Publishing sample data using the GBIF IPT

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

Contents

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]; 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. **parentEventID***: An event identifier for the super event which is composed of one or more sub-sampling events; the value must refer to an existing eventID. If the identifier is local it must exist within the given dataset. Example: "A1" identifying the main Whittaker Plot in nested samples, each with their own eventID (e.g., "A1:1", "A1:2").
3. **samplingProtocol**: the name of, reference to, or description of the method or protocol used during a sampling event.
4. **sampleSize***: a numeric value for the time duration, length, area or volume involved in the sampling event. (new)
5. **sampleSizeUnit***: the unit of measurement used in the sampling event, e.g., minute, hour, day, metre, square metre, cubic metre.
6. **organismQuantity***: a number or enumeration value for the quantity of organisms. Use with organismQuantityType to indicate the type of entity that is being quantified.
7. **organismQuantityType***: the type of entity to which the number or enumeration value reported for the quantity of organisms in organismQuantity refers.

**Indicates proposed new term*

Five of the seven terms are new. Four of them are required to be used in pairs: sampleSize with sampleSizeUnit, organismQuantity with organismQuantityType.

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 1).

Table 1. 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 event.	**sampleSizeUnit**: the unit of measurement used in the sampling event, e.g., minute, hour, day, metre, square metre, cubic metre.
2	hour
3	m2
17	km
1	litre

[1] http://www.standardsingenomics.org/index.php/sigen/article/view/sigs.4898640

organismQuantity and organismQuantityType

The terms, organismQuantity and organismQuantityType, are also required to be used as a pair. Table 2 lists some example values. The value of organismQuantity is a number or enumeration, e.g., "27" for an organismQuantityType "individuals", "12.5" for an organismQuantityType "%biomass", or "r" for an organismQuantityType "BraunBlanquetScale". The value of organismQuantityType (i.e., the entity being measured) is expected to be drawn from a small controlled vocabulary with terms such as "Individuals", "%Biomass", "%Biovolume", "%Species", "%Coverage", "BraunBlanquetScale", "DominScale". Examples when combined with organismQuantity values: + on DominScale; 5 on BraunBlanquetScale; 45 for %Biomass.

Table 2. organismQuantity and organismQuantityType must be used together, e.g., a count of 14 individuals, or a code value "r" on the Braun Blanquet scale.

organismQuantity: A number or enumeration value for the quantity of organisms. Use with organismQuantityType to indicate the type of entity that is being quantified.	**organismQuantityType**: The type of entity to which the number or enumeration value reported for the quantity of organisms in organismQuantity refers.
14	Individuals
r	BraunBlanquetScale
0.4	%Species
31	%Biomass

eventID and parentEventID

The terms eventID and parentEventID, while related, are not required to be used as a pair. eventID is used to hold an identifier for a single sampling event. Sampling events can be related to each other (e.g., nested samples) via a common parent identifier. For example, several sub-sampling events within a Whittaker Plot[2] each with their own eventID (e.g., "A1:1", "A1:2") would share a common parentEventID (e.g., "A1") thus enabling them to be linked together easily (see Table 4 and Figure 3).

Further information on the nature of the relationship (e.g. part of a monitoring series) can be described in the project section of the accompanying metadata.

Darwin Core Archive

The GBIF Integrated Publishing Toolkit (IPT)[3] uses a particular text format called Darwin Core Archive (DwC-A)[4]. 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[5] and

[2] http://www.niiss.org/cwis438/websites/1niiss/FieldMethods/ModWhit.php
[3] http://www.gbif.org/ipt
[4] http://rs.tdwg.org/dwc/terms/guides/text/index.htm
[5] http://rs.gbif.org/core/dwc_occurrence.xml

Occurrence[6]. 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).

In order to encode sample-based data, we here propose a third, new core, the **Event**[7] (i.e. sampling event) core, and an associated **Occurrence** extension, which is identical in structure to the Occurrence core[8] but includes two additional terms, organismQuantity and organismQuantityType (Figure 2). In the core table, each row is a sample identified by a unique eventID and other columns hold sampling protocol, sampleSize, sampleSizeUnit, date, location, parentEventID, 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 quantity measurement (organismQuantity and organismQuantityType). Following the one-to-many star schema, one Event row can link to many Occurrence rows.

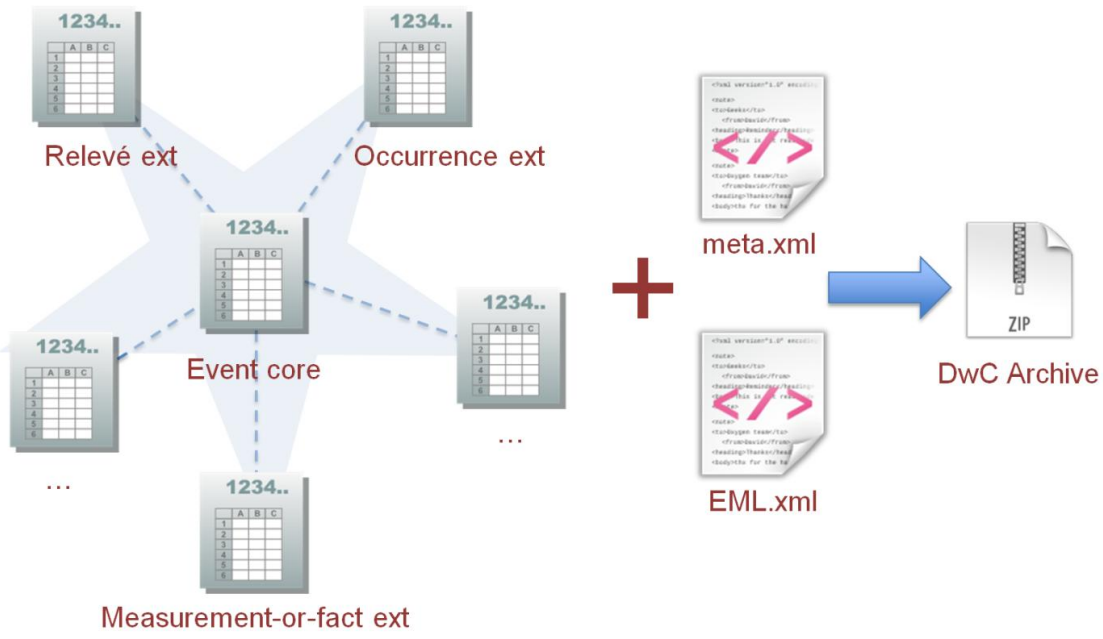

Figure 1. The components of a Darwin Core Archive.

[6] http://rs.gbif.org/core/dwc_taxon.xml

[7] 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.

[8] http://rs.gbif.org/core/dwc_occurrence.xml

The Event[9] core elements are mainly drawn from the DwC classes Event, Location and Geological Context with the addition of the three new terms sampleSize , sampleSizeUnit and parentEventID (Table 3). The Occurrence extension draws from the Occurrence, Taxon and Identification classes with the addition of the two new terms organismQuantity and organismQuantityType. 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, parentEventID*, samplingProtocol, sampleSize*, sampleSizeUnit*
Occurrence Extension	eventID, organismQuantity*, organismQuantityType*

The * symbol indicates proposed new terms.

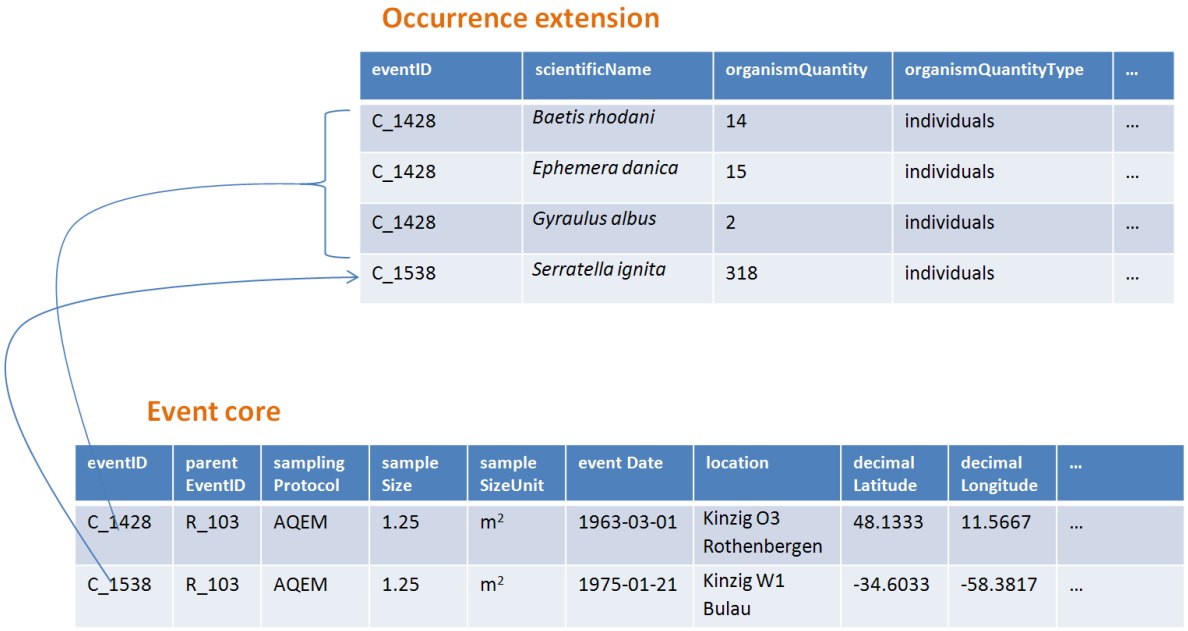

Figure 2. A sampling event uses a particular samplingProtocol , sampleSize, sampleSizeUnit etc. and can record one or more taxa, each of which has a measurement (organismQuantity and organismQuantityType) associated with it.

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.

[9] http://rs.gbif.org/sandbox/core/dwc_event.xml

Freshwater invertebrate survey

Core (Event) table

eventID	sampling Protocol	sample Size	sample SizeUnit	event Date	location	decimal Latitude	decimal Longitude	…
C_1428	AQEM	1.25	m^2	21/06/2006	Kinzig O3 Rothenbergen	50.18689	9.100369	
B_1538	AQEM	1.25	m^2	11/06/2008	Kinzig W1 Bulau	50.1316	8.9657	

Extension (Occurrence) table

eventID	scientificName	organismQuantity	organismQuantityType	…
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 square 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	sampling Protocol	sample size	sample SizeUnit	startDay OfYear	end Day OfYear	year	location	decimal Latitude	decimal Longitude	…
IA1	hand operated van Veen grab	0.04	m^2	147	154	1995	Gialova lagoon	36.9564	21.6661	
IA3	hand operated van Veen grab	0.04	m^2	147	154	1995	Gialova lagoon	36.9564	21.6661	

Extension (Occurrence) table

eventID	scientificName	organismQuantity	organismQuantityType	…
IA1	*Abra ovata*	57	individuals	
IA3	Bittium reticulatum	113	individuals	

Extension (Measurement-or-Fact) table

eventID	measurement Type	measurement Value	measurement Unit	measurement Remarks	…
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 square 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	sampling Protocol	sample Size	sample SizeUnit	event Date	location	decimal Latitude	decimal Longitude	…
1001	Braun Blanquet	100	m^2	09/08/2012	Kinzig O3 Rothenbergen	50.18689	9.100369	

Extension (Occurrence) table

eventID	scientificName	organismQuantity	organismQuantityType	…
1001	*Acer pseudoplatanus*	r	BraunBlanquetScale	

Extension (Relevé) table

eventID	syntaxon Code	inclination	cover Total	trees Cover	cover Shrubs	highTree LayerHeight	highHerb LayerMeanHeight	mosses Identified	…
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	sampling Protocol	sample size	sample SizeUnit	startDay OfYear	end Day OfYear	year	location	decimal Latitude	decimal Longitude	…
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	organismQuantity	organismQuantityType	…
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	sampling Protocol	sample Size	sample SizeUnit	eventDate	location	decimal Latitude	decimal Longitude	…
1014-tr023m	Pollard walks	250	m^2	2012-10-11	Ramat Hanadiv botanik garden	32.553191	34.947492	
1012-tr006-s5	Pollard walks	250	m^2	2012-05-02	Carmel Hurshan haarbaim	32.75789805	35.02697333	

Extension (Occurrence) table

eventID	scientificName	organismQuantity	organismQuantityType	…
1014-tr023m	*Pieris rapae*	1	individuals	
1014-tr023-s5	*Maniola telmessia*	2	individuals	

Extension (Measurement-or-Fact) table

eventID	measurement Type	measurement Value	measurement Unit	measurement Remarks	…
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	sampling Protocol	sample Size	sample SizeUnit	event Date	location	decimal Latitude	decimal Longitude	…
506003329	Reef Life Survey methods	500	m^2	2006-09-02	Cocos Islands	5.56187	-87.04693	
57003326	Reef Life Survey methods	500	m^2	2006-12-11	Panama Bight	4.008553	-81.605377	

Extension (Occurrence) table

eventID	scientificName	organismQuantity	organismQuantityType	…
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.

Nested samples

Table 4. Several sub-plots may be related to the parentEventID as in this example of a Whittaker plot consisting of 13 sub-plots (see Figure 3 for plot layout).

eventID	parent EventID	sampling Protocol	sample size	sample SizeUnit	eventDate	location	decimal Latitude	decimal Longitude	…
A1		Modified Whittaker Plot[10]	1000	m^2	18/03/84	Monarch	55.72	13.28	
A1.1	A1		100	m^2					
A1:2	A1		10	m^2					
A1:3	A1		10	m^2					
A1:4	A1		1	m^2					
A1:5	A1		1	m^2					
A1:6	A1		1	m^2					
A1:7	A1		1	m^2					
A1:8	A1		1	m^2					
A1:9	A1		1	m^2					
A1:10	A1		1	m^2					
A1:11	A1		1	m^2					
A1:12	A1		1	m^2					
A1:13	A1		1	m^2					

[10] http://www.niiss.org/cwis438/websites/1niiss/FieldMethods/ModWhit.php

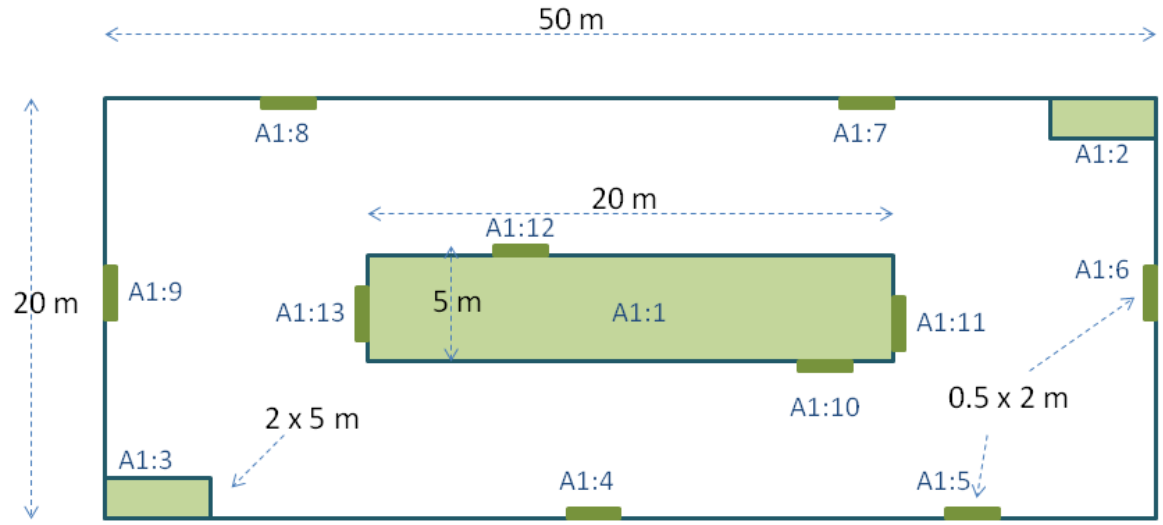

Figure 3. A schematic of a Whittaker plot consisting of 13 sub-plots of varying area.