Publishing sample data using the GBIF IPT

Latest published version: <http://links.gbif.org/ipt-sample-data-primer>

Contents

[Summary 1](#_Toc398301841)

[What is sample-based data? 1](#_Toc398301842)

[Expressing sample data in Darwin Core 2](#_Toc398301843)

[sampleSize and sampleSizeUnit 2](#_Toc398301844)

[quantity and quantityType 2](#_Toc398301845)

[Darwin Core Archive 3](#_Toc398301846)

[Examples 5](#_Toc398301847)

[Freshwater invertebrate survey 5](#_Toc398301848)

[Brackish water invertebrates survey 5](#_Toc398301849)

[Macrophyte survey 6](#_Toc398301850)

[Lepidoptera survey I 7](#_Toc398301851)

[Lepidoptera survey II 7](#_Toc398301852)

[Reef fish survey 8](#_Toc398301853)

# Summary

This document describes how the Darwin Core vocabulary, extended with a small number of additional terms, can be used in a Darwin Core Archive to express sample-based data sets.

# What is sample-based data?

Sample-based data is a type of data available from thousands of environmental, ecological, and natural resource investigations. These can be one-off studies or monitoring programmes. Such data are usually quantitative, calibrated, and follow certain protocols so that changes and trends of populations can be detected. This is in contrast to opportunistic observation and collection data, which today form a significant proportion of openly accessible biodiversity data.  Sample-based data are often not shared because the underlying protocols have been hard to encode in a standardised way.

# Expressing sample data in Darwin Core

The Darwin Core vocabulary already provides a rich set of terms, organised into several classes (e.g., Occurrence, Event, Location, Taxon, Identification). Many of these terms are relevant for describing sample-based data. Synthesising several sources of input (GBIF organised workshop on sample data, May 2013[[1]](#footnote-1); discussions on the TDWG mailing list; discussions on the EU BON mailing list), a small set of terms relating to sample data were identified as essential, some of which are already present in the DwC vocabulary. These terms are:

1. **eventID**: an identifier for the set of information associated with an Event; may be a global unique identifier or an identifier specific to the data set.
2. **samplingProtocol**:the name of, reference to, or description of the method or protocol used during a sampling event.
3. **sampleSize**: a numeric value for the time duration, length, area or volume involved in the sampling.
4. **sampleSizeUnit**: the unit of measurement used for sampling, e.g., minute, hour, day, metre, metre^2, metre^3.
5. **quantity**: the number or enumeration value of the quantityType per sample.
6. **quantityType**: the entity being referred to by quantity, e.g., individuals,  a percentage  (e.g., species, biomass, biovolume),  a scale type such as Braun Blanquet or Domin.

Four of the six terms are new and are required to be used in pairs: sampleSize together with sampleSizeUnit, quantity together with quantityType.

## sampleSize and sampleSizeUnit

The value of sampleSize is a number. The value of sampleSizeUnit could be restricted to use only SI units/derived units or other non-SI units accepted for use within the SI (e.g. minute, hour, day, litre).

Table . sampleSize and sampleSizeUnit must be used together, e.g., 3 square metres, or 1 litre.

|  |  |
| --- | --- |
| **sampleSize**: a numeric value for the time duration, length, area or volume involved in the sampling. | **sampleSizeUnit**:  the unit of measurement used for sampling, e.g., minute, hour, day, metre, metre^2, metre^3. |
| 2 | hour |
| 3 | m^2 |
| 17 | km |
| 1 | litre |

## quantity and quantityType

The terms, quantity and quantityType, are also required to be used as a pair. Table 1 lists some example values. The value of quantity is a number or enumeration while the value for quantityType (i.e., the entity being measured) is expected to be drawn from a small controlled vocabulary. Typical values include: Individuals, % Biomass, % Biovolume, % species, % coverage, BraunBlanquetScale, DominScale.

Table . quantity and quantityType must be used together, e.g., a count of 14 individuals, or a code value “r” on the Braun Blanquet scale.

|  |  |
| --- | --- |
| **quantity**: the number or enumeration value of the quantityType per sample | **quantityType**:  the entity being referred to by quantity, e.g., individuals,  a percentage  (e.g., species, biomass, biovolume),  a scale type |
| 14 | Individuals |
| r | BraunBlanquetScale |
| 0.4 | %Species |
| 31 | %Biomass |

# Darwin Core Archive

The GBIF Integrated Publishing Toolkit (IPT)[[2]](#footnote-2) uses a particular text format called Darwin Core Archive (DwC-A)[[3]](#footnote-3). DwC-A (Figure 1) imposes a relatively simple, one-to-many relational model (i.e., star schema) in which a row in a (central) *core* table can be linked to many rows in one or more (surrounding) *extension* tables. Table column headers typically map to Darwin Core terms although terms from other vocabularies can also be used. Currently, GBIF uses two cores: Taxon[[4]](#footnote-4) and Occurrence[[5]](#footnote-5). Thus, e.g., a row in the Taxon core typically including a Linnaean binomial could be linked to several rows in a “VernacularNames” extension, each row providing a different vernacular name for the species named in the row in the core. The core and extension files are compressed into an archive together with a descriptor file (meta.xml) which describes the mappings, and a data set level metadata document in Ecological Metadata Language (EML.xml).



Figure 1. The components of a Darwin Core Archive.

In order to encode sample-based data, we here propose a third, new core, the **Event**[[6]](#footnote-6) (i.e. sampling event) core, and an associated **Occurrence** extension, which is identical in structure to the Occurrence core[[7]](#footnote-7) but includes two additional terms, quantity and quantityType (Figure 2). In the core table, each row is a sample identified by a unique eventID and other columns describe sampling protocol, sampleSize, sampleSizeUnit, date, location, etc. The rows in the Occurrence extension table reference a sampling event in the core (via eventID) and list the taxa in the sample together with associated measurement (quantity and quantityType). Following the one-to-many star schema, one Event row can link to many Occurrence rows.



Figure . A sampling event uses a particular samplingProtocol , sampleSize, sampleSizeUnit etc. and can record one or more taxa, each of which has a measurement (quantity and quantityType) associated with it.

The Event[[8]](#footnote-8) core elements are mainly drawn from the DwC classes Event, Location and Geological Context with the addition of the two new terms sampleSize and sampleSizeUnit (Table 3). The Occurrence extension draws from the Occurrence, Taxon and Identification classes with the addition of the two new terms quantity and quantityType. For reasons of consistency, the Occurrence extension includes all terms found in the Occurrence core. Thus Event, Location and Geological Context terms are also listed for the Occurrence extension but are actually redundant.

Table 3. Placement of the five sample related terms in the Event core and Occurrence extension.

|  |  |
| --- | --- |
| **Event Core** | eventID, samplingProtocol, sampleSize\*, sampleSizeUnit\* |
| **Occurrence Extension** | eventID, quantity\*, quantityType\* |

The \* symbol indicates proposed new terms.

# Examples

Following are some examples of typical sample data sets. In each case, the key fields in the Event core and Occurrence extension are provided. For some examples, additional extensions such as Relevé and measurement-or-fact are also included.

## Freshwater invertebrate survey

**Core (Event) table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| eventID | samplingProtocol | sampleSize | sampleSizeUnit | eventDate | location | decimalLatitude | decimalLongitude | … |
| C\_1428 | AQEM | 1.25  | m2 | 21/06/2006 | Kinzig O3 Rothenbergen | 50.18689 | 9.100369 |  |
| B\_1538 | AQEM | 1.25  | m2 | 11/06/2008 | Kinzig W1 Bulau | 50.1316 | 8.9657 |  |

**Extension (Occurrence) table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| eventID | scientificName | quantity | quantityType | … |
| C\_1428 | *Baetis rhodani* | 14 | individuals |  |
| C\_1428 | *Ephemera danica* | 15 | individuals |  |
| C\_1428 | *Gyraulus albus* | 2 | individuals |  |
| B\_1538 | *Serratella ignita* | 318 | individuals |  |

**Explanation**

***Ephemera danica*** : A total of 14 individuals from 1.25 sq metres were obtained in this sampling event. The derived individuals per sq metre count is 11.2 (14/1.25).

## Brackish water invertebrates survey

**Core (Event) table**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| eventID | samplingProtocol | samplesize | sampleSizeUnit | startDayOfYear | end DayOfYear | year | location | decimalLatitude | decimalLongitude | … |
| IA1 | hand operated van Veen grab | 0.04  | m2 | 147 | 154 | 1995 | Gialova lagoon | 36.9564 | 21.6661 |  |
| IA3  | hand operated van Veen grab | 0.04  | m2 | 147 | 154 | 1995 | Gialova lagoon | 36.9564  | 21.6661 |  |

**Extension (Occurrence) table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| eventID | scientificName | quantity | quantityType | … |
| IA1 | *Abra ovata* | 57 | individuals |  |
| IA3  | Bittium reticulatum | 113 | individuals |  |

**Extension (Measurement-or-Fact) table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| eventID | measurementType | measurementValue | measurementUnit | measurementRemarks | … |
| IA1 | Tmp (sed) | 21.5 | Degree C | temperature at the bottom surface |  |
| IA1  | Rdx(sed)0 | 170 | mv | Eh value at the bottom surface (0cm) |  |

**Explanation**

***Abra ovata*** : A total of 57 individuals from 0.04 sq metres were obtained in sampling event IA1.

Each event can also have measurements or facts associated with it, e.g., environmental measurements like sediment temperature and redox potential (Eh).

## Macrophyte survey

**Core (Event) table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| eventID | samplingProtocol | sampleSize | sampleSizeUnit | eventDate | location | decimalLatitude | decimalLongitude | … |
| 1001 | Braun Blanquet | 100  | m2 | 09/08/2012 | Kinzig O3 Rothenbergen | 50.18689 | 9.100369 |  |

**Extension (Occurrence) table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| eventID | scientificName | quantity | quantityType | … |
| 1001 | *Acer pseudoplatanus* | r | BraunBlanquetScale  |  |

**Extension (Relevé) table**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| eventID | syntaxonCode | inclination | coverTotal | treesCover | coverShrubs | highTreeLayerHeight | highHerbLayerMeanHeight | mossesIdentified | … |
| 1001 | 843200 | 40 | 100 | 95 | 50 | 25 | 40 | Y |  |

**Explanation**

***Acer pseudoplatanus***: In the 100 sq metres surveyed, the abundance of the species was reported as “r” on the Braun Blanquet scale.

Additional vegetation plot measurements such as vegetation community type (syntaxon) % coverage values that are typical of TurboVeg type databases are captured in a Relevé (vegetation-plot) extension.

## Lepidoptera survey I

**Core (Event) table**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| eventID | samplingProtocol | samplesize | sampleSizeUnit | startDayOfYear | end DayOfYear | year | location | decimalLatitude | decimalLongitude | ... |
| 2320 | Jalas-model light trap with 160W ML matt lamp | 16  | day | 164 | 180 | 1999 | Kungsmarken | 55.72 | 13.28 |  |

**Extension (Occurrence) table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| eventID | scientificName | quantity | quantityType | … |
| 2320 | *Opisthograptis luteolata* | 11 | Individuals  |  |

**Explanation**

***Opisthograptis luteolata***: 11 individuals were observed over the sampling period of 16 days. The derived number of individuals per day is 0.68 (11/16).

## Lepidoptera survey II

**Core (Event) table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| eventID | samplingProtocol | sampleSize | sampleSizeUnit | eventDate | location | decimalLatitude | decimalLongitude | … |
| 1014-tr023m | Pollard walks | 250  | m2 | 2012-10-11 | Ramat Hanadiv botanik garden | 32.553191 | 34.947492 |  |
| 1012-tr006-s5  | Pollard walks | 250  | m2 | 2012-05-02 | Carmel Hurshan haarbaim | 32.75789805  |  35.02697333 |  |

**Extension (Occurrence) table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| eventID | scientificName | quantity | quantityType | … |
| 1014-tr023m | *Pieris rapae* | 1 | individuals |  |
| 1014-tr023-s5 | *Maniola telmessia* | 2 | individuals |  |

**Extension (Measurement-or-Fact) table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| eventID | measurementType | measurementValue | measurementUnit | measurementRemarks | … |
| 1014-tr023m | Temp | 20 | Degree C |  |  |
| 1014-tr023m | Wind speed | light |  |  |  |
| 1014-tr023m | Cloudiness | 0 | Level 1 of 8 |  |  |
| 1014-tr023m | AvgAltitude | 10 | m | Average altitude |  |

**Explanation**

***Pieras rapae*** : A total of 1 individual from 250 sq metres was obtained in this sampling event.

Several environmental measurements (e.g., temperature, wind speed, cloudiness) are included in a measurement-or-facts extension.

## Reef fish survey

**Core (Event) table**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| eventID | samplingProtocol | sampleSize | sampleSizeUnit | eventDate | location | decimalLatitude | decimalLongitude | … |
| 506003329 | Reef Life Survey methods | 500  | m2 | 2006-09-02 | Cocos Islands | 5.56187  | -87.04693 |  |
| 57003326 | Reef Life Survey methods | 500  | m2 | 2006-12-11 | Panama Bight | 4.008553 | -81.605377 |  |

**Extension (Occurrence) table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| eventID | scientificName | quantity | quantityType | … |
| 506003329 | *Acanthurus nigricans* | 42 | individuals |  |
| 506003329 | *Acanthurus xanthopterus* | 1 | Individuals |  |
| 506003329 | *Aulostomus chinensis* | 4 | Individuals |  |
| 57003326 | *Axoclinus cocoensis* | 1 | individuals |  |

**Explanation**

***Aulostomus chinensis***: A total of 4 individuals from 500 sq metres were obtained in this sampling event.

1. <http://www.standardsingenomics.org/index.php/sigen/article/view/sigs.4898640> [↑](#footnote-ref-1)
2. <http://www.gbif.org/ipt> [↑](#footnote-ref-2)
3. <http://rs.tdwg.org/dwc/terms/guides/text/index.htm> [↑](#footnote-ref-3)
4. <http://rs.gbif.org/core/dwc_occurrence.xml> [↑](#footnote-ref-4)
5. <http://rs.gbif.org/core/dwc_taxon.xml> [↑](#footnote-ref-5)
6. Note that Occurrence can also be used as the core for sample data. This allows additional information relating to the taxon occurrences to be captured in a MeasurementOrFact extension. Using Event as the core is preferable if a plot or site is the main focus of a study. See the GBIF workshop report referenced on page 1 for discussion of an alternative sample core based on the Occurrence core. [↑](#footnote-ref-6)
7. <http://rs.gbif.org/core/dwc_occurrence.xml> [↑](#footnote-ref-7)
8. <http://rs.gbif.org/sandbox/core/dwc_event.xml> [↑](#footnote-ref-8)