a more effective and efficient way to find relevant data on GBIF. The selection of freshwater assemblage data for analysis still remains a barrier to the use of GBIF data. We have recommended the use of the Humboldt Ecological Inventory extension field targetTaxonomicScope, but it is important to consider whether a field more specifically designed to indicate assemblage group (e.g., benthic macroinvertebrates, phytoplankton) would improve data findability.

Sampling method details are currently captured in a single field in the sampling-event data (`samplingProtocol`). However, we recommend the creation of fields specifically for sampling equipment (e.g. type of net or sampler), mesh size of nets, and sample processing protocols. Each of these details has been shown to be vital to selecting data for meta-analysis (Lento et al. 2019; Goedkoop et al. 2022), and including separate fields for them instead of grouping them all within the protocol field increases the chances that complete information will be provided without ambiguities.

If there is a need for other fields beyond these recommendations, i.e. to capture additional information about the sampling event, there are DwC extensions that may provide guidance on publishing these additional data. For example, Humboldt extension and Darwin Core Measurement or Facts extension.

3.1.1 Recommended terms for improving freshwater data publishing

Most terms that we suggest are urgently needed for other realms as well, which is why most terms do not have a "freshwater" precursor. Those terms that are specific to freshwater or specifically needed to support assessment of freshwater data include "freshwater" as a precursor. For all recommended terms, we have provided freshwater examples.

biome

definition: ecologically similar components of freshwaters with broadly similar features
examples: One of "lakes", "rivers", "wetlands", "groundwater", "adjacent to freshwater", ”interstitial”
status: Required
comment: Please classify your event accordingly based on where the observation was made. If the observation was in a terrestrial habitat adjacent to freshwater, indicate "adjacent to freshwater".
inclusion: Occurrence

ecosystemFunctionalGroup

definition: typology within biomes that classifies ecosystems by joining those with similar ecological conditions.

examples: "lowland rivers", "large lakes", "ponds" (see typology for full list)
status: Required
comment: Please follow the definitions of the IUCN Global Ecosystem Typology for consistency
inclusion: Occurrence

freshwaterLakeZone

definition: typology within the lake biome that classifies this ecosystem into habitat zones
examples: One of "littoral", "sub-littoral", "pelagic", "profundal"
status: Share if available (based on biome)
inclusion: Occurrence

freshwaterRiverMesohabitat

definition: typology within the river biome that classifies this ecosystem into habitat zones
examples: One of "riffle", "run", "pool"
status: Share if available (based on biome)
inclusion: Occurrence

typeOfContribution

definition: category based on the type of data contribution
examples: one of "professional data"; "community-based research data"; "citizen science data"
status: Required *inclusion*: Occurrence

typeOfObservation

definition: category of occurrence and sampling-event data based on the type of observation recorded
examples: one of "opportunistic observation"; "targeted sampling"; "assemblage sampling"
status: Required
inclusion: Occurrence

freshwaterOrganismGroup

definition: collections of biologically and ecologically similar organisms that are generally grouped together and described as an assemblage
examples: "fungi"; "microbes"; "benthic algae"; "phytoplankton"; "macrophytes"; "zooplankton"; "benthic invertebrates"; "decapods"; "fish"; "amphibians"; "reptiles"; "birds"; "mammals"
status: Required
inclusion: Occurrence, Checklist

season

definition: indicates the season in which a sample was collected
examples: one of "winter"; "spring"; "summer"; "autumn"; "wet"; "dry"
status: Recommended
inclusion: Occurrence

samplingEquipment

definition: name or description of the sampling instrument that was used for collecting the organisms, including mesh sizes where applicable
examples: "light trap"; "500 µm mesh kick net"; "80 µm mesh plankton net"; "6.25, 8, 10, 12.5, 15.5, 19.5, 24, 29, 35, 43, 55 mm mesh gill net"
status: Required
comment: It is important that both the sampling equipment and the net mesh size (if nets were used) are provided, as mesh size gives an indication of the size of organisms retained.
inclusion: Occurrence

sampleProcessing

definition: name or description of the sample processing protocol (e.g. procedures followed after sample collection to sort and identify taxa)
examples: "20x microscope magnification"; "subsampled with Marchant box until 300 organisms identified - abundance estimated based on the number of cells processed"; "samples filtered on 45 µm pore size filter paper prior to identification"; "samples mounted on slide and random transects identified under 500x inverted microscope until 300 individuals filaments or colonies counted and identified"
status: Share, if available (based on freshwaterOrganismGroup (fungi, microbes, benthic algae, phytoplankton, zooplankton, benthic macroinvertebrates)
comment: Provide as much detail as possible about procedures followed in the lab to process and idenfity samples, including any sub-sampling procedures, sample treatment/staining, slide mounting, and magnifications used. If relevant, include a reference to the protocol used.
inclusion: Occurrence

3.2. Freshwater data tagging

Data portals such as GBIF.org offer a great variety of data but still show limitations in terms of freshwater species. This relates mostly to the fact that freshwater species and freshwater datasets are not specifically tagged and therefore hard to find among millions of terrestrial and marine species and occurrence records. Looking for entire freshwater datasets (e.g. recordings of whole assemblages) often requires searching for specific freshwater species, which is a time-consuming task.

Freshwater datasets that published through a GBIF node or uploaded using IPT software should therefore be tagged as “freshwater” to make the dataset more visible to the freshwater community. This can be done by allocating the specific dataset to the Freshwater Network during the publication process, after registering it with GBIF.

3.3. Importance of reliable taxonomy

The use of organismal names is ubiquitous in a wide range of research, environmental management and policy domains. Expert-curated taxonomic databases and tools to query these data are therefore essential for ensuring the quality of biological data. Species information systems for monitoring status and trends of biodiversity (e.g. GBIF) and those dealing with policy concerns (e.g. European Water Framework Directive, Natura 2000 species, commercial, invasive alien species and pest species) benefit from such high-quality tools and databases ensuring the interoperability of data. The last global taxonomic assessment of freshwater species dates back to the year 2008 (Balian et al. 2008). This Freshwater Animal Diversity Assessment (FADA) comprises a global, extensive set of taxa lists for freshwater animal groups (125,530 described species and 11,388 genera). However, these lists were never fully integrated into GBIF. As taxonomy is a living scientific discipline where new taxa are being described and existing taxa are being placed in new taxonomic positions, the FADA database is currently being updated with the ultimate goal to serve as up-to-date freshwater animal taxonomic backbone for GBIF as well as for other international infrastructures like the Catalogue of Life or the data portal of the Freshwater Information Platform (FIP), which is currently rebuilt as “FIPbio”.

3.4. Interaction and linkages between data infrastructures

Species observed in freshwaters are typically good indicators of the health and status of these ecosystems and are therefore frequently analyzed as part of ecological monitoring programs. The

biodiversity data generated during such monitoring routines, in combination with data from other ecological studies in freshwaters, can form an invaluable source of information to support sustainable management and conservation of aquatic ecosystems. However, a large amount of data still remains scattered on individual researchers' computers and institute servers as well as in different data infrastructures depending on the type of data. This has led to a variety of calls for intense freshwater data mobilization activities as well as a better and more connected infrastructure landscape where data publishing follows the FAIR Principles (e.g. Van Rees et al 2021; Maasri et al. 2022).

While findability through web search seems to be less of a pressing issue, accessibility of data, interoperability between data infrastructures and reusability still play a major role. This guide seeks to streamline data publication in terms of data reuse and accessibility by making them available through GBIF and by including a specific set of fields for freshwater-relevant information. Alternatively, other publishing platforms that guarantee exchange with GBIF like the data portal of the Freshwater Information Platform (FIPbio) or the South African Freshwater Biodiversity Information System, which both focus on freshwater data, can be used. In any case, we advise that priority be given to infrastructures that provide biogeographic information and are well-connected with GBIF, rather than using simple repositories for data publishing.

Once freshwater data can be more easily filtered within GBIF (through respective tagging of freshwater species), it will be possible to more easily assess global freshwater taxa coverage and to actually identify data and/or research gaps in freshwater biodiversity.

Glossary

assemblage

A collection of biologically and ecologically related taxa within a community (i.e. all individuals in an organism group), following the definition of Fauth et al. 1996.

benthic

The ecological region at the bottom of a water body (such as a lake, river, or ocean) or a wetland. Also used to refer to organisms that live on the bottom of a body of water or wetland, whether on or in the substrate.

biome

According to the IUCN Global Ecosystem Typology there are three biomes in freshwater: 1) rivers and streams; 2) lakes; and 3) artificial wetlands. Groundwater, brackish water, palustrine wetlands, and coastal systems are grouped within transitional realms.

DwC

Darwin Core data exchange standard

extension

extensions provide a way to capture additional information outside of the DwC core fields, including additional fields and the ability to map one to many relationships. GBIF has a number of registered extensions.

ecosystem functional group

According to the IUCN Global Ecosystem Typology, ecosystem functional groups describe ecological conditions within the realms e.g. permanent, seasonal or episodic/ephemeral; freeze-thaw; upland or lowland; large or small; etc.

FIP

Freshwater Information Platform

IPT

Integrated Publishing Toolkit, localized repository software developed and maintained by GBIF for managing and publishing open biodiversity data

lake zone

Lake habitat zones defined based on depth and characteristics related to light penetration, oxygen levels, substrates and temperature; includes littoral, sub-littoral, profundal, and pelagic.

microhabitat

Fine-scale habitat differences within a water body, such as areas with different substrate composition.

organism group

Collections of biologically and ecologically similar organisms that are generally grouped together and described as an assemblage, e.g. benthic invertebrates.

realm

According to the IUCN Global Ecosystem Typology, there are five realms: 1) terrestrial; 2) freshwater; 3) marine; 4) subterranean; and 5) atmospheric components of the biosphere, as well as transitional zones between realms.

river mesohabitat

Zones in a river differentiated based on types of flow, including riffles (fast-flowing, shallow, rocky areas), runs (deeper, fast-flowing areas), and pools (areas of slow-flowing or standing water).

References

- Balian, E. V., Lévêque, C., Segers, H. & Martens, K, eds. (2008) Freshwater animal diversity assessment. Hydrobiologia 595/Developments in Hydrobiology 198. Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-1-4020-8259-7
- Beno M, Figl K, Umbrich J & Polleres A (2017) Perception of Key Barriers in Using and Publishing Open Data. JeDEM - eJournal of eDemocracy and Open Government 9(2): 134-165. https://doi.org/10.29379/jedem.v9i2.465
- Comte L & Olden JD (2018) Fish dispersal in flowing waters: A synthesis of movement- and genetic-based studies. Fish and Fisheries 19(6): 1063-1077. https://doi.org/10.1111/faf.12312
- Carvajal-Quintero J, Villalobos F, Oberdorff T et al. (2019) Drainage network position and historical connectivity explain global patterns in freshwater fishes' range size. Proceedings of the National Academy of Sciences 116(27): 13434-13439. https://doi.org/10.1073/pnas.1902484116
- Darwall W, Bremerich V, De Wever A et al. (2018) The *Alliance for Freshwater Life*: A global call to unite efforts for freshwater biodiversity science and conservation. Aquatic Conservation: Marine and Freshwater Ecosystems 28(4): 1015-1022. https://doi.org/10.1002/aqc.2958
- Dudgeon D, Arthington AH, Gessner MO et al. (2006) Freshwater biodiversity: importance, threats, status and conservation challenges. Biological Reviews 81(2): 163-182. https://doi.org/10.1017/S1464793105006950
- Fauth JE, Bernardo J, Camara M, Resetarits WJ, Van Buskirk J & McCollum SA (1996) Simplifying the Jargon of Community Ecology: A Conceptual Approach. The American Naturalist 147(2): 282-286. http://www.jstor.org/stable/2463205
- GEO BON & FWBON (2022) Inland Waters in the Post-2020 Global Biodiversity Framework. https://geobon.org/science-briefs/
- Gido KB, Whitney JE, Perkin JS & Turner TF (2016) Fragmentation, connectivity and fish species persistence in freshwater ecosystems. In Closs GP, Krkosek M & Olden J, eds. Conservation of freshwater fishes. Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9781139627085 Goedkoop W, Culp JM, Christensen T et al. (2022) Improving the framework for assessment of ecological change in the Arctic: A circumpolar synthesis of freshwater biodiversity. Freshwater Biology 67(1): 210-223. https://doi.org/10.1111/fwb.13873
- Gotelli NJ & Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4(4): 379-391. https://doi.org/10.1046/j.1461-0248.2001.00230.x
- Harper M, Mejbel HS, Longert D et al. (2021) Twenty-five essential research questions to inform the protection and restoration of freshwater biodiversity. Aquatic Conservation: Marine and Freshwater Ecosystems 31(9): 2632-2653. https://doi.org/https://doi.org/10.1002/aqc.3634
- Jarvis SG, Mackay EB, Risser HA et al. (2023) Integrating freshwater biodiversity data sources: Key challenges and opportunities. Freshwater Biology 68(9): 1479-1488. https://doi.org/10.1111/fwb.14143
- Ledger SEH, Loh J, Almond R et al. (2023) Past, present, and future of the Living Planet Index. npj Biodiversity: 2(12). https://doi.org/10.1038/s44185-023-00017-3
- Lento J, Goedkoop W, Culp J et al. (2019) State of the Arctic Freshwater Biodiversity Report. Akureyri, Iceland: Conservation of Arctic Flora and Fauna (CAFF) International Secretariat. https://caff.is/freshwater
- Mack HR, Conroy JD, Blocksom KA, Stein RA & Ludsin SA (2012) A comparative analysis of zooplankton field collection and sample enumeration methods. Limnology and Oceanography: Methods 10(1): 41-53. https://doi.org/https://doi.org/10.4319/lom.2012.10.41

- Maasri A, Jähnig SC, Adamescu MC et al. (2022). A global agenda for advancing freshwater biodiversity research. Ecology Letters 25(2): 255-263. https://doi.org/10.1111/ele.13931
- Pansera M, Granata A, Guglielmo L, Minutoli R, Zagami G & Brugnano C (2014) How does mesh-size selection reshape the description of zooplankton community structure in coastal lakes? Estuarine, Coastal and Shelf Science 151: 221-235. https://doi.org/https://doi.org/10.1016/j.ecss.2014.10.015
- Reid AJ, Carlson AK, Creed IF et al. (2019) Emerging threats and persistent conservation challenges for freshwater biodiversity. Biological Reviews 94(3): 849-873. https://doi.org/10.1111/brv.12480
- Sarremejane R, Cid N, Stubbington R et al. (2020) DISPERSE, a trait database to assess the dispersal potential of European aquatic macroinvertebrates. Scientific Data 7(1): 386. https://doi.org/10.1038/s41597-020-00732-7
- Schmidt-Kloiber A & De Wever A (2018) Biodiversity and freshwater information systems. In Schmutz S. & Sendzimir J., eds. Riverine Ecosystem Management: Science for Governing Towards a Sustainable Future. Aquatic Ecology Series, vol 8. Springer Open. https://doi.org/10.1007/978-3-319-73250-3_20
- Sholler D, Ram K, Boettiger C & Katz DS (2019) Enforcing public data archiving policies in academic publishing: A study of ecology journals. Big Data & Society 6(1): 205395171983625. https://doi.org/10.1177/2053951719836258
- Tickner D, Opperman JJ, Abell R et al. (2020) Bending the Curve of Global Freshwater Biodiversity Loss: An Emergency Recovery Plan. BioScience 70(4): 330-342. https://doi.org/10.1093/biosci/biaa002
- Turak E, Walters M, Pienaar M & van Deventer H (2023) Final report to GBIF on the FWBON-GBIF freshwater biodiversity data mobilization project. February 2023.
- Van Rees CB, Waylen KA, Schmidt-Kloiber A et al. (2021) Safeguarding freshwater life beyond 2020: Recommendations for the new global biodiversity framework from the European experience. Conservation Letters: 14(1): e12771. https://doi.org/10.1111/conl.12771
- Ward JV (1998) Riverine landscapes: Biodiversity patterns, disturbance regimes, and aquatic conservation. Biological Conservation 83(3): 269-278. https://doi.org/https://doi.org/10.1016/S0006-3207(97)00083-9
- Wilkinson MD, Dumontier M, Aalbersberg IJ et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1): 160018. https://doi.org/10.1038/sdata.2016.18