Extending the GBIF Integrated Publishing Toolkit for sample data

# Éamonn Ó Tuama, GBIF Secretariat

***Please note that the modifications proposed here to Darwin Core and the IPT are experimental: all changes will need to undergo acceptance testing and ratification.***

# Introduction

This document scopes the requirements for extending Darwin Core (DwC) for publishing sample-based data using the GBIF Integrated Publishing Toolkit (IPT). The driving context are the tasks within the GBIF 2014 work programme “demonstration of mobilization of sample-based data through the GBIF network” and the EU BON project “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.

# 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 and also, possibly, a term for linking associated events: “eventSeries”.

# 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 type of measurement.

Table . Some proposals for terms to handle "abundance" in Darwin Core as discussed on the TDWK mailing list.

|  |  |  |
| --- | --- | --- |
| **Term** | **Definition** | **Comments** |
| abundance  abundanceValue  quantity | 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”. |
| abundanceUnit  abundanceType | 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 |  |  |
| abundanceAsPercent  abundanceAsNumber | 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”.

# Darwin Core Archive data model

The GBIF IPT uses a particular text format called Darwin Core Archive (DwC-A)[[2]](#footnote-2). 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. Currently, GBIF uses two cores: Taxon or Occurrence. 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.

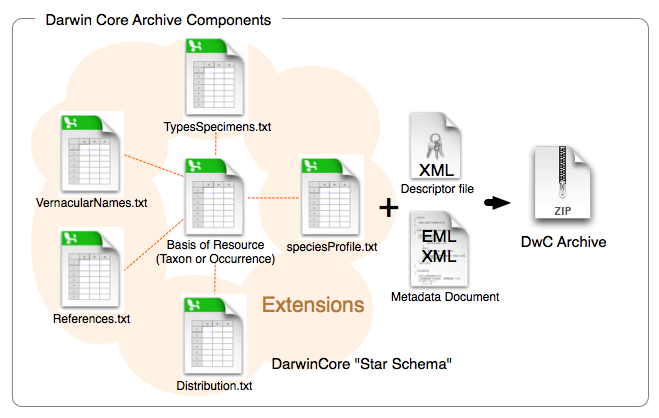


Figure . The components of a Darwin Core Archive.

Here, we are proposing a third, new core for sample data (Figure 2) consisting of a **core** Event[[3]](#footnote-3) (i.e. sampling event) table and an event-occurrence **extension** table. In the core table, each row is a sample identified by a unique eventID and other columns describe sampling protocol, date, location, etc. The rows in the extension table reference a sample 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 sample row can link to many taxon rows.

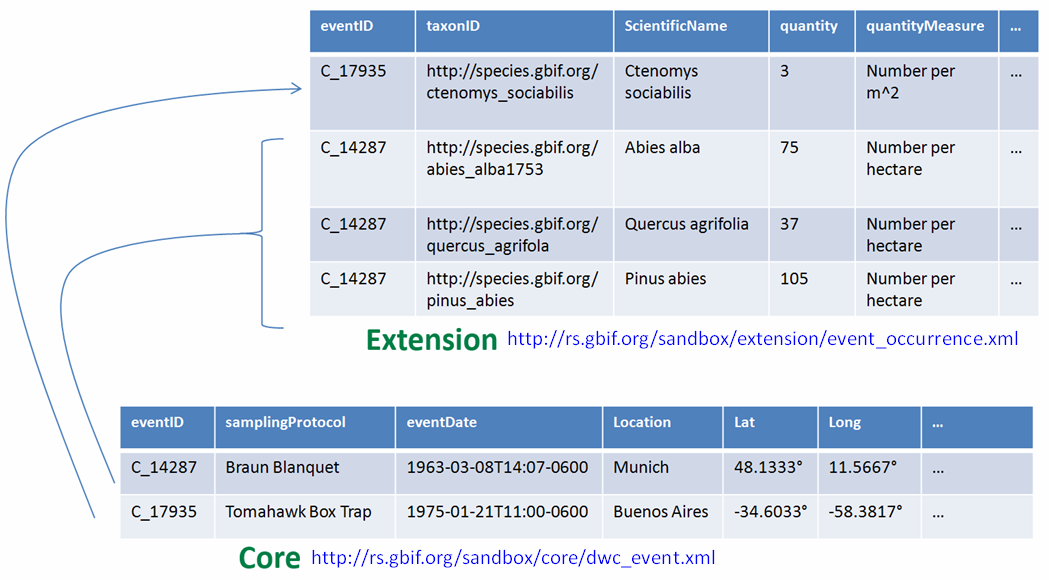


Figure 2. The data model proposed for the Darwin Core Archive for sample data.

# Requirements

## Standards

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[[4]](#footnote-4)** (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. **eventSeries** (new; to be defined; “a set of events that are associated in some way”), e.g. as part of a monitoring series; may be a global unique identifier or an identifier specific to the series
6. Decide on the most machine efficient way to refer to a sampling protocol (controlled vocabulary; publication reference; URL?).
7. Decide on the terms in a controlled vocabulary for quantityType. Examples: % of species, % of biovolume, % of biomass, individuals/l, individuals/m^2, individuals/m^3, individuals per day, etc.

## 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 by GBIF.
2. Identify at least 5 uses cases where data publishers want to use/extend DwC for publishing sample based data using the above model. Invite 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.

## Discovery and access

It will be necessary to demonstrate access to sample data elements through the EU BON portal and services. This requires indexing of the new properties associated with sample based data sets (viz *eventID*, *eventSeries*, *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).

### Desirable functionality

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.standardsingenomics.org/index.php/sigen/article/view/sigs.4898640/1067> [↑](#footnote-ref-1)
2. <http://rs.tdwg.org/dwc/terms/guides/text/index.htm> [↑](#footnote-ref-2)
3. See the GBIF workshop report referenced on page 1 for discussion of an alternative sample core based on the Occurrence core. [↑](#footnote-ref-3)
4. 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-4)