Darwin Core for sample data

Contents

[Introduction 1](#_Toc381007943)

[Abundance 1](#_Toc381007944)

[Use cases 2](#_Toc381007945)

[What is sample-based data? 2](#_Toc381007946)

[New Darwin Core terms versus MeasurementOrFact 2](#_Toc381007947)

[Defining new terms 3](#_Toc381007948)

[Absolute vs. relative counts 4](#_Toc381007949)

[What class is abundance a property of? 4](#_Toc381007950)

[Sample data in the GBIF work programme 4](#_Toc381007951)

[Goals & Milestones in WP 2014-2016 4](#_Toc381007952)

[Sample data in EU BON 5](#_Toc381007953)

[Work plan proposal 5](#_Toc381007954)

[Tasks 5](#_Toc381007955)

[Standards and tools 5](#_Toc381007956)

[Publishing 6](#_Toc381007957)

[Discovery and access 6](#_Toc381007958)

# Introduction

This document scopes the requirements for extending Darwin Core (DwC) for publishing sample-based data to the GBIF network drawing on synergies between the GBIF work programme for 2014-2016 and the goals of the EU BON project on tools for data sharing.

# Abundance

Following the outcomes of the GBIF workshop[[1]](#footnote-1) on extending DwC for sample-based data, a new term “abundance” was proposed for discussion by the community on the TDWG mailing list. Abundance was defined as “*The number of individuals of a taxon found in a sample. This is typically expressed as number per unit of area or volume*”. The term was meant to reflect, at least in part, the definition used by GEO BON in its work on Essential Biodiversity Variables – “*quantity of individuals or biomass of a given taxon or functional group at a given location*”. Abundance is typically measured by taking a sample of a population and it is assumed that a sample has a particular context (time, place, protocol).

However, in the ensuing discussions on the TDWG list, it became obvious that the term was the cause of confusion and that it is better not to conflate the concept of abundance as normally used by ecologists to describe the absolute or relative presence of a taxon in an ecosystem with the raw numbers in the sample data that are used to derive the abundance. The aim in DwC should be to report the quantities and their types in a sample and let a measure of abundance be derived from those as a separate exercise.

# Use cases

What are our use cases?

1. To support work flows requiring quantitative data for, e.g., BioVel, that will enable automated processing of Essential Biodiversity Variables in order to generate biodiversity indicators.
2. To support data aggregators such as GBIF in developing new data visualisations, statistics.

# 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.

DwC is typically used to share specimen-occurrence data as exemplified by museum / herbarium specimens and simple observation events. What are the additional terms, if any, required in DwC in order to be able to express sample-based data? The critical features for documenting sample data are the **sampling protocol** (method) and a **quantity** measure. DwC already provides terms for Event, Location, Time, etc. In particular, the Event class provides the following terms:

**eventID** | **samplingProtocol** | **samplingEffort** | eventDate | eventTime | startDayOfYear | endDayOfYear | year | month | day | verbatimEventDate | habitat | fieldNumber | fieldNotes | eventRemarks

So, what appears to be missing in DwC is a way to express quantities[[2]](#footnote-2).

# New Darwin Core terms versus MeasurementOrFact

There are two possible approaches to expressing quantities in DwC:

1. Add new terms to DwC for capturing a quantity value and the units that value is measured in;
2. Use the terms in the already existing MeasurementOrFact DwC class to capture quantities:

measurementType | measurementValue | measurementAccuracy | measurementUnit | measurementDeterminedDate | measurementDeterminedBy | measurementMethod | measurementRemarks

There is a constraint on using MeasurementOrFact, however. For Simple Darwin Core (typically as used in the GBIF Integrated Publishing Toolkit (IPT)), only a single instance can be used[[3]](#footnote-3).

It can also be argued that quantity is such a much needed and common property that it should be made to work easily in Simple Darwin Core (i.e., by adding specific terms for it to DwC).

# Defining new terms

What is the minimal set of new terms needed in DwC to express quantity in sample data? The following table lists terms proposed on the TDWG mailing list. It should be noted that most of the terms make use of the word “abundance” as the original proposal from the GBIF workshop used that term. However, as explained in the introduction, it is probably better to drop the word “abundance” and instead use something like *quantity* or *countInSample* to refer to the number of individuals, counting units, etc., in a sample and *quantityType* or *countType* to define the unit of measurement.

Table . Proposed new terms to handle "abundance" in Darwin Core.

|  |  |  |
| --- | --- | --- |
| **Term** | **Definition** | **Comments** |
| abundanceabundanceValuequantity | A numeric value | [HS] Keep quantity and unit separate; no need for abundanceAsPercent; use quantity instead of abundance as it better describes things like “a catch of 1000kg of fish”, “a harvest of 200 cubic metres of timber”. |
| abundanceUnitabundanceType | Specifies whether the abundanceValue/quantity is in % of species, % of biovolume, % of biomass, individuals/l, individuals/m^2, individuals/m^3, individuals/sampling effort, etc. | [AdeW] Use separate terms for value and type.[EOT] Need to define a controlled list of types for abundanceValue |
| abundanceMethod |  |  |
| abundanceAsPercentabundanceAsNumber | A numeric value referring to the % count of a taxon within a sample | [DH] This is covered by using abundance with abundanceType where value is “% of species” |
| quantityRange |  | [HS]Provide a range for the quantity as uncertainty is so common in sightings |
| individualCount |  | [HS] should be deprecated  |

## Absolute vs. relative counts

How useful is it to take a simpler approach and just provide a measure of relative presence of a taxon in a sampling event, e.g., the number of individuals of a species in a Malaise trap as a % of the number of individuals of all species present? This would require three properties to be associated with an occurrence: i) *samplingEventID*, ii) *samplingProtocolIdentifier*, iii) *relativeAbundanceValue*. An alternative, more flexible solution as suggested in Table 1, is to use, in place of *relativeAbundanceValue*, *abundance* and *abundanceType* where the value of the latter would be “% of species”.

Note that % counts are already partially derived and thus not strictly primary data. “… *the statistician in me compels me to say that it’s better to report raw numbers (with metadata) than already partially computed statistics (like proportions)* [Stan Blum, TDWG mailing list]”

## What class is abundance a property of?

Which Darwin Core class should *abundance* be a property of? This is a complex area and encroaches on the (deliberate) limitations of the semantic model in Darwin Core and the need to develop a proper TDWG Ontology where classes (types) of objects and their properties are clearly defined.

Neither the Occurrence class nor the newly ratified MaterialSample[[4]](#footnote-4) class are appropriate as the former refers to individual s (observations, specimens) while the latter is “*The category of information pertaining to the physical results of a sampling (or subsampling) event. In biological collections, the material sample is typically collected, and either preserved or destructively processed*”. Thus, samples that do not remove anything from the environment cannot be placed in the MaterialSample class. The BioCollections Ontology[[5]](#footnote-5) at the BioPortal is working on some new BCO terms to describe surveys and plot sampling but these are not yet available and it remains to be worked out how they could be used with DwC.

For the purposes of publishing sample data using the GBIF IPT, it will be necessary to define the equivalent of a new “sample” core, additional to the existing occurrence and taxon cores.

# Sample data in the GBIF work programme

The GBIF work programme for 2014-2016 lays out the following milestones and delivery dates relating to sample based data.

## Goals & Milestones in WP 2014-2016

1. Enable **identification** of data records which have been collected as part of a project/campaign with **consistent sampling methodologies** and use of this criterion to **discover** and **download** data.
2. Work with **interested data publishers** which hold suitable data to trial the use of **DwC extensions** for **sample** based data
3. Demonstrate **access** to sample data elements through the **GBIF portal** and **services**

**By December 2014**

* Demonstration of publishing of sample based data through the GBIF network (mobilization and discovery of additional data elements from a sample of GBIF data publishers with suitable data)

**In 2015 – 2016**

* Expand mobilization of sample based data and explore options for improved discovery and visualization of these data.

# Sample data in EU BON

Task 2.3[[6]](#footnote-6) in EU BON deals with tools for data sharing and will scope and build new releases of data sharing tools. Requirements include:

1. Tools should implement selected interoperability mechanisms, e.g., for monitoring data;
2. Tools should implement data publishing mechanisms, e.g., for producing data papers;
3. Tools should provide registration and query functions towards the GCI;
4. Base development of tools for metadata, occurrence data and ecological data on those from GBIF and LTER;
5. Investigate new tools for sharing habitat data;
6. Adopt distributed model of development.

# Work plan proposal

As the GBIF and EU BON work programmes share some common requirements and outcomes regarding sample-based data, we propose to align their activities in a mutually beneficial way. The plan is outlined below.

## Tasks

### Standards and tools

1. Prepare section on the GBIF IPT and sample-based data for the EU BON milestone report (due April 2014 with a first draft by end of March).
2. Use the EU BON training workshop in Crete to introduce partners to DwC-A and IPT for sample data.
3. Agree on a new **“sample” core** for DwC archives. The existing two cores (occurrence and taxon) are published in <http://rs.gbif.org/core/> but, until ratified, the new sample core prototype is better placed in <http://rs.gbif.org/sandbox/> on a temporary URL where it can be referenced from a DwC-A.
4. Agree on the essential terms to be associated with a taxon occurrence in order to be able to express sample-based data in DwC. The minimum appears to be:
	1. **eventID** (“An identifier for the set of information associated with an Event (something that occurs at a place and time). May be a global unique identifier or an identifier specific to the data set.”)
	2. **samplingProtocol** (“The name of, reference to, or description of the method or protocol used during an Event.”)
	3. **quantity[[7]](#footnote-7)** (new; to be defined; “the number of units found in the sample”)
	4. **quantityType** (new; to be defined; “the type of unit used to quantify the sample”)
5. Define the permitted list of terms in the new **sample** core: these are from the Event, Location and GeologicalContext DwC classes.
6. Define the permitted list of terms in the new **SampleTaxonOccurrence** extension”: these are from the Occurrence, Identification and Taxon classes and include a new “Quantity” grouping/class with two new terms *quantity* and *quantityType*.
7. Decide on the most machine efficient way to refer to a sampling protocol (controlled vocabulary; publication reference; URL?).
8. Decide on the most machine efficient way to refer to the types of quantity (controlled vocabulary?). Examples: % of species, % of biovolume, % of biomass, individuals/l, individuals/m^2, individuals/m^3, individuals/sampling effort, etc.
9. If it is decided to use controlled vocabularies for samplingProtocol and quantityType, then choose a platform to develop these (e.g. <http://terms.tdwg.org>) and once ready, prepare them in the GBIF format for publishing on <http://rs.gbif.org/vocabulary/>

### Publishing

1. Set up a test instance of the IPT so that it recognises the new sample core and use it to prepare some data sets for publication (this will require some re-coding of the IPT).
2. Identify at least two uses cases where data publishers want to use/extend DwC for publishing sample based data using the above model. Try to get some EU BON partners to act as publishers. The five LTER sites would be good candidates for publishing sample data as they are likely to share a common sampling protocol.
3. Ensure that each sample data set has complete metadata in EML including full description of project, sampling protocols and methods.
4. Invite the TDWG list to test the IPT with example data sets.

### Discovery and access

It is necessary to demonstrate access to sample data elements through the GBIF portal and services. This requires indexing of the new properties associated with sample based data sets (viz *eventID*, *samplingProtocol*, *quantity*, *quantityType*) so information about them can be retrieved through a search filter (in combination with other search parameters already available in the portal).

1. Portal indexes new DwC elements
2. Ability to filter on a core (sample vs. occurrence or taxon)
3. Ability to select records associated with a particular sampling protocol (use drop down list?)
4. Ability to select records associated with a particular quantityType (use drop down list?)
5. Automate workflow for BioVel
	1. Ability to select sampling protocol via web service (e.g. an LTER sites protocol)
	2. Ability to return records associated with the selected sampling protocol (with other arbitrary filters
	3. BioVel consumes sample data and does something interesting (e.g., calculates a Biodiversity Indicator).
1. <http://www.gbif.org/resources/2245> [↑](#footnote-ref-1)
2. The Darwin Core term “individualCount” defined as “the number of individuals represented present at the time of the Occurrence” is not suitable as a quantity measure for sample data. [↑](#footnote-ref-2)
3. See <http://rs.tdwg.org/dwc/terms/simple/index.html#rules> for explanation. [↑](#footnote-ref-3)
4. <http://rs.tdwg.org/dwc/terms/MaterialSample> [↑](#footnote-ref-4)
5. <http://bioportal.bioontology.org/ontologies/BCO> [↑](#footnote-ref-5)
6. An EU BON milestone report “Specifications for data sharing tools” is due in May 2014 but the task runs on until Feb 2017 by which time actual software tools must be released. [↑](#footnote-ref-6)
7. The names “quantity” and “quantityType” are not necessarily those that will be used. Alternatives will also be considered before a final decision is made. [↑](#footnote-ref-7)