EXTERNAL SCIENTIFIC REPORT

APPROVED: 01 April 2020

doi:10.2903/sp.efsa.2020.EN-1841

Applying the Darwin core standard to the monitoring of wildlife species, their management and estimated records

ENETWILD consortium[1], Guillaume Body[a,b], Mathilde Mousset[b], Emmanuelle Chevallier[b], Massimo Scandura[c], Sophie Pamerlon[ade], Jose Antonio Blanco-Aguiar[f], Joaquín Vicente[f]

[a] Direction de la surveillance des évaluations et des données, Office Français de la Biodiversité, France, [b] Direction de la Recherche et de l'Expertise, Office National de la Faune Sauvage, France, [c] Department of Veterinary Medicine, University of Sassari, Sassari, Italy., [d] Unité Mixte de Service Patrimoine Naturel OFB-MNHN-CNRS, France, [e] GBIF France, France, [f] National Institute on Wildlife Research (IREC), University of Castilla-La Mancha and Consejo Superior de Investigaciones Científicas, Ciudad Real, Spain.

Abstract

Enetwild consortium aims at aggregating data on occurrence, abundance and hunting bag of wildlife in Europe, either as raw data or as results of statistical estimation. These data come from a large community of researchers, hunters and wildlife managers. A flexible and robust data standard is therefore necessary to present the large diversity of data and collection method. We evaluated the possibilities offered by the Darwin Core Standard. The Event core, the occurrence extension and the extended measurement or fact extension proved their utility for our purpose. However, these were not able to record statistical estimation values. We proposed to extend the measurement or fact extension to allow them to be nested among themselves. Any confidence interval or precision measure is indeed a measurement about the punctual estimate, another measurement. We proposed controlled vocabularies adapted to wildlife survey in data and metadata. This will be aligned with the EFSA data model harmonisation under the SIGMA project.

Key words: Data standard, Darwin core, population monitoring, statistical estimation

Question number: EFSA-Q-2020-00220

Correspondence: alpha@efsa.europa.eu

[1] ENETWILD Consortium: www.enetwild.com

Acknowledgements: Thanks to Philippe Aubry and Clément Calenge for their help on statistical concepts and Solène Robert, Rémy Jomier, Silvère Camponovo, Yvan Le Bras, Frédéric Vest for their helpful discussion about standards, and Anaïs Just for her help on the way to record units.

Suggested citation: ENETWILD consortium, Guillaume Body, Mathilde Mousset, Emmanuelle Chevallier, Massimo Scandura, Sophie Pamerlon, Jose Antonio Blanco-Aguiar, Joaquin Vicente, 2020. Applying the Darwin core standard to the monitoring of wildlife species, their management and estimated records. EFSA supporting publication 2020:EN-1841. 81 pp. doi:10.2903/sp.efsa.2020.EN-1841.

ISSN: 2397-8325

The present document has been produced and adopted by the bodies identified above as authors. This task has been carried out exclusively by the authors in the context of a contract between the European Food Safety Authority and the authors, awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

Summary

Our first standard, adapted from the Darwin Core archive to store data about wild boars, has showed that this format has a good enough flexibility to be adapted to the multiple data types that the Enetwild project aims at storing at the same place. However, the consortium now extends the research for data to a large set of new species, which are studied in various way and on various scales, and to more protocoled and estimated data. The introduction of data coming from statistical procedure interrogates the suitability of the Darwin Core standard to record such values and their methodological details. We came back to the original Darwin Core structure to think of new improvements and adaptations and make it useable to face this new challenge.

The Darwin core standard and its star structure successfully proved its ability to record complex and very structured raw and summarized data using the Event core and two extensions, the *occurrence* and the *extended measurement or fact* extensions. Using the current state of the Darwin core standard, both biotic and abiotic information can be recorded, allowing the description of the protocol implementation.

We proposed a light improvement of the *occurrence* extension to be able to record partial-data, i.e. different views of the same occurrence when it corresponds to a group of individuals. The inclusion of estimated values required a fundamental, but light, improvement of the *extended* measurement, allowing to nest record in each other in this extension. This allows to record statistical information about another statistical values, such as the confidence interval of an estimation. We named it the *nested extended measurement or fact* extension. We also completed the metadata and the distinction of the information that have to be find in data vs metadata.

To be applicable to Enetwild purpose, we complete this report with three annexes:

- the **list of variables** that are of particular interest for our objectives to include in the *Event Core*, the *occurrence* and the *nested extended measurement or fact* extensions (Appendix A). While it is still allowed to include other variables from the Darwin core according to its rules, this selection should be sufficient for Enetwild goal.
- the list of **controlled vocabulary**, which particularly focus on terms adapted for wildlife surveys (Appendix B). These vocabularies where as much as possible based on international references such as ISO, and linked to the International Statistical Institute (ISI), or discussed with expert biometrician for statistical notions.
- an excel file presenting the **proposed standard** corresponding to this report (Appendix C).

In this report, we focus on data standard, and we included some considerations on metadata fields. Further discussion is needed before including the proposed standards into the current main international metadata in ecology: the Ecological Metadata Language (EML)[2].

We proposed to name this new standard the "Wildlife monitoring data standard", a version of the Darwin Core standard.

[2] https://eml.ecoinformatics.org/eml-schema.html

Table of contents

1 Introduction

1.1 Background and Terms of Reference as provided by the requestor

This contract was awarded by EFSA to Universidad de Castilla-La Mancha, contract title: Wildlife: collecting and sharing data on wildlife populations, transmitting animal disease agents, contract number : OC/EFSA/ALPHA/2016/01 – 01

The terms of reference of the present report were to develop standards for data collection on presence, abundance, density of wild ruminants, wild carnivores in Europe.

1.2 Scope of the report

The general goal of Enetwild (EFSA/ALPHA/2016/01 Contract, Wildlife: collecting and sharing data on wildlife populations, transmitting animal disease agents) is to take a harmonised approach to data collection activities regarding population data (distribution and abundance) of selected species of wildlife that are relevant because of the pathogens they may transmit to domestic animals and humans (Network Strategic Plan, ENETWILD 2017).

The first objective of the project consists in collecting existing published or unpublished data on the geographical distribution, abundance and structure of selected wildlife hosts, to validate and to aggregate them in a harmonized way in a common database.

An important step to "*aggregate [data] in a harmonised way in a common database*" is to define the requested data, their structure, their acquisition method and to offer a common framework to insert them, i.e. a standard. This data report describes how to enlarge the current Wild Boar Data Model of Enetwild Project toward all species, and the record of presence, abundance, hunting bag and sampling effort based on the former data standard analysis[3].

1.3 Challenges & problem framing

The current standard were primarily developed by the consortium and derive from the international Darwin Core (DwC) and Ecological Metadata Language (EML) biodiversity standards for sharing data and metadata. They consist mainly in DwC and EML fields with some additions, dispatched in several files depending on the data type.

The DwC standard, primarily created by the Biodiversity Information Standards (TDWG) community to share occurrence and taxonomic records, underwent important evolutions in the last years. Now, it allows the inclusion of a wider range of data, especially in regards to measurements and sampling events. In the case of Enetwild, an adaptation of those developed standards has allowed the easy storage of data specific to wild boars, through the Wild Boar Data Model (WBDM).

However, this model is today specific to one species and focuses on a restricted number of study types and contextual questions (ex: presence of pig husbandry). EFSA, and therefore Enetwild aims to broaden the current model to include numerous species (i.e. European ungulates and carnivores) and different data type (e.g. raw data about presence or calculated densities). This brings forward several challenges, and forces our reflection towards the general concepts underlying data structure, if we aim to store them in a unique place.

[3] Body G., Cohen Nabeiro A. (2018) Proposal of presence data model for Enetwild, based on international data standards. Report Enetwild 2.2 for EFSA

The new standards should allow the inclusion of a wider range of data, which we cannot fully know in advance; in a structure common for all types of records, explicit and understandable, and usable in different contexts. For instance, the marine OBIS (Ocean Biogeographic Information System[4]) and the terrestrial GBIF (Global Biodiversity Information Facility[5]) initiatives are based on the Darwin Core standard.

To rethink the organisation of our standards, we have to consider the design of each study. More specifically, we need to be able to describe the hierarchy of events leading to this design and to store facts and measures relative to each of them, as well as showing the links between every element of this hierarchy and the intensity of the data collection. This introduced hierarchy also needs to be simple to navigate through.

For instance, we must be able to describe that a particular observation was collected in a particular point within a particular transect which itself results from a random sampling protocol in a larger administrative area, which itself is one among others. The campaign of data collection can also be repeated. An efficient standard will allow the user to describe precisely all of these study steps and will allow the re-user to understand it.

We need to be able to store various type of data: occurrence data (a presence of a species at a particular place and time), technical data (e.g. angle and distance toward an observation while performing a distance sampling protocol), and calculated data (e.g. density estimation including their precision); each of them presenting their own difficulties. Although we have tested the storage of occurrence and hunting bag records for wild boar, it requires generalisation for new species. Including technical raw data means to include many different variables that we will not be able to explicitly included as fields, while calculated data comes with a wide range of modalities that also need to be stored.

Differentiate data and metadata is another challenge, as metadata definition is only "data about data". One good approach to make such a distinction is to determine what is essential to find the dataset and understand it (i.e. metadata), and what is technically needed to use the data once understood (i.e. data).

In the former report[6], we explained how the Darwin Core Archive presented good characteristics for answering such a challenge, as well as the EML for metadata.

We will here in a first part describe how to use the DwC-A for our objectives, issues we met and proposed solutions. We start with a reminder of what is the DwC-A and its recent evolution, and we then go on how to use it to describe a study design, to report a biological, then a technical observation. We finally discuss ways on how to report calculated results, to store methods and sampling protocols.

In a second part, we will expose the list of variables we selected from the DwC-A and its proposed evolution. The third part proposes lists of controlled vocabularies to be able to fit all the possible data in a common standard: **the wildlife monitoring standard**.

[4] https://obis.org/

[5] https://www.gbif.org/

[6] Body G., Cohen Nabeiro A. (2018) Proposal of presence data model for Enetwild, based on international data standards. Report Enetwild 2.2 for EFSA

1.4 Description and recent developments

The Darwin Core is an international standard to share data about biodiversity, occurrence of organisms and links to their environments. It appeared around 1999 as a set of loosely defined terms, and progressed with the help of many groups, until the Darwin Core Task Group of the community of Biodiversity Information Standards (TDWG) took in charge to decide on a formal set of terms and processes that was ratified as a standard in October 2009[7]. The community initially developed this standard to store occurrence records (i.e. one species at precise place and time, the "Occurrence Core"), as simple and open as possible, only developing terms in the event of a shared demand. Following the initiative of EUBON in 2014, the community thus agreed on important developments to extend the standard to sampled-based protocol ("Event Core") and then with OBIS in 2017 to the abiotic context of the observations ("Extended Measurements of Facts Extension").

The Darwin core now plays a fundamental role in the sharing of open access biodiversity data and represents for instance a large majority of the 1.4 billion of species occurrence records shared by the Global Biodiversity Information Facility (GBIF), published by more than 1561 organizations in 59 countries in January 2020. It is also the base of the 3,000 datasets shared in the Ocean Biogeographic Information System (OBIS), another main international and collaborative database about biodiversity.

In practice, using the Darwin Core comes **down to using a standard file format, the Darwin Core Archive (DwC-A).** The Darwin Core corresponds to one or several flat tables (typically csv documents). In the case of several tables, as we will describe below, they are organized around a "star" schema with a central table named "Core" linked to all the others named "Extensions" (Figure 1). Together with a metadata file (based on the Ecological Metadata Language), they form an Archive that contains a group of coherent data (a dataset) and the necessary information to discover and understand the dataset (the metadata).

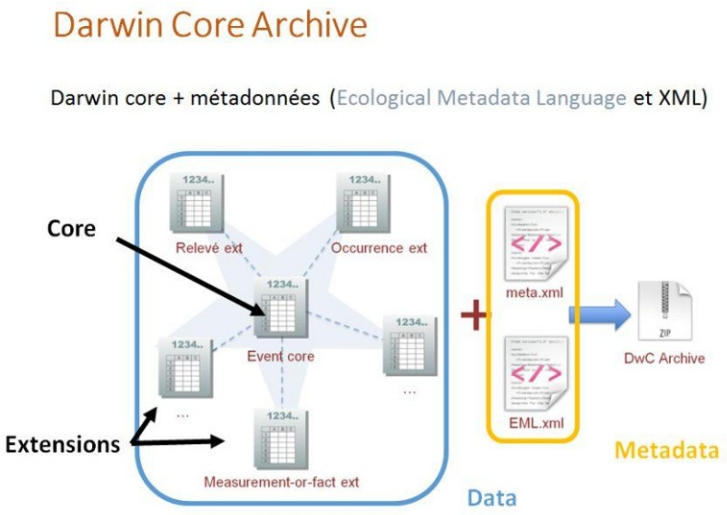


Figure 1: The star organisation of a Darwin Core Archive

[7] Wieczorek J. et al. (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE 7(1): e29715. doi:10.1371/journal.pone.0029715
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029715

The Darwin core uses a set of clearly defined classes and terms that are either specific to an extension or generic to any "record", building a common terminology which guarantees that data will keep their meaning while used by different people or machines. Most of the terms can be found on the Darwin core website (https://dwc.tdwg.org/terms/), and are organized into nine categories (often referred to as "classes"):

- Record-level Terms, which concern Dublin Core terms, institutions, collections and nature of data records, they can be used in any core or extension;
- Occurrence, about the evidences of species in nature, observers, behaviour, associated media, and references;
- Event, on the sampling protocols and methods, data, time and field notes;
- Location, about the geography, locality descriptions and spatial data, they can be used in any core or extension;
- Taxon, with terms describing the scientific names, vernacular names, names usage, taxon concepts, and the relationship between them;
- GeologicalContext, about geologic time, chrono-stratigraphy, biostratigraphy and lithostratigraphy;
- ResourceRelationship, which explicit establishes relationships between identified resources (e.g. taxon to location)
- MeasurementOrFact to store measurements, facts, characteristics, assertions and references.

The two last categories are part of the recent developments we evoked, and require a more complex data structure than the original flat structure. We will here propose additional ones, filling up gaps where needed by the Enetwild project.

If we only use the original and most simple version of the Darwin core, it is composed of a single table, with a list of standardized column names and definitions. The user then picks variables according to his needs. For instance, the observation of a wild boar at a precise location that is stored in our own original presentation in French (Table 1a) would be presented in a standardized and shareable way using the Darwin core (Table 1b).

Table 1a: Simple occurrence data as recorded in the producer format (original variable name, language and date format)

Date	***Localité***	***Espèce***	***Nombre d'individus***
06/05/2019	Rambouillet	Sanglier	2
06/07/2018	Auffargis	Sanglier	1

Table 1b: Simple occurrence data as translated into the Darwin core standard

occurrenceID	**basisOfRecord**	**eventDate**	**scientificName**	**locality**	**individualCount**

00001	Human Observation	2019-05-06	*Sus scrofa*	Rambouillet	2
00002	Human Observation	2018-07-06	*Sus scrofa*	Auffargis	1

Note that here, the **basisOfRecord** column can only be filled in with values from the following list, as it is a variable associated with a controlled vocabulary: *PreservedSpecimen, FossilSpecimen, LivingSpecimen, MaterialSample, Event, HumanObservation, MachineObservation, Taxon, Occurrence* (https://dwc.tdwg.org/terms/#dwc:basisOfRecord)

However, and especially due to the need of standardizing more than simple occurrence facts, datasets can be restructured into groups of records: the Core and its Extensions. Identification of records will therefore play a key role to link them into Parent-Child relationships.

We identified that three groups are necessary for Enetwild purposes:

- The "Event core" can be used **to describe the structuration of the sampling process**, the "events": campaign, transect, point count, study area, period of study etc.; The purpose of the sampling and the design method are recorded in metadata
- The "Occurrence extension" can be used **to describe the biological record**: species, location, time, observer, identifier etc. These records are therefore observed during an event;
- The "Extended measurement or fact extension" (eMoF) can be used **to add quantitative or qualitative information either about an event or about an occurrence**. This information therefore describes abiotic conditions or the material used while applied to an event whereas it describes abundance, sex, life stage, morphology, body weight while applied to an occurrence.

In the Darwin Core[8], every extension has to be linked to the Core directly, through an "eventID" while using the Event core. These links allow the archive to model 1-to-n relationships (for example, one transect and n observations in this transect). It is also possible to link one Core event to another (for example in the case of subevents, such as points inside a study field), with the notion of "parentEventID" coupling a child event to its parent event. The eMoF[9] is linked to both the eventID and to the occurenceID while applied to an occurrence, but only to an eventID while applied to the event (Figure 2). The original Darwin Core presentation uses the term "eventID" in Occurrence and in eMoF extensions to refer to the parent Event record, the term "occurenceID" to link an eMoF record to an Occurrence record, but it uses the term "parentEventID" to link an Event record to its parent Event record. We propose here to be more intuitive in the Parent-Child relationship of records. Therefore we choose to replace the term "eventID" by "parentEventID" in the Occurrence and eMoF extensions, and the term "occurrenceID" by "parentOccurrenceID" in the eMoF extension.

[8] for properties relative to identification (http://rs.gbif.org/)

[9] De Pooler D. et al. (2017) Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. Biodiversity Data Journal 5: e10989 https://bdj.pensoft.net/articles.php?id=10989

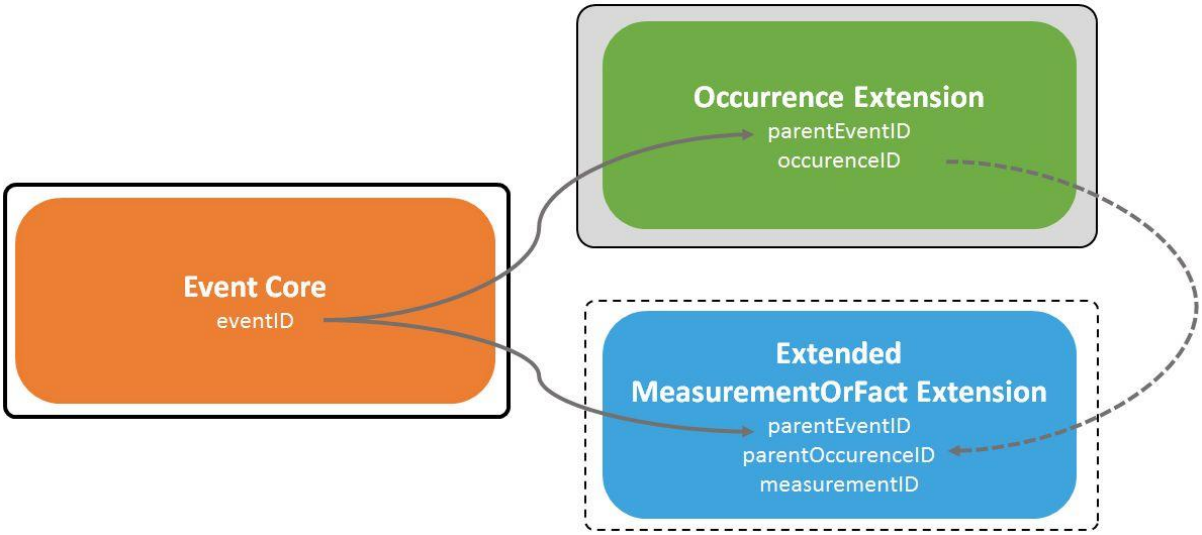


Figure 2: Schema of the notions used by the Enetwild standard (De Pooler et al 2017, adapted)

These identifications can be either explicit and unique within the dataset, or random using a universally unique identifier (UUID). Current best recommendations suggest using UUID for all identifier, but using an internal dataset unique identifier remains a common practice. We suggest that at least the dataset identifier to be an UUID. Online generator are available[10].

1.5 Case of the Wild boar data model

In our first standard, the Wild Boar Data Model, we have already used notions from the Event Core, the Occurrence extension and the Extended Measurement or Fact extension. For instance, a drive hunt was stored using notions belonging to an Event Core, while the position of each killed animal and other information such as the sex, life stage or hunting conditions were stored using notion from the Occurrence and Measurement or Fact extensions. For easier application, we have flattened this DwC structure into a simple table containing all the information that was required to store simple data for wild boars.

This adaptation has made the Enetwild standard very useful to store data about wild boars, with specific and straightforward notions such as "pig Husbandry" or the "number of dogs" and "beaters" during the hunt. The Darwin Core indeed showed robust and flexible enough to adapt to our data, which was mainly occurrence and hunting bag records.

As new types of data will now have to be included, such as estimated densities, relative abundance values, or survey specificities for other species, we come back in this report to the original and non-specific organization of the DwC. We will discuss notions to keep and/or to add according to the new challenges we cited above.

[10] https://www.uuidgenerator.net/

2 Study design or the Event Core

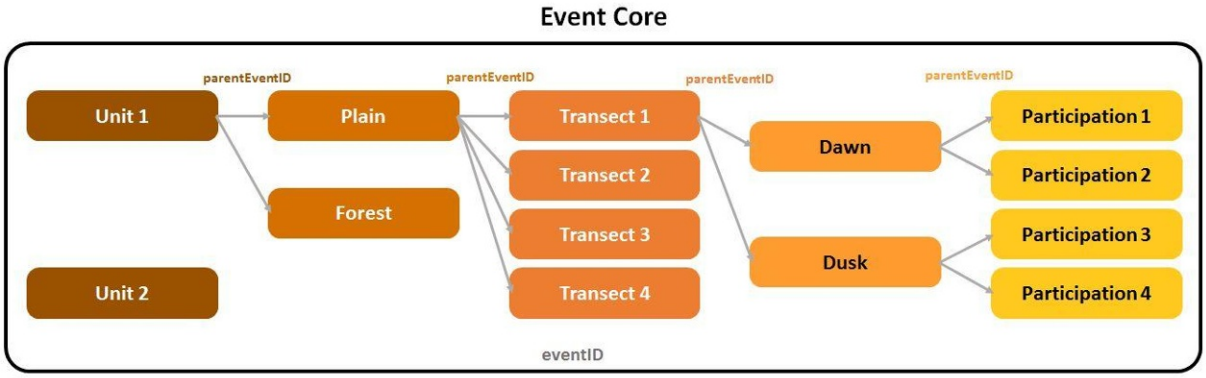
2.1 Description

The recent evolution of the Darwin Core standard induced by the EU BON project allowed it to take into account robust sample-based studies, rather than simple opportunistic species occurrence records. The Event Core allows the description of the structure of the reported study or survey. Using nested event, one can record information about high level "events" such as the global study area, intermediate level "event" such as habitat type that correspond to strata of a subsample, or such as transects, and lower level "events" such as performed protocol on a transect when observations are made, and that we refer to as "participation" hereafter.

Example: one wildlife manager team is leading a survey of a roe deer population using a Pedestrial Kilometric Index, an index of abundance from the Indicator of Ecological Change methods, which has been validated for this species in plains. It allows surveying the change in relative abundance of the population[11].

The team applies these methods to two different management unit (Unit 1 and Unit 2), within which two habitats are present (forest and plain). Between 4 and 6 transects are defined in each habitat of each unit, and they are repeated 4 times each, two at dawn and two at dusk. The method is then applied for three years before interpreting the observed trend.

We can draw this setup using events in different ways, an exhaustive one being (Figure 3):



[11] http://www.oncfs.gouv.fr/Ongules-ru220/Colloque-ICE-2015-ar1806

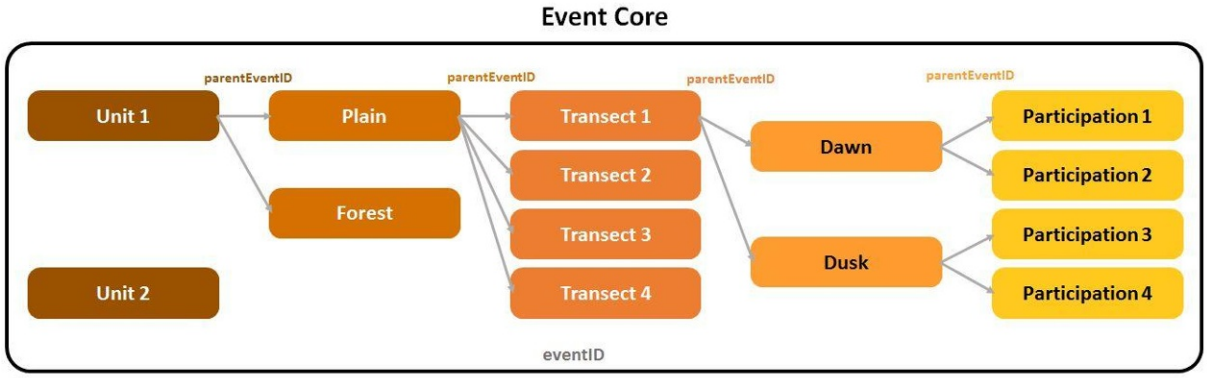


Figure 3: Modelling of a Pedestrial Kilometric Index protocol: Four repetitions at dawn and dusk of a transect are performed, each transect belonging to a habitat of a management unit.

The resulting table looks therefore like Table 2:

Table 2: Structuration of the event data records of a Pedestrial Kilometric Index protocol.

eventID	parentEventID
unit1	
unit2	
unit1:plain	unit1
unit1:forest	unit1
unit1:plain:transect1	unit1:plain
unit1:plain:transect2	unit1:plain
unit1:plain:transect3	unit1:plain
unit1:plain:transect4	unit1:plain
unit1:plain:transect1:dawn	unit1:plain:transect1
unit1:plain:transect1:dusk	unit1:plain:transect1
unit1:plain:transect1:dawn:participation1	unit1:plain:transect1:dawn
unit1:plain:transect1:dawn:participation2	unit1:plain:transect1:dawn
unit1:plain:transect1:dusk:participation3	unit1:plain:transect1:dusk
unit1:plain:transect1:dusk:participation4	unit1:plain:transect1:dausk
…	…

Carefully designing this scheme is necessary to be sure in particular to differentiate performed transects with no observations made (i.e; a true 0) from the fact that the transect was not performed. An issue which has emerged with the first models created by the consortium has indeed been the lack of data about "absence" or "non-observation" of particular species in a studied area.

Different schemes arise according to the different information we want to associate to each event, and to the degree of precision we want. One existing variable that we can use in the event core corresponds to the "*habitat*". Therefore, one could consider removing the event level "Plain vs Forest" and use the *habitat* variable to describe each transect. However, if a particular information has to be linked to the strata corresponding to the habitat, such as a different sampling weight, it would be better to keep this event level. We recommend in a general way to keep as much as possible the steps of the sampling design. The structure of the Event Core facilitates such a design.

Occurrences are then linked to the most precise event (here, the transect on which observers perform the protocol). Using an event corresponding to the "participation", i.e. when an observer performs the protocol, allows to fix all of the upper level event, while only writing once information about the participation, such as the date and time period, the number of observers, weather condition, etc…

2.2 Variables describing events

The Event core allows a number of variables to be included directly at the event level to describe it. We will further see that we can record more information using the eMoF extension.

Description of an event includes its identity, including the identity from the original database, temporal description (either a precise date or time, or a range of date or time) and geographical description (decimal coordinates, linear or polygonal shape, reference to a geographical external referential such as NUTS), but also more precise information such as habitat.

Example: The "Réseau Oiseaux de passage ONCFS-FNC-FDC" in France surveys 19 hunted bird species and one protected. Since 1996, a systematic sampling is performed. France is divided into a 1,067 cells grid, and a road was initially drawn in the centre of each cell. In each road, five counting points were identified at every kilometre (Figure 3). The program "ACT" applies this method during spring and birds are counted twice per year.

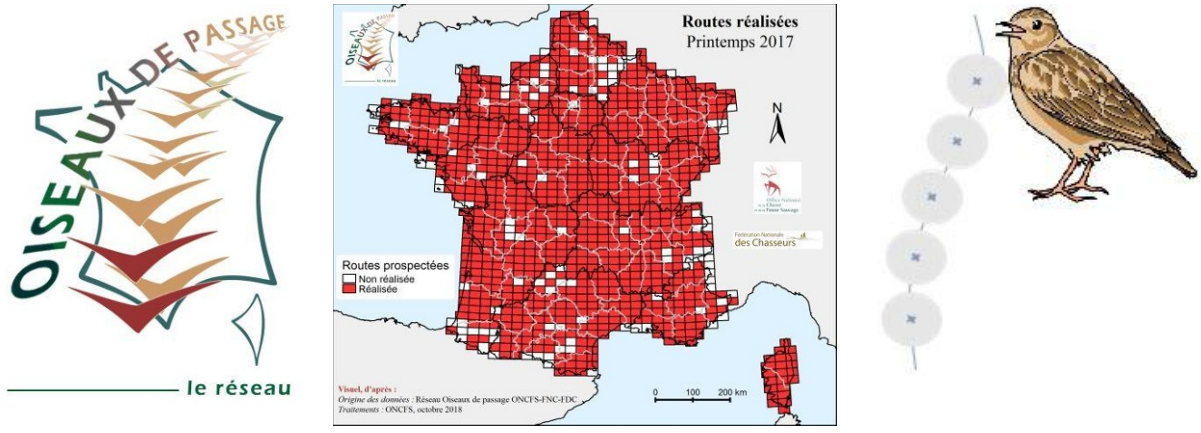


Figure 3: The ACT protocol of the Réseau Oiseaux de passage ONCFS-FNC-FDC: logo, performed roads in 2017, and representation of a road with 5 counting points.

The survey design can easily be drawn with three events levels: roads, counting points and participation (Figure 4, Table 3).

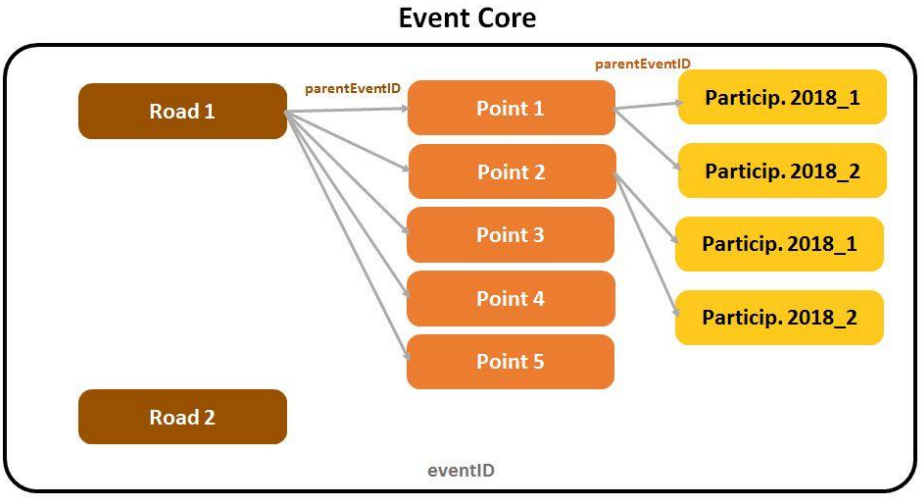


Figure 4: Schematic representation of the structure of the ACT protocol using the Event Core

Table 3: Sample of the raw data of the ACT protocol according to the producer format

CD_SIG	Année	Espèce	CD_SPC	EFFECTIF	X	Y
0316	2008	Alouette des champs	AC	1	132019.813666889	6852878.40406315
0317	2009	Alouette des champs	AC	4	132322.927958752	6838906.09544092

Roads have a spatial information: a linear that can be expressed using Well Known Text (WKT), while points have their own spatial information, different from the road: it consists in X-Y coordinates. The data, once integrated into the Event core looks like Table 4, other information will also be included into the occurrence extension:

Table 4: Structure of the events of the ACT data according to the Event core.

eventID	parentEventID	footprint WKT	decimalLatitude	decimalLongitude	Date
road1		LINE(…)			
road1:point1	road1		6852878.40406315	132019.813666889	
road1:point2	roade1		6838906.09544092	132322.927958752	
road1:point1:part1	road1:point1				2008-05-15
road1:point1:part2	road1:point1				2009-05-16

Similar models can be made for a large variety of protocol. For instance, OBIS use the event core to model data from cruises[12], automatic sensor[13], telemetry data[14] or video plankton recorder[15].

3 Record of a biological organism: the Occurrence extension

3.1 Description

The Occurrence extension is defined as "The category of information pertaining to evidence of an occurrence in nature, in a collection, or in a dataset (specimen, observation, etc.)".

We can summarize this idea as the presence of a species at a given place and date, which means that the extension includes general concepts to store information on species, location and temporal coverage as well as the number of individuals. The occurrence is linked to the event using a Parent-Child relationship: eventID-occurenceID, very similar to the parentEventID-eventID relationship.

For instance, the biological records of the two former examples, i.e. the total number of roe deer seen on a transect, and the number of individuals detected for each bird species on a count point are recorded into the occurrence table (Figure 5, Table 5).

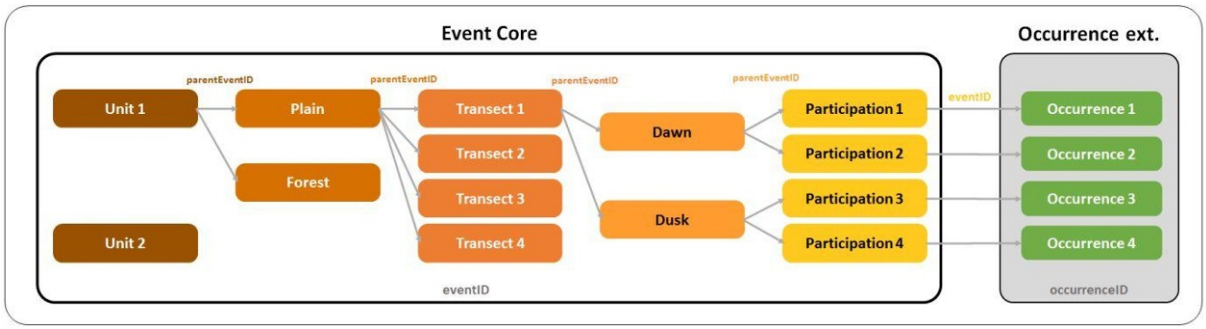


Figure 5: Schematic representation of occurrence records (number of individuals seen) in the Pedestrial Kilometric Index protocol.

Table 5: Organisation of occurrence records of the Pedestrial Kilometric Index protocol within the Occurrence extension (the … of ids correspond to "unit1:plain:transect1:dawn:")

[12] https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5345125_bdj-05-e10989-g001.jpg

[13] https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5345125_bdj-05-e10989-g015.jpg

[14] https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5345125_bdj-05-e10989-g016.jpg

[15] https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5345125_bdj-05-e10989-g017.jpg

occurenceID	parentEventID	basisOfRecord	scientificName	individualCount
…:participation1:occ1	…:participation1	human observation	*Capreolus capreolus*	10
…:participation2:occ1	…:participation2	human observation	*Capreolus capreolus*	11

Similar representations can be established for the ACT protocol (Figure 6, table 6).

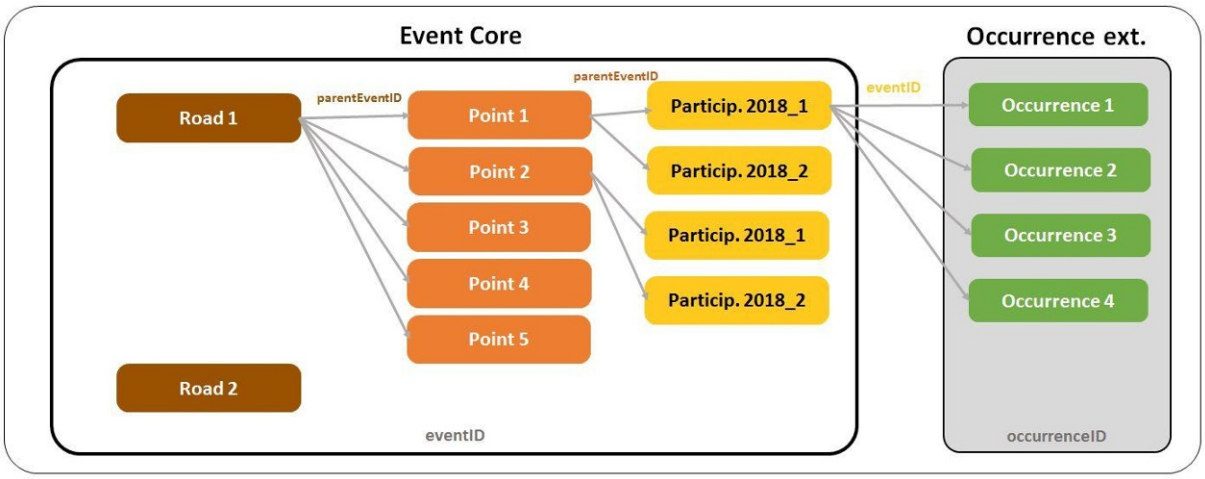


Figure 6: Schematic representation of occurrence records (number of individuals detected per species) according to the ACT protocol of the Réseau Oiseaux de Passage ONCFS-FNC-FDC. There are multiple occurrences for a single participation, one for each detected species.

Table 6: Organisation of occurrence records of the ACT protocol within the Occurrence extension the "…" of ids correspond to "road1:point1".

eventID	occurenceID	basisOfRecord	scientificName	individualCount
…:particip.2018_1	…:particip.2018_1:occ1	human observation	*Lullula arborea*	2
…: particip.2018_1	…:particip.2018_1:occ2	human observation	*Columbus palumbus*	11
…: particip.2018_1	…:particip.2018_1:occ2	human observation	*Streptopelia turtur*	5
…: particip.2018_1	…:particip.2018_1:occ2	human observation	*Coturnix coturnix*	4

As for the event core, we can add various columns to describe the occurrence, according to the Darwin core: scientificName, basisOfRecord, information about geography and time, occurenceStatus, occurenceType, individualCount, lifeStage, sex. It is recommended to use controlled vocabularies for these information.

The simplest situation corresponds to recording opportunistic and naturalist observation of species (Figure 7). We create an event corresponding to "opportunistic" protocol (which can be empty apart

from the identification, or indicate location and time if they exists), and then we fit in any observation of species with all of the details: species, location, time, number, life stage, observers…

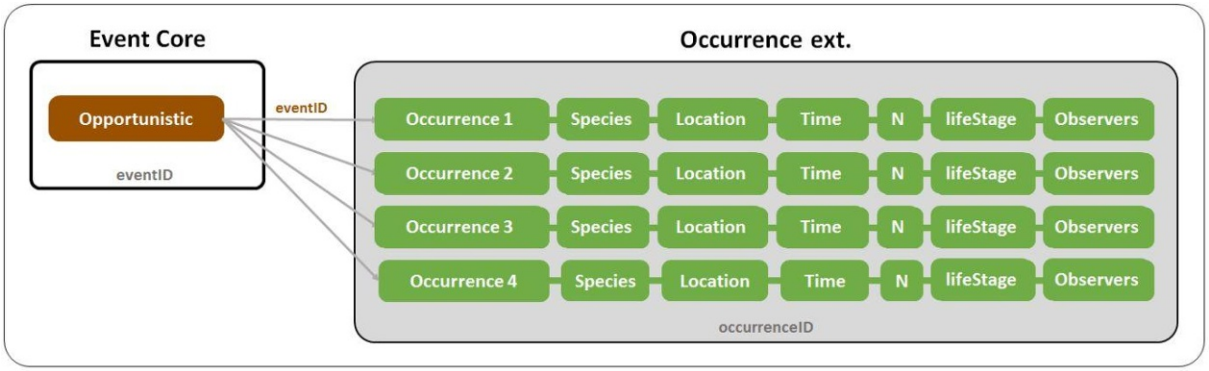


Figure 7: Simple structuration of event corresponding to an opportunistic protocol. Details of the observations are recorded in the occurrence extension.

3.2 Storing hunting bags

Hunting bags in their simplest form can be considered similarly to occurrence records, as they correspond to a number of (killed) animals, at a certain location and date.

For example, French hunting bags are collected every year at the departmental level by the ONCFS (French National Hunting and Wildlife Office). The department- hunting bags are then transmitted to ENETWILD (Table 7).

Table 7: Sample of data about French departmental hunting bag for wild boar, according to the producer format

Année de début de campagne	Espèce	Département	Nom du département	Réalisation hors parc et enclos[16]
1973	Sanglier	04	ALPES DE HAUTE-PROVENCE	400
1973	Sanglier	06	ALPES-MARITIMEs	627
1973	Sanglier	07	ARDECHE	247
1973	Sanglier	08	ARDENNES	1909
1973	Sanglier	09	ARIEGE	220
1973	Sanglier	10	AUBE	508

[16] hunting bag outside parcs and enclosure

1973	Sanglier	11	AUDE	1146
1973	Sanglier	12	AVEYRON	365

We recommend using two events levels, one corresponding to the administrative unit and the other to the hunting season, -". Hunting bags of the different species are then recorded as occurrences (Figure 8, Table 8ab).

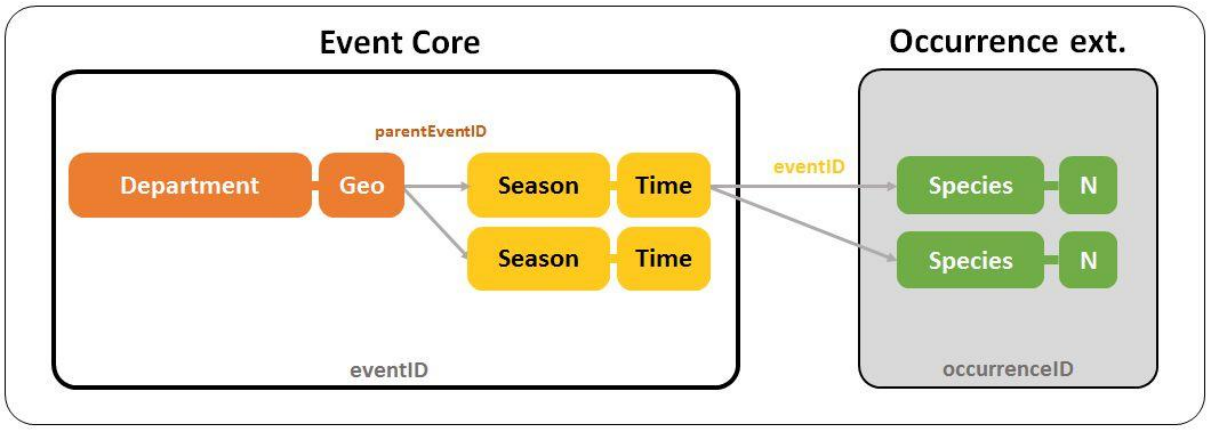


Figure 8: Modelling of the French departmental hunting bag for different seasons and time using the event core and the occurrence extension

Table 8a: Organisation of the French departmental hunting bag events in the Event Core

Parent EventID	eventID	footprint WKT	eventDate
	dep1	POLYGON()	
dep1	dep1:season2017		2017-09-01:2018-02-28
dep01	dep1:season2018		2018-09-01:2019-02-28

Table 8b: Organisation of the French departmental hunting bag events in the Occurrence extension

eventID	occurenceID	basisOfRecord	scientificName	individualCount
dep1:season2017	dep1:season2017:occ1	Hunting bag	*Sus scrofa*	5 200
dep1:season2017	dep1:season2017:occ2	Hunting bag	*Capreolus capreolus*	3 183

Using a combination of event structure and occurrence, it is therefore easy to describe hunting bags at different levels of aggregation, for instance at the departmental and management unit level, even when both are recorded within the same data set (Figure 9).

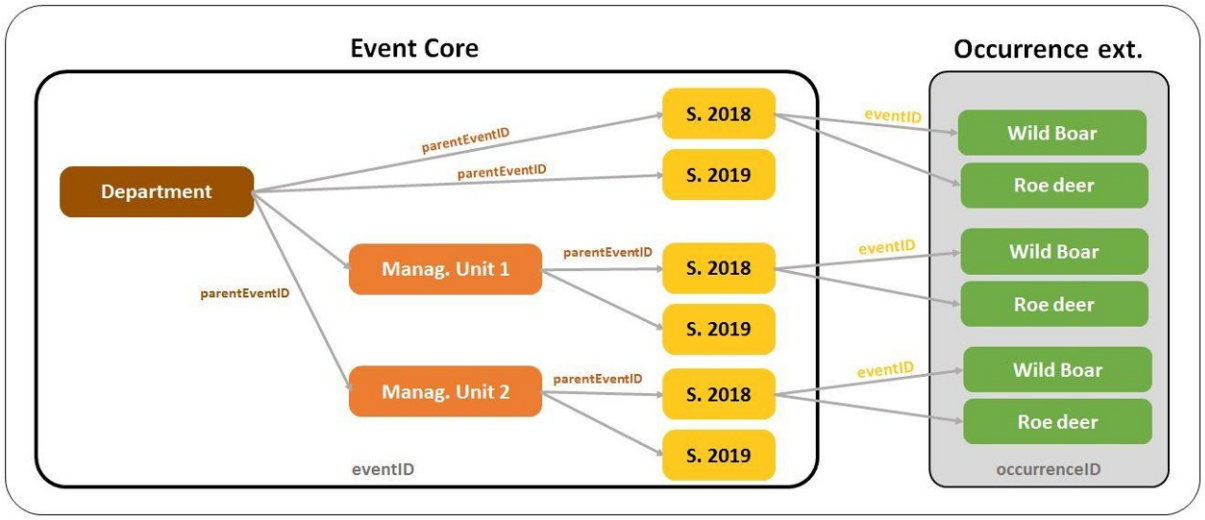


Figure 9: Modelling of hunting bag data at two different level of organisation

3.3 Case of the partial data

Another issue that we came across is the possibility of having to store partial data (i.e. different split of the same data). This especially happens when the main occurrence record corresponds to a group of individuals. Let says that we observe a group of five individuals, among which three were males, and two females, two were adults, two juveniles and one undetermined. It is easy, within the occurrence extension to record that five individuals were seen together, but it is not straightforward to record the rest of the information. Recording different simple occurrences is not a simple solution, as the link between these records would be lost, and as it will create artificial double counts. Hunting bags often include additional information about some particular individuals, but not for all. The way to record these "partial" information corresponds to the same problematic.

This information was formerly recorded in the *sex* and *lifeStage* variables as free text such as *sex* = "3 males, 2 females". The Darwin Core does not accept anymore such unstandardized values (there were more than one thousand ways to enter information in the sex variable). It makes using the column contents nearly impossible at international databases scale, and difficult to treat at our scale, even if we try to control the format within the column.

A possibility to solve this issue is to extend the Occurrence extension and allow hierarchy for occurrence records. This could be easily done by adding a *parentOccurrenceID* column in the Occurrence extension on the same model than the *ParentEventID* allows to have a hierarchy within events. This simple addition enables us to store data as structured as possible (Figure 10).

It is important to note that the occurrence having a parent occurrence id does not therefore represent pure occurrence, but different views on the parent occurrence. As a result, the sums may not be equal to the main occurrence value (i.e. a juvenile female can be counted as female, juvenile, and female-juvenile category).

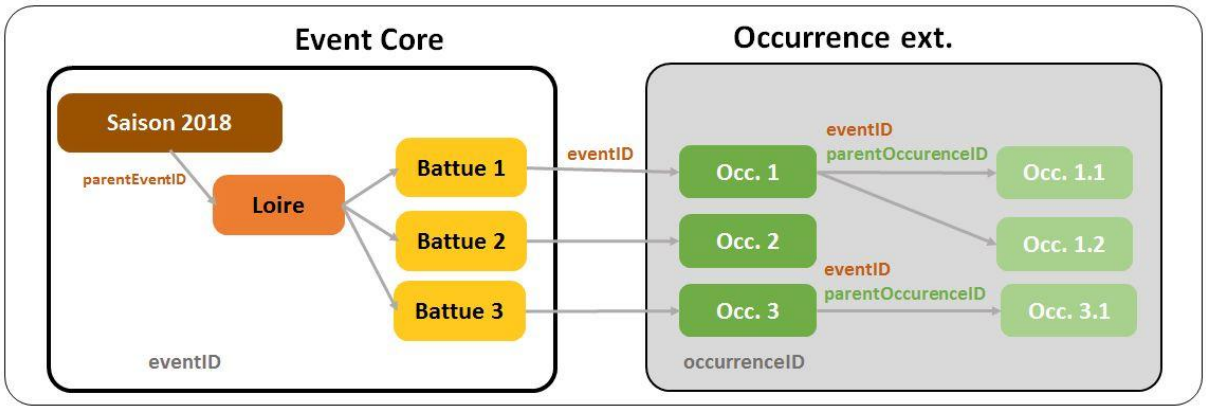


Figure 10: Modelling of partial data using nested occurrences

The data corresponding to such a structure could look like this for grouped data (Table 9ab):

Table 9a: Organisation of partial data records in the event core (no change)

eventID	parentEventID
saison2018	
saison2018/loire	saison2018
saison2018/loire/battue1	saison2018/loire
saison2018/loire/battue2	saison2018/loire
saison2018/loire/battue3	saison2018/loire

Table 9b: Organisation of partial data record in the occurrence extension

occurenceID	parentEventID	parentOccurenceID	scientificName	individualCount	sex	lifeStage
saison2018/loire/battue1/occ1	saison2018/loire/battue1		Sus scrofa	7		
saison2018/loire/battue2/occ2	saison2018/loire/battue2		Sus scrofa	5	male	adult
saison2018/loire/battue3/occ3	saison2018/loire/battue3		Sus scrofa	2		
saison2018/loire/battue1/occ1.1	saison2018/loire/battue1	saison2018/loire/battue1/occ1	Sus scrofa	3	female	
saison2018/loire/battue1/occ1.2	saison2018/loire/battue1	saison2018/loire/battue1/occ1	Sus scrofa	3		adult
saison2018/loire/battue3/occ3.1	saison2018/loire/battue3	saison2018/loire/battue3/occ3	Sus scrofa	1	female	juvenile

We believe that this suggestion enhances efficiently the Occurrence extension of the Darwin Core Archive. We would thus not be surprised if other people find uses to it. We will propose to the Darwin Core community this modification to be used at the global level.

3.4 Detailed occurrence records

Some hunting bag data sets can come with more information than those we described in the case of French department hunting bags. It is for example the case for Bern hunting bag data which provides information on each individual shot:

- Location information: hunting area which is the structuring unit, municipality corresponding to the hunt, sometimes coordinates of the animal, verbatim locality and its context such as the presence of pig husbandry on the area;
- Temporal information: date and hour;
- Individual information: weight, age class, sex, death causes.

So far, the standard does not allow to clearly catch the organisation of hunting in Switzerland, and many information are repeated, such as the geographic information about the hunting area. Some information were also recorded in the free text "notes" variable, which is not easy to use.

Using the event core and the occurrence extension now allows to model correctly this organisation (Figure 11).

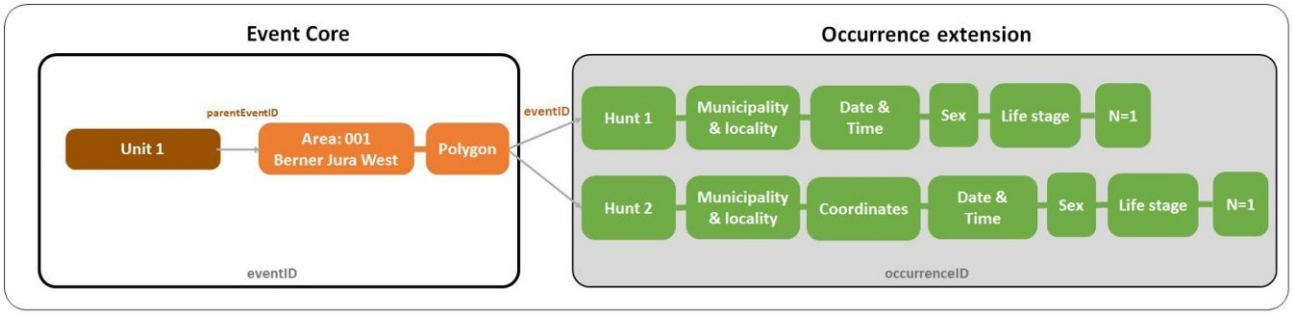


Figure 11: Modelling of data from Switzerland hunting bag records at the drive hunt level

However, the nature of the additional information provided can be variable, and it would be very difficult to know in advance all the notions that would be necessary to store it. For instance, on the above scheme of the Bern data, neither the presence of pig husbandry nor the weight of individual can be informed. We thus need a new extension, which would allow the storage of a great number of different and not yet known variables, as well as information about the event itself, such as the type of hunt or the sampling; as we will discuss in the next chapter.

4 The storage of technical records through the "extended Measurement of Facts"

As we mentioned above, the occurrence extension alone presents limitations when it comes to storing complex information, for example about the sampling effort, the methods, individual weights, or the reproductive status.

To remediate this problem, we have already used in the first Enetwild standard notions from another extension, the "MeasurementOrFact" extension. For example, notions such as "Recording method" or "Effort Value" allowed us to store some additional material on the sampling frame.

We now study the possibility offered by this frame, as well as its recent extension, to store additional and supposedly unknown variables, as well as technical and calculated data (density, abundance…).

4.1 The MeasurementOrFact extensions

The original **MeasurementOrFact (MoF)** extension from the Darwin Core Archive already allows the storage of any punctual measurement or fact such as the surface water temperature in Celsius or the kind of nest used for fishing. It must be linked directly to the "Core" of the Archive, which is most often the Event core to respect the star structure of a Darwin core archive. It is particularly useful when a wide range of information is available on the sampling procedure. The name "Measurement or Fact" means that we can record quantitative values (measurement), or qualitative values (facts).

The MoF extension consists in a small number of variables compared to other extension, which however are very flexible. The two first corresponds to the identification: *eventID*, for the link to the event measured, and the *measurementID* itself. The three others form a group describing the measurement: the column *measurementType* describes the meaning of the measurement (e.g. nest type, temperature). It is followed by the *measurementValue* which correspond to the value measured or stated (e.g. nest type, 15). Finally, the *measurementUnit* is used to store the unit, or can be left empty if the record is a fact (e.g. empty, degree Celsius). Of course, while this extension is very flexible, we need to agree on control vocabularies to easily share and use data within a project. However, data are still usable and well described for other projects to use them.

A drive hunt can therefore be described with two events (one being the management unit, and the other the drive hunt itself). The many technical information about the hunt (number of baiters, of dogs, of hunters, nº animals sighted, beaten surface, habitat, weather) can then be stored separately within the MoF extension and the hunting bag of this drive hunt, while the information on each species will be stored using the occurrence extension (Figure 12, Table 10abc).

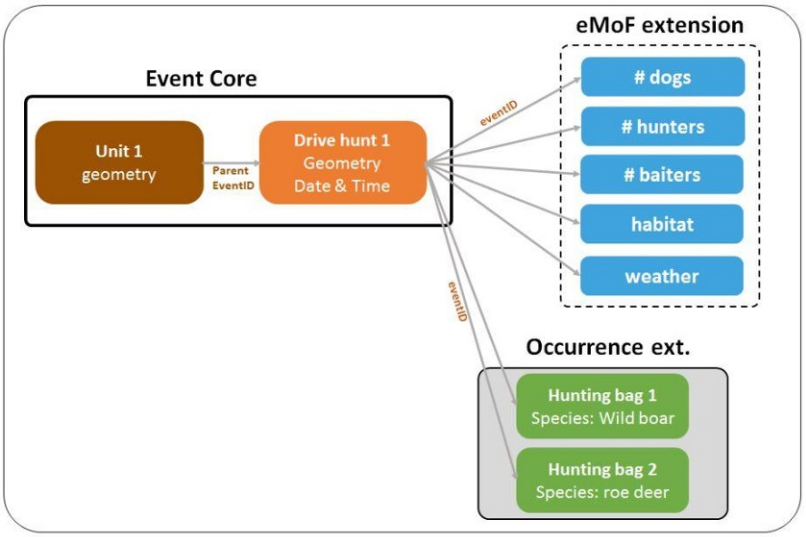


Figure 12: Modelling of drive hunt data hunting bag and hunting pressure

Table 10a: Organisation of the drive hunt data in the Event core

eventID	parentEventID	footprint WKT	eventDate
Unit1		POLYGON (…)	
Unit1/driveHunt1	Unit1	POLYGON (…)	16-10-2019

Table 10b: Organisation of the drive hunt data in the Occurrence extension

occurenceID	parentEventID	scientificName	individualCount
unit1/driveHunt1/occ1	unit1/driveHunt1	Sus scofra	10
unit1/driveHunt1/occ2	unit1/driveHunt1	Capreolus capreolus	5

Table 10c: Organisation of the drive hunt data in the Measurement or Fact extension

measurementID	parentEventID	measurementType	measurementValue	measurementUnit
unit1/driveHunt1/mea1	unit1/driveHunt1	effort in dogs	10	individual
unit1/driveHunt1/meas2	unit1/driveHunt1	effort in hunters	15	individual
unit1/driveHunt1/meas3	unit1/driveHunt1	effort in baiters	17	individual
unit1/driveHunt1/meas4	unit1/driveHunt1	habitat	forest	
unit1/driveHunt1/meas5	unit1/driveHunt1	weather	rain	

4.2 The extended Measurement or Fact extension

OBIS recently introduced the extended MeasurementOrFact extension (hereafter eMoF) (Pooter et al. 2017). It extends the MeasurementOrFact extension by adding several variables: *occurrenceID, measurementTypeID, measurementValueID* and *measurementUnitID*.

The *measurementTypeID*, *measurementValueID*, *measurmementUnitID* offers a way to present values coming from normalized protocol, where the value and unit are internationally fixed for a given type of measurement, and for which a unique identification exists. We have little use for those variables for now.

The new column that is of most interest to us is the *occurrenceID* column, which creates a link between an eMoF record and an occurrence record. Consequently, the eMoF extension can be used to store qualitative and quantitative data about sampling events and now about occurrences, such as the weight of an organism in kilograms or the cause of death (Figure 13, table 11abc).

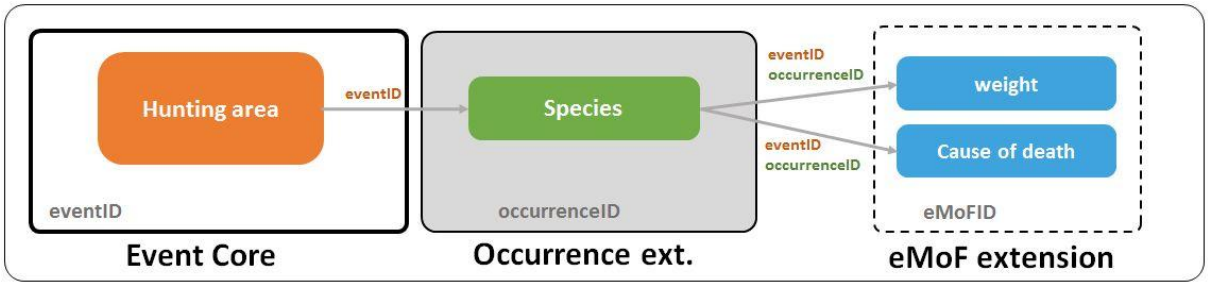


Figure 13: Modelling of hunting data attached to each individual

Table 11a: Organisation of the individual information from the hunting data in the Event core

eventID
HuntingArea1

Table 11b: Organisation of the individual information from the hunting data in the Occurrence extension

occurenceID	parentEventID	scientificName
HuntingArea1/occ1	HuntingArea1	Sus scrofa

Table 11c: Organisation of the individual information from the hunting data in the eMoF extension

measurementID	parentEventID	parentOccurrenceID	measurementType	measurementValue	measurementUnit
HuntingArea1/occ1/mea1	HuntingArea1	HuntingArea1/occ1	weight	53	kilogram
HuntingArea1/occ1/mea2	HuntingArea1	HuntingArea1/occ1	cause of death	hunting shot	

This enhancement to the classical MeasurementOrFact extension is really powerful as it allows us to store additional data about any other record from the Core or from the Occurrence extension. The nature of this additional information does not have to be known in advance, and thus does not requires the use of several predefined columns in the Occurrence Core. At the management unit level, this information could concern the area of the sampling unit, the presence of dogs or pig husbandry. At the transect level, it could be the sampling effort in number of observers or in kilometres drove, or the weather on that day. For each observation of a distance sampling protocol, the distance to the animal, angle, and some covariates if needed could be stored as well using the same table (Figure 14, table 12abc).

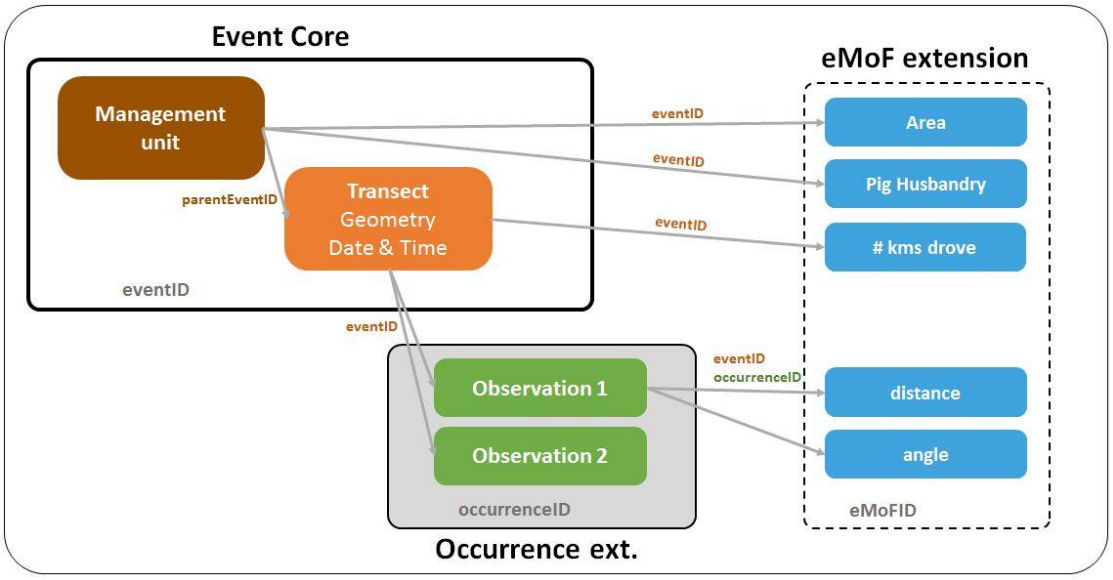


Figure 14: Modelling of data from a distance sampling protocol including studied area details

Table 12a: Organisation of the information from the distance sampling protocol in the Event core

eventID	parentEventID	footprintWKT	eventDate
Unit1		POLYGON (…)	
Unit1/Transect1	Unit1	LINE (…)	16-10-2019

Table 12b: Organisation of the information from the distance sampling protocol in the Occurrence extension

occurenceID	parentEventID	scientificName
Unit1/Transect1/occ1	Unit1/Transect1	Sus scrofa
Unit1/Transect1/occ2	Unit1/Transect1	Sus scrofa

Table 12c: Organisation of the information from the distance sampling protocol in the eMoF extension

measurementID	ParentEventID	parentOccurrenceID	measurementType	measurementValue	measurementUnit
Unit1/mea1	Unit1		Area	150	square kilometer
Unit1/mea2	Unit1		Pig husbandry presence	Yes	
Unit1/Transect1/mea1	Unit1/Transect1		effort distance	15	kilometer
Unit1/Transect1/occ1/mea1	Unit1/Transect1	Unit1/Transect1/occ1	distance	26	meter
Unit1/Transect1/occ1/mea1	Unit1/Transect1	Unit1/Transect1/occ1	angle	45	degree

Consequently, in the case of the Bern hunting bags that we mentioned above, all of the information recorded can now be stored in the DwC: individual weights will go in the eMoF linked to the occurrence extension as well as the presence of pig husbandry, both linked to the Event core.

4.3 The storage of statistical values: introducing the "nested eMoF"

The Enetwild project is also interested in sharing density estimates and abundance indexes assessed through various protocols. Estimation of population size or estimation of total hunting bag obtained from different national surveys are very similar data. We can generalise the subject saying that we need to store data associated with statistical analyses in the Darwin core Archive.

We can store summary values about an event or an occurrence using the eMoF extension: for example, the population density estimated by a distance sampling analysis can be stored in an eMoF attached to an occurrence attached to the surveyed area (Figure 15, table 13).

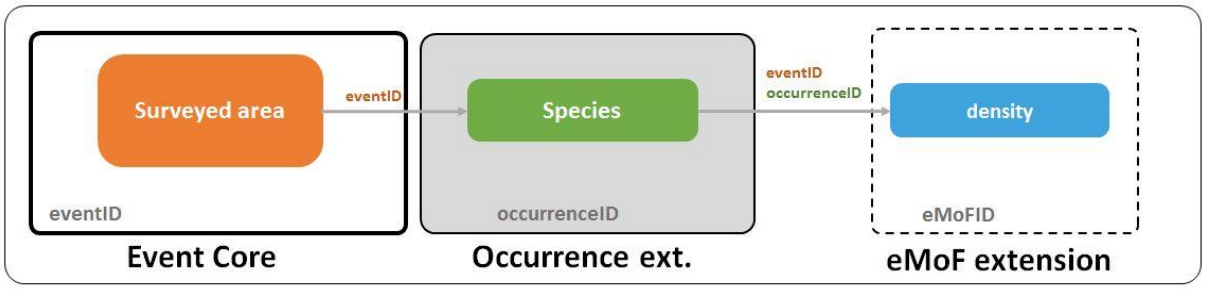


Figure 15: Modelling of a density data

Table 13: Organisation of a density data in the eMoF extension. The organisation of the event core and of occurrence extension are not displayed

measurementID	parentEventID	parentOccurrenceID	measurementType	measurementValue	measurementUnit
Unit1/occ1/mea1	Unit1	occ1	density	3	individual per square kilometer

This is not enough to truly assess the reliability of this estimate. Estimates are necessarily associated with different measures of their precisions, such as 95% confidence intervals, variances, standard errors, coefficient of variations. We can go further saying that the form of the distribution of the estimation (Gaussian, binomial, non-parametric), the inference type (frequentist, Bayesian), or even the variance of the estimation of the variation of the estimate (for pure statistician) could be useful for a good use of the data or for further analyses, and therefore need to be recorded in the Darwin Core Archive. At the best of our knowledge, these data cannot be recorded in the Darwin core archive as it is not possible to link a measurement (e.g. a variance) to another measurement (e.g. the punctual estimate).

This observation leads us to propose a small adjustment of the extended measurement or fact extension. We propose to add a new variable corresponding to the "*parentMeasurementID*". Therefore, there would be four variables for the measurement identification: parentE*ventID*, parentO*ccurrenceID*, *parentMeasurementID*, *measurementID*, in addition to the *datasetID* which is unvariable for the whole dataset and link it to the metadata.

This proposition is similar to the one proposed by EU-BON for nesting events within other, and to the one above where occurrence could also be nested within others. It would allow us to model a 1 to n relationship between measure, which means nesting measurements within each other, and thus making it possible to describe the real pattern (the estimate, and its precision(s)). We propose to name this extension the nested measurement or fact extension (neMoF hereafter).

For instance, after long sequences of capture in the Trois-Fontaines forest (an enclosed and experimental research station of the Office Français de la Biodiversité), the research team was able to estimate the density of roe deer population and its evolution through years using a Capture Mark Recapture method (Table 14).

Table 14: Density estimation of roe deer in Trois Fontaines Forest associated with their precisions, according to the producer database format.

Année	N	N min	N max
2005	440	357	560
2006	306	247	394
2007	200	164	255
2008	145	118	186
2009	176	142	227
2010	164	141	199
2011	197	169	240
2012	270	214	353
2013	233	196	285
2014	240	202	297
2015	144	122	177

This information can be stored in the Darwin Core, taking advantages of the proposed neMoF extension (Figure 16, Table 14abc). As for any variable in the standard containing multiple values, the two limits of the interval are concatenated by a vertical bar "|". A light data manipulation will then be necessary to separate them for further use.

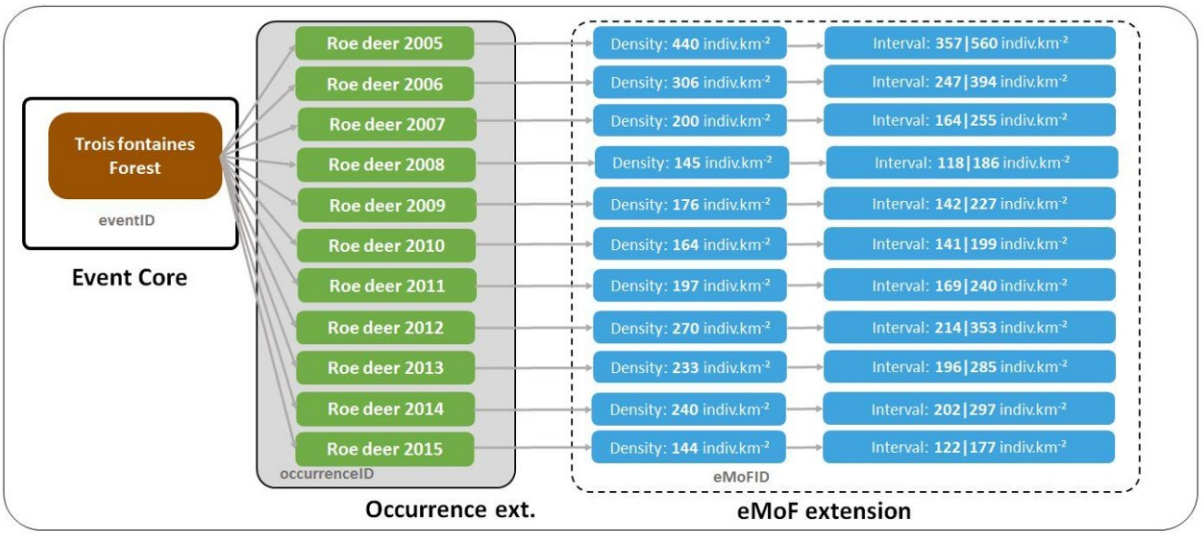


Figure 16: Modelling of roe deer density estimation data in Trois Fontaines Forest using the proposed neMoF extension

Table 14a: Organisation of the roe deer density estimation data in the Trois Fontaines Forest (3FF) the Event core

eventID
3FF

Table 14b: Organisation of the roe deer density estimation data in the Occurrence extension

occurenceID	parentEventID	scientificName	date
3FF/occ2005	3FF	Capreolus capreolus	2005
3FF/occ2006	3FF	Capreolus capreolus	2006
3FF/occ2007	3FF	Capreolus capreolus	2007
3FF/occ2008	3FF	Capreolus capreolus	2008
3FF/occ2009	3FF	Capreolus capreolus	2009
3FF/occ2010	3FF	Capreolus capreolus	2010
3FF/occ2011	3FF	Capreolus capreolus	2011
3FF/occ2012	3FF	Capreolus capreolus	2012

3FF/occ2013	3FF	Capreolus capreolus	2013
3FF/occ2014	3FF	Capreolus capreolus	2014
3FF/occ2015	3FF	Capreolus capreolus	2015

Table 14c: Organisation of the roe deer density estimation data in the proposed neMoF extension

measurementID	parentEventID	parentOccurrenceID	parentMeasurementID	measurementType	measurementValue	measurementUnit
3FF/occ2005/mea1	3FF	3FF/occ2005		density	440	individual per square kilometer
3FF/occ2006/mea1	3FF	3FF/occ2006		density	306	individual per square kilometer
3FF/occ2007/mea1	3FF	3FF/occ2007		density	200	individual per square kilometer
3FF/occ2008/mea1	3FF	3FF/occ2008		density	145	individual per square kilometer
3FF/occ2009/mea1	3FF	3FF/occ2009		density	176	individual per square kilometer
3FF/occ2010/mea1	3FF	3FF/occ2010		density	164	individual per square kilometer
3FF/occ2011/mea1	3FF	3FF/occ2011		density	197	individual per square kilometer
3FF/occ2012/mea1	3FF	3FF/occ2012		density	270	individual per square kilometer
3FF/occ2013/mea1	3FF	3FF/occ2013		density	233	individual per square kilometer
3FF/occ2014/mea1	3FF	3FF/occ2014		density	240	individual per square kilometer
3FF/occ2015/mea1	3FF	3FF/occ2015		density	144	individual per square kilometer
3FF/occ2005/mea1/mea1	3FF		3FF/occ2005/mea1	interval	357|560	individual per square kilometer
3FF/occ2006/mea1/mea1	3FF		3FF/occ2006/mea1	interval	247|394	individual per square kilometer
3FF/occ2007/mea1/mea1	3FF		3FF/occ2007/mea1	interval	164|255	individual per square kilometer
3FF/occ2008/mea1/mea1	3FF		3FF/occ2008/mea1	interval	118|186	individual per square kilometer
3FF/occ2009/mea1/mea1	3FF		3FF/occ2009/mea1	interval	142|227	individual per square kilometer
3FF/occ2010/mea1/mea1	3FF		3FF/occ2010/mea1	interval	141|199	individual per square kilometer

3FF/occ2011/mea1/mea1	3FF		3FF/occ2011/mea1	interval	169|240	individual per square kilometer
3FF/occ2012/mea1/mea1	3FF		3FF/occ2012/mea1	interval	214|353	individual per square kilometer
3FF/occ2013/mea1/mea1	3FF		3FF/occ2013/mea1	interval	196|285	individual per square kilometer
3FF/occ2014/mea1/mea1	3FF		3FF/occ2014/mea1	interval	202|297	individual per square kilometer
3FF/occ2015/mea1/mea1	3FF		3FF/occ2015/mea1	interval	122|177	individual per square kilometer

Similarly, the Réseau Loup-Lynx in France, a participative surveying network, estimates, based on capture-mark-recapture method on genetic samples, the number of wolfs at 256 individuals [IC95 189-296] in 2016, 357 [265-402] in 2017, 430 [387-477] in 2018, 527 [477-576] in 2019 for France.

These values can be entered in the DwC-A using the neMoF extension (Figure 17, Table 15abc).

Figure 17: Modelling of the population size estimation data for French wolves using the proposed neMoF extension

Table 14a: Organisation of the population size estimation data for French wolves in the Event core

eventID	country
France	FR

Table 14b: Organisation of the population size estimation data for French wolves in the Occurrence extension

occurenceID	parentEventID	scientificName	date
France/occ2016	France	Canis lupus	2016
France/occ2017	France	Canis lupus	2017
France/occ2018	France	Canis lupus	2018
France/occ2019	France	Canis lupus	2019

Table 14c: Organisation of the population size estimation data for French wolves in the proposed neMoF extension

measurementID	parent EventID	parent occurrenceID	parentMeasurementID	measurementType	measurementValue	measurementUnit
France/occ2016/mea1	France	France/occ2016		Population size	256	indiv.km^{-2}
France/occ2017/mea1	France	France/occ2017		Population size	357	indiv.km^{-2}
France/occ2018/mea1	France	France/occ2018		Population size	430	indiv.km^{-2}
France/occ2019/mea1	France	France/occ2019		Population size	527	indiv.km^{-2}
France/occ2016/mea1/mea1	France	France/occ2016	France/occ2016/mea1	Interval	189|296	indiv.km^{-2}
France/occ2017/mea1/mea1	France	France/occ2017	France/occ2017/mea1	Interval	265|402	indiv.km^{-2}
France/occ2018/mea1/mea1	France	France/occ2018	France/occ2018/mea1	Interval	387|477	indiv.km^{-2}
France/occ2019/mea1/mea1	France	France/occ2019	France/occ2019/mea1	Interval	477|576	indiv.km^{-2}

Advantages offered by the neMoF extension compared to other solutions (for instance using the resource relationship extension, or creating a new extension) are not fully discussed here. There are however three important characteristics to highlight. Firstly, this extension respects the spirit of the OBIS enhancement, and just extends its capacity. Secondly, the hierarchy of measurement is powerful, safe and understandable while using the *parentMeasurementID* and it represents actual statistical values: level 1 corresponds to punctual estimates, level 2 to variances/confidence intervals of punctual estimates, level 3 to variances/confidence intervals of variances of punctual estimates, etc…. Thirdly, we need a relatively short list of controlled vocabulary to describe all of these factors (variance, and confidence intervals on the above example), which can be fixed by the international statistical community.

5 Recording methodological details in data and metadata

Estimated values in ecology strongly rely on the use of the appropriate field and statistical methods. The purpose of a data standard cannot be to define which method is a good fit for particular goals, nor which data should be recorded according to their reliability. The standard must guarantee that a scientist will find all of the necessary information to judge /the fitness-for-use of the data for the analysis. However, data standards are not methodological papers, and all of the information, sometimes very specific, cannot be recorded. The amount of methodological details about the sampling and the statistical analysis to be included as data or metadata is therefore an essential question.

One extreme answer is to record everything, for the best use of the data and its full comprehension. There are two ways to describe all of the methodological details about an estimation: a link to the published associated article, and a free text protocol in the EML metadata. Methods are currently stored in semi-organized text storage in the EML metadata file with tags: <methods>, <methodStep>, <sampling>, <qualityControl>. The drawback of this choice is that information, while

being here, is hard to find and query and need a large amount of time to be read and understood. So the extreme answer "everything" is very close to the other extreme being "none". A link to the published paper (which could be a data paper) will be a useful and essential option in the metadata, but is not sufficient.

It seems that there are two types of information that we can store: information on the **type of sampling** (particularly important for surveys, but not only); and information on the **statistical methods**. The idea is here to have vocabulary guidelines or controlled vocabularies to facilitate the sharing of these data sets. To reduce the number of information to control, we could fix a target to this information. They must allow the reader to correctly use the data, but they do not have to allow the reader to perform again the analysis, this would be the aim of the published paper and tools developed for a reproducible science.

Another aspect to take into account corresponds to the location of the information: either in the data themselves, or in the metadata. To correctly facilitate the decision, we should keep in mind that metadata describe the entirety of the dataset and are used while searching datasets and, once found, to understand them. However, while using the dataset, the metadata can be disregarded and the only information that will be used are found in the data themselves. For instance, the metadata will allow a researcher to set up and parametrize a script analysing or using the data, but once done, the script only runs on data. Some types of information can be data and/or metadata depending on whether they concern the whole data set, or only a part of it (e.g. sample sizes or analysis methods) without problem.

There are **four concepts** requiring methodological details: Sampling scheme, sampling effort, analysis, and estimated values..

The first concept corresponds to the *sampling scheme* and will describe type of sampling (e.g. random, non-random), the sampling frame, the planned and realised sample size, and sample weights given to data. All of this information could be useful while describing a dataset of raw data, but only the most essential ones are necessary while describing a dataset of estimated values. The sampling frame and the type of sampling correspond to metadata, and the planned and realised sample sizes can be useful in the metadata to select for instance large datasets. These sample sizes can be also included into the data in case of multistage sampling if appropriate. Sample weights are data and not metadata, and can describe the different step of a sampling scheme. The names of the variables used to stratify the sampling scheme are metadata, as well as names of measured covariates, if any, while values of these variables for each event or occurrence are recorded in the data. Sampling information within data can be described in the neMoF extension attached to the appropriate event (Figure 18).

The second concept corresponds to the *sampling effort*, a key component for analysis. The notion of effort regroups any information about the force put on a sampling unit, and therefore can take many aspects: area or length surveyed, duration of the survey, number of observers, number of visits to a particular transect or point. These are particular notions that are often summarized into a single total effort value according to each project, for example by multiplying the number of observers by the number of sampled areas. Efforts mostly correspond to values attached to the finest level of events, where occurrences are recorded. However, the theoretical protocol can be described in the metadata using these values. Summarized values have to be defined in metadata, and the total effort value can be recorded in the metadata, and at the different event levels in the data if pertinent, using the neMoF extension. Additional types of effort can be identified for specific fields, such as the effort given in number of hunters, baiters and dogs for a drive hunt (Figure 19).

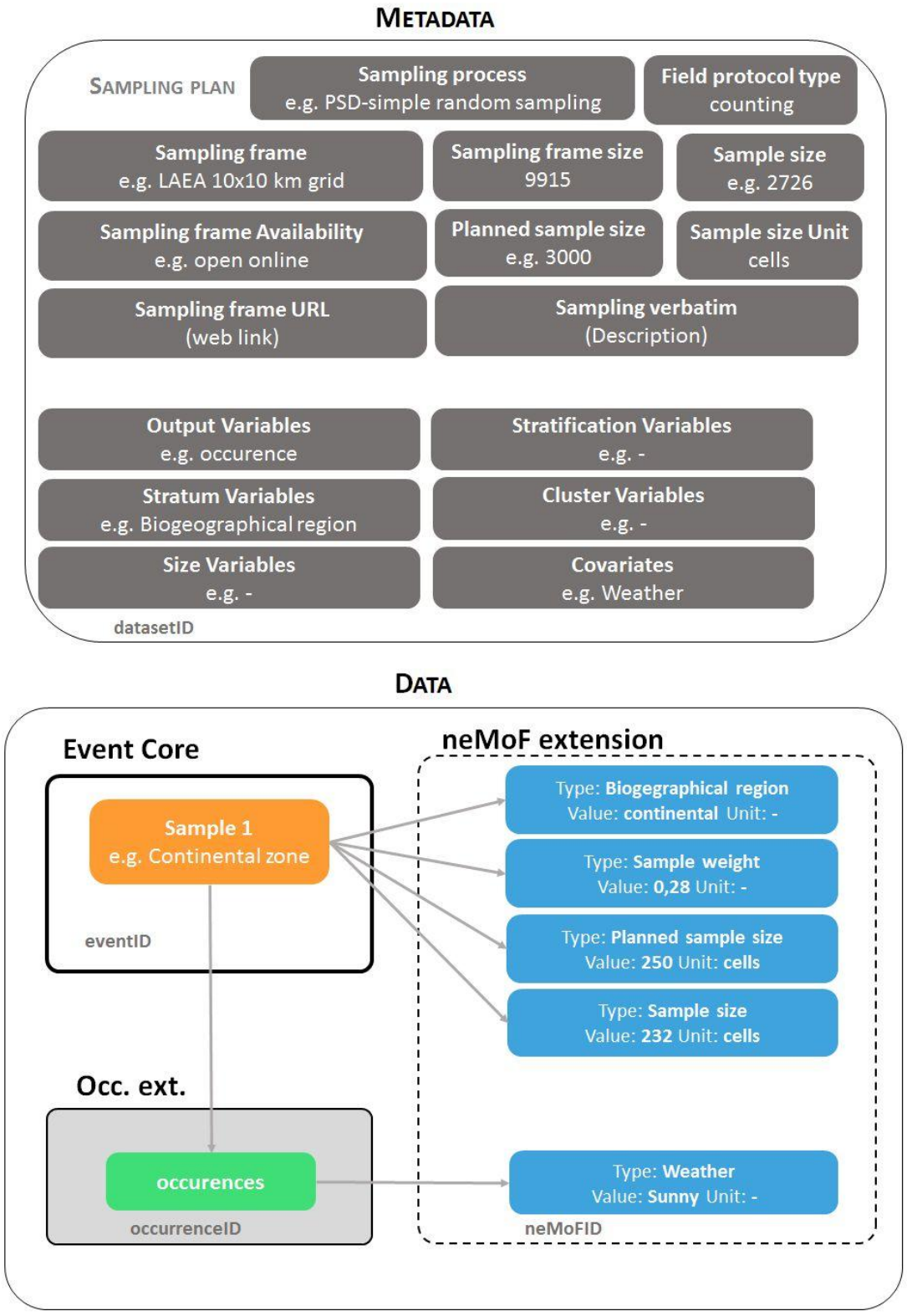


Figure 18: Scheme of variables in data and metadata with a focus on variables conceptually related to sampling scheme

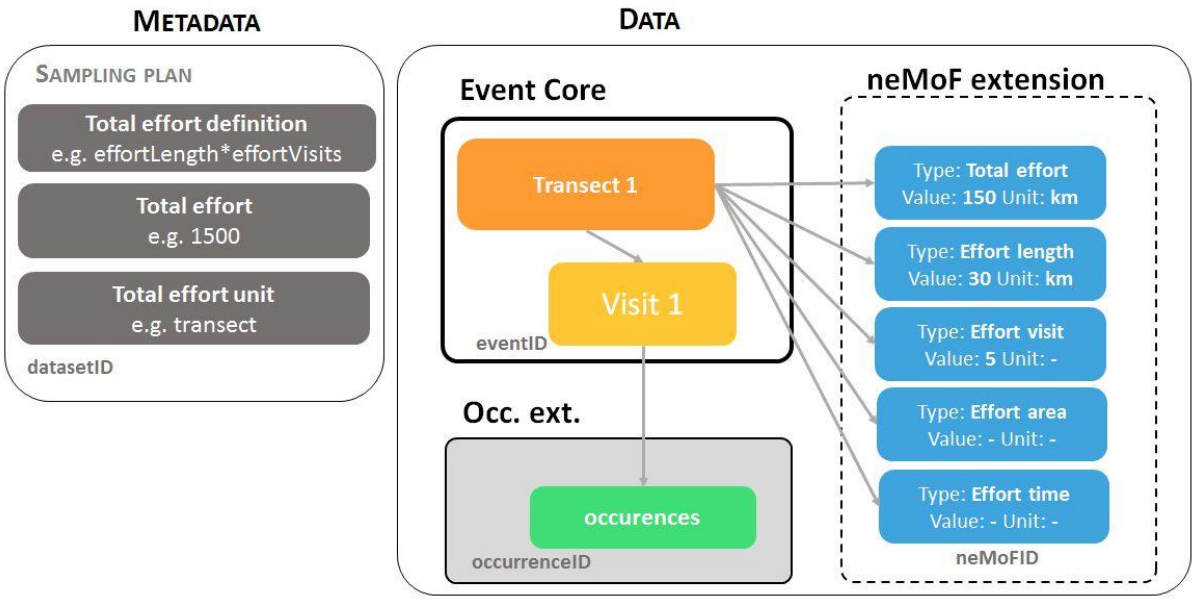


Figure 19: Scheme of variables in data and metadata with a focus on variables conceptually related to sampling effort. Example is for 10 transects of 30 km each, visited 5 times each. Of course, if transects lengths are different, this information will be recorded to each transect level and not in the metadata, but the total effort value and the number of visits per transect can still be recorded in the metadata. In this example, the area surveyed and the duration of the survey are not pertinent information.

The third domain corresponds to the *description of the analysis*: sourcing the data used (e.g. link to other datasets, data sample size), and identifying the broader analysis family (e.g. CMR, distance sampling, census). Then we can record information about the method itself: inference type, model selection method, degrees of freedom, included covariates, methods for interval estimations and even the script and software used. Most of these variables are found in the metadata, except if different methods are used. In such a case, this information can be recorded into the neMoF extension attached to the appropriate event or occurrence (Figure 20).

The fourth domain corresponds to details about the *resulting values* themselves: the type of variable estimated (e.g. population size, density, relative abundance, hunting bag, survival rate, capture rate) and their precisions which include the statistical distribution of the value, the interval born and type, variance, standard errors, but also p-quantile or model/distribution parameters. The type of variables estimated (i.e. the outputs of the analysis) are recorded in the metadata as it is a key information to select the dataset, and it is also included in the neMoF extension in the data associated with the estimated value. Interval type and their distribution are also recorded in metadata, and if there are different type for each in the output, they can also be recorded in the neMoF attached to the interval values as a nested fact. All other information corresponds to data and are included in the neMoF extension attached to the estimated value (Figure 21).

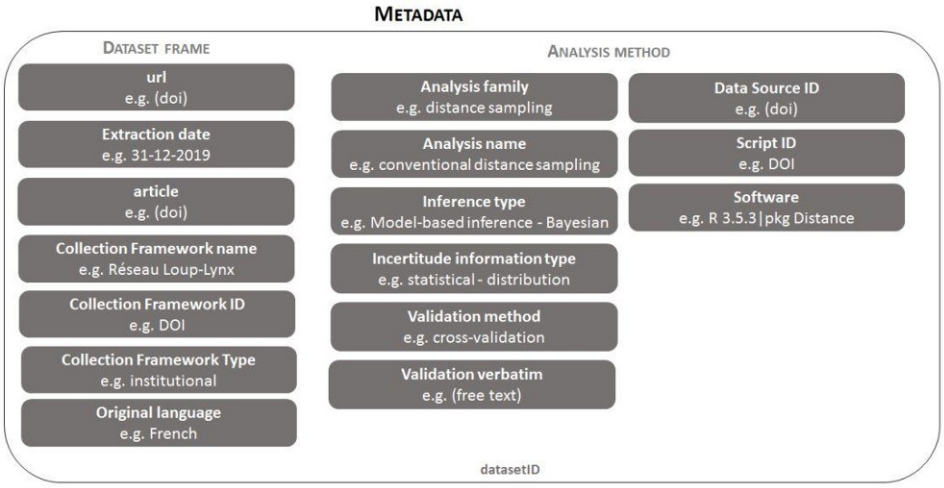


Figure 20: Scheme of variables in metadata with a focus on variables conceptually related to an analysis. The results of the analysis correspond to data.

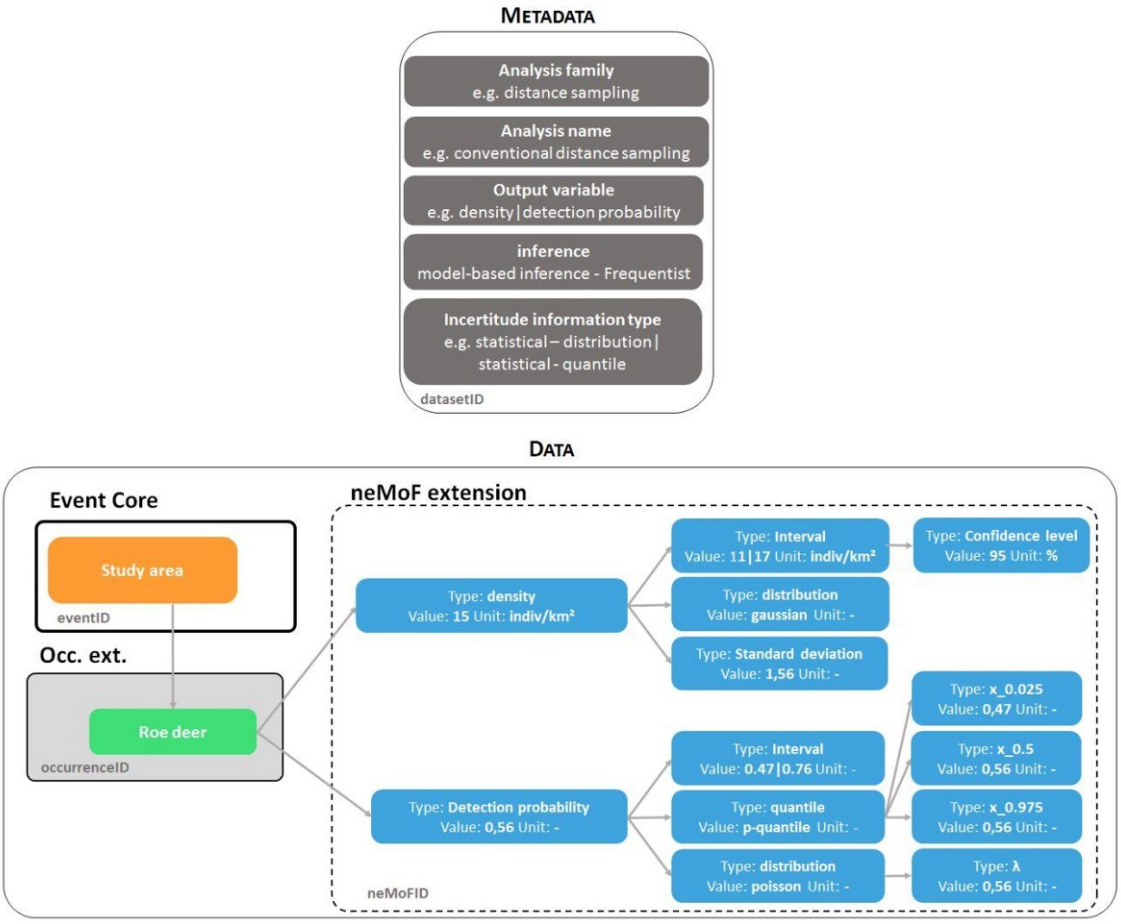


Figure 21: Scheme of variables in data and metadata with a focus on variables conceptually related to outputs of a statistical analysis within the Darwin core. In the example, an estimation of a roe deer population on a study area using a distance sampling method provides two results: the density itself and the detection probability, as well as their precision.

6 Advices while working on datasets

6.1 Defining datasets

A common difficulty of researchers or data managers wanting to share their data is to correctly define the perimeter of a given dataset: are raw and result data a single dataset, are data resulting from the same protocol but applied to different place a single dataset, are yearly data of a program a single dataset? There are many ways to organize data and to split them in different datasets according to protocols, analysis, articles, survey episode, or even according to the targeted public. The first rule is that there are no wrong ways to group data, and no problem if some data are found in different datasets (e.g. datasets presenting data used in different articles can overlap). Any way that suits the defined purpose is acceptable.

The second rule is that the data grouped together will have the same metadata in common. For best clarity, we advise to separate "raw data" from "summarized data" from "results of analysis". Indeed, this information will rarely be used together at the same time, and workflow can be best followed using this separation. Writing metadata, and structuring events will also be easier. Indeed, data structure from raw, summarized and results data can be different. Raw data are data as they are collected on the field, after being cleaned up and verified, summarized data are secondary data, processed from raw or other summarized data with little or no statistics used (e.g. sum or number of records), while results of analysis are processed data used after a particular statistical procedure. The basis of records used in the occurrence extension must therefore be completed for the results of analysis data, as they are neither human observation nor machine observation. Results of analysis are not even evidence of presence, as it could only be a probability or suitability of presence. We suggest adding the term *StatisticalAnalysis* to the accepted terms, with the definition "An output of a statistical analysis", and with example "Probability of presence, measure of suitability, density estimation". In contrast, summarized and raw data are well described by either the *HumanObservation* or the *MachineObservation* values.

These different datasets share a common basis. The *Collection Framework* variable of metadata offers a way to link them under the same framework. As we will see further (recording methodological details), a *Data source* variable allows referencing datasets ids used by an analysis to produce the datasets corresponding to results.

6.2 Structure of a dataset

The organization of metadata, events, occurrences and measurements or fact may appear complex. We try here to summarize the associated concepts and purposes.

First, information in metadata are applicable to the whole dataset, and are used to find the dataset and to understand it.

Second, events describe the structure of the data and can therefore be defined before the collection of data. Measurements or facts associated to an event correspond to abiotic variables, either environmental values, or methodological, sampling information. Location and time must describe the event itself, the protocol of collection, independently of the observations made. The structuration of the event would best be organized with the most stable concepts on top, and the most variable (the participation) at the bottom of the structure.

Third, occurrences are biological records, the only location where we can find the species field (i.e. scientificName). Location and time included in the occurrence extension describe the occurrence itself, not the protocol of collection. Furthermore, if a location or a time is known before the protocol is performed (for example, a transect that was decided to be walked through), it is an event. By contrast, if the location and time could not be known before the field implementation (for example,

the observation of an individual within the transect on a precise point), it will be recorded as an occurrence. It is thus common to store records without any indication on the location or time in the Occurrence extension if this information is already contained in the Event core (for example, in the case of hunting bags where the frame of the survey is decided before the collection starts and is therefore an event). It is also possible to have information on date and time only in the Occurrence extension, for example in the case of opportunistic data (observations made at unpredictable places and dates). Measurements or facts which are attached to an occurrence must describe this occurrence, and thus are biological records.

Last, measurements or facts attached to other measurements of facts are logically statistical details ("measures about measures").

7 Recommendations

We recommend using the Darwin Core Standard as the basis for the wildlife monitoring standard for sharing data about wildlife occurrence, abundance and hunting bags. We advocate enhancing the Darwin Core Standard with our proposition to allow it to record data coming from statistical estimations, such as density, reproductive success, mortality rate.

It seems necessary to use at a European level an international well developed and used data standard such as the Darwin Core Standard.

Appendix A Lists of fields

New variables relative to the Wild boar data model are indicated in bold.

Metadata

TABLE	CONCEPT	VARIABLE	DEFINITION	TYPE	N	EXAMPLE	EML	ISO
METADATA	Dataset Identification	title	Title of the dataset	string	1		<title>	
		citation	A verbatim reference for the resource as a statement indicating how this record should be cited (attributed) when used.	string	1		<citation>	
		datasetID	An identifier for the set of data. May be a global unique identifier or an identifier specific to a collection or institution.	string/doi	1		<dataset>	
	Dataset frame	url	URL (Uniform Resource Locator) where the dataset can be found	url	1	https://enetwild.com/reports-docs/	<onlineUrl	
		extractionDate	Date on which the dataset was extracted from its original location.	date	1	01-02-2019		
		article	Name of the article from which the data is extracted	string/doi	1		<article>	

		collectionFrameworkName	Name of the program or collection framework from which the dataset come from	string	1	Réseau Ongulés sauvages ONCFS-FNC-FDC		
		collectionFrameworkID	identification, if any, of the program or collection framework	string/doi	1			
		collectionFrameworkType		mtd_fwktyp	1			
		originalLanguage	Language in which the data was originally written	string	1	French	xml:lang=	
	Provider identification	provName	Name of the person who sent the dataset	string	1	Smith	<givenName>	
		provSurname	Surname of the person who sent the dataset	string	1	John	<surName>	
		provAffiliation	Name of the organization which the data provider comes from	string	1	ONCFS	<organizationName>	
		provAddress	Adress of the data provider	string	1	5, saint-Thibault str.	<address>	
		provEmail	Email of the data provider	string	1	John.smith@example.com	<electronicMailAddress>	
		provPhone	Phone number of the data provider	string	1	01 01 02 04 01	<phone>	
		provPostCode	Postal Code of the data	string	1	78610	<postalCode>	

		provider					
	provCity	City of the data provider	string	1	Auffargis	<city>	
	provCountry	Country of the data provider	loc_country	1	France	<country>	ISO3166-1 alpha2
	role	Role the data provider played in the acquirement of the data	mtd_role	1	point of contact	<role>	
Accessibility	dsaSigner	Name of the responsible person who signed the data sharing agreement	string	1	John Smith		
	dsaDate	Data the data sharing agreement was signed	date	1	01-02-2019		
	accessibility	Definition of the frame in which the data can be used and shared	mtd_access	1	public database on request		
	gbifAuthorization		yes/no	1	yes		
Technical	descriptionVerbatim		string	1			
	datasetType	Type of data composing the dataset	mtd_datasettyp	1	summarized data		
	updateFrequency	Frequency at which the data is collected	mtd_upfreq	1	biannually		

Geographical coverage	countryCode	NUT code of the country where the data was collected	loc_country	n by|	FR		ISO3166-1 alpha2
	geographicScale	Georgraphic coverage of the data set	loc_geoscale	1	municipality		
Temporal coverage	beginDate	Date on which the collecting of the data started	date	1	01-02-2017		
	endDate	Date on which the collecting of the data ended, if relevant	date	1	01-03-2018		
	temporalResolution	The resolution of temporal information	time_scale	1	month		
Biological coverage	Taxon	List of the species covered by the data set	bio_species	n by|	Capreolus capreolus|Sus scrofa|Cervus elaphus		
Sampling plan	**samplingProcess**	what it the process used to obtain the sample	samp_process	1	PSD-simple random sampling		
	samplingFrame	what is the statistical population sampled	string	1	Grid_ETRS89_LAEA_10K_FR		
	samplingFrameAvailability	Are the sampling frame data available	samp_availability	1	open online		
	samplingFrameUrl	link to the sampling frame	url/doi	1	https://www.eea.europa.eu/data-and-maps/data/eea-reference-		

		if online			grids-2		
	samplingFrameSize	how large is the sampling frame	integer	1	9915		
	plannedSampleSize	what was the planned sample size	integer	1	3000		
	sampleSize	Realized or resulting number of sampling units	integer	1	2726		
	sampleSizeUnit	Common unit for the frame, the planned and realized/effective sample size	string	1	cell		
	fieldProtocolType	type of protocol applied to collect data per sample unit	samp_prot	1	counting		
	totalEffortDefinition	Technical definition of the total effort	string	1	EffortLength*EffortVisits		
	totalEffort	Value of the total effort	numeric	1	1762		
	totalEffortUnit	Unit of the total effort	list_unit	1	km		
	samplingVerbatim	Literal description of the sampling, and notes	string	1			
Variables	**outputVariables**	names as found in the dataset. Could be either	var_out / string	n by	density|detection probabiliity		

		the results of the analysis or the kind of variable collected on ground		|			
	stratificationVariables	names as found in the dataset	string	n by |	habitat		
	stratumVariables	names as found in the dataset	string	n by |	habitat		
	clusterVariables	names as found in the dataset	string	n by |	clusterName		
	sizeVariables	names as found in the dataset	string	n by |			
	covariates	names as found in the dataset	string	n by |			
Analysis	**analysisFamily**	Name of the statistical family used for the analysis.	ana_fam / string	1	Species distribution model		
	analysisName	Name of the statistic used for the analysis.	string	1	MaxEnt		
	inference	type of inference (applied either for an analysis or for raw data sampling)	inference	1	Model-based inference - Frequentist		

		incertitudeInformationType	how is recorded the information about uncertainty	ana_incert	n by |	statistical – distribution		
		validationMethod	method of validation of the analysis	ana_val		cross validation		
		dataSourceID	name or id of data sources used to perform the analysis	string/doi	n by |	doi		
		scriptID	name or id of the script	string/doi	1			
		sofware	Software on which was written and run the script used for the analysis	string	n by|	R (Distance)		
		analysisVerbatim	free description of the analysis					

Events

TABLE	CONCEPT	VARIABLE	DEFINITION	TYPE	N	EXAMPLE	DWC	ISO
EVENT	Identification	datasetID	The identifier of the dataset as found in the metadata	string/doi	1		http://rs.tdwg.org/dwc/terms/datasetID	

		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.	string/doi	1	IT001A1000004	http://rs.tdwg.org/dwc/terms/Event	
		parentEventID	An identifier for the broader Event that groups this and potentially other Event	string/doi	1	IT001A1000001	http://rs.tdwg.org/dwc/terms/parentEventID	
		eventName	The common name of the Event	string	1	Manoria		
		originalID		string/doi	1			
	Origin	recordedBy	A person, group, or organization responsible for recording the original Event.	string	1	Mario Rossi	http://rs.tdwg.org/dwc/iri/recordedBy	
	Location	locality	The specific description of the	string	1	San Luigi	http://rs.tdwg.org/dwc/terms/locality	

			place. This term may contain information modified from the original to correct perceived errors or standardize the description.					
		locationType	Type of delineation of the geographic information	loc_typ	1	polygon		
		xyType	Nature of the X,Y values (real or estimated) to be reported	loc_xytyp	1	estimated		
		x	Geographic Longitude (in decimal degree, using the spatial reference system in "Reference system")	numeric	1	7.210178	http://rs.tdwg.org/dwc/terms/decimalLongitude	
		y	Geographic Latitude (in decimal degree, using the spatial reference system in "Reference system")	numeric	1	45.520042	http://rs.tdwg.org/dwc/terms/decimalLatitude	

	xyUncertainty	The horizontal distance class (in meters) from the given X and Y describing the smallest circle containing the whole of the Location. If the uncertainty is unknown or cannot be estimated select option NA (not applicable)	loc_uncertainty	1	100m-1km	http://rs.tdwg.org/dwc/terms/coordinateUncertaintyInMeters	
	footprintWKT	A Well-known Text (WKT) representation of the shape that defines the location	string	1	POLYGON((5 5,28 7, 44 14, 47 35,40 40,20 30,5 5))	http://rs.tdwg.org/dwc/terms/footprintWKT	
	locationID	Identifier of the line in the provided shapefile or Code corresponding to the cell of the reference grid to which the data refers or to the NUT	string	1	ID3526 *or* 10kmE283N286 *or* ITI18 - Arezzo	http://rs.tdwg.org/dwc/terms/locationID	

			identification(see locationAccordingTo)					
		locationAccordingTo	The name for the shape file provided (assigned by data provider) or to the reference grid or reference NUT	string	1	CR12082.shp *or* Grid_ETRS89_LAEA_10K_FR *or* NUTS3-2016	http://rs.tdwg.org/dwc/terms/locationAccordingTo	
		referenceSystem	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the X,Y coordinates or polygon are given	string	1	EPSG: 23030	http://rs.tdwg.org/dwc/terms/geodeticDatum	
		country	Country where data were collected	loc_country	1	IT	http://rs.tdwg.org/dwc/terms/countryCode	
		areaType	Type of the area which data refer.	loc_areatyp	1	hunting ground		
		areaSize	Size of this area	numeric	1	3250		
		areaSizeUnit	Unit of the area	unit_area		hectare		
	Time	timeLevel	Level of temporal	time_level	1	day		

			aggregation of data.				
		dayBeginDate	Day of observation or starting day of data collection (if interval).	integer	1	14	
		monthBeginDate	Month of observation or starting month of data collection (if interval).	integer	1	10	
		yearBeginDate	Year of observation or starting year of data collection (if interval).	integer	1	2017	
		dayEndDate	Ending day of data collection (if interval).	integer	1	12	
		monthEndDate	Ending month of data collection (if interval).	integer	1	11	
		YearEndDate	Ending year of data collection. (if interval).	integer	1	2019	

		dataTime	Original representation of the date-time or interval during which an Event occurred. It can be a precise date-time or a range (e.g., hunting season)	string	1	2017/2019	http://rs.tdwg.org/dwc/terms/eventDate	
	Notes	notes	Notes to the record	string		Drive hunt locally called monteria		

Occurrence

TABLE	CONCEPT	VARIABLE	DEFINITION	TYPE	N	EXAMPLE	DWC	ISO
OCCURRENCE	Identification	datasetID	The identifier of the dataset as found in the metadata.	string/doi	1		http://rs.tdwg.org/dwc/terms/datasetID	
		occurrenceID		string/doi	1			
		parentEventID		string/doi	1			
		parentOccurrenceID		string/doi	1			
	Origin	originalID		string/doi	1			
		recordedBy		string	1	Maro Rossi		
	Location	locality	The specific description of the place. Less specific geographic information can be provided in other geographic terms (higherGeography, continent, country, stateProvince, county, municipality,	string	1	San Luigi	http://rs.tdwg.org/dwc/terms/locality	

		waterBody, island, islandGroup). This term may contain information modified from the original to correct perceived errors or standardize the description.					
	locationType	Type of delineation of the geographic information	loc_typ	1	polygon		
	xyType	Nature of the X,Y values (real or estimated) to be reported	loc_xytyp	1	estimated		
	x	Geographic Longitude (in decimal degree, using the spatial reference system in "Reference system")	numeric	1	7.210178	http://rs.tdwg.org/dwc/terms/decimalLongitude	
	y	Geographic Latitude (in decimal degree, using the spatial reference system in	numeric	1	45.520042	http://rs.tdwg.org/dwc/terms/decimalLatitude	

			"Reference system")				
		xyUncertainty	The horizontal distance class (in meters) from the given X and Y describing the smallest circle containing the whole of the Location. If the uncertainty is unknown or cannot be estimated select option NA (not applicable)	loc_uncert	1	100m-1km	http://rs.tdwg.org/dwc/terms/coordinateUncertaintyInMeters
		footprintWKT	A Well-known Text (WKT) representation of the shape that defines the location	string	1	POLYGON((5 5,28 7, 44 14, 47 35,40 40,20 30,5 5))	http://rs.tdwg.org/dwc/terms/footprintWKT
		locationID	Identifier of the line in the provided shapefile or Code corresponding to the cell of the reference grid to	string	1	ID3526 *or* 10kmE283N286 *or* ITI18 - Arezzo	http://rs.tdwg.org/dwc/terms/locationID

			which the data refers or to the NUT identification(see locationAccordingTo)					
		locationAccordingTo	The name for the shape file provided (assigned by data provider) or to the reference grid or reference NUT	string	1	CR12082.shp *or* Grid_ETRS89_LAEA_10K_FR *or* NUTS3-2016	http://rs.tdwg.org/dwc/terms/locationAccordingTo	
		referenceSystem	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the X,Y coordinates or polygon are given	string	1	EPSG: 23030	http://rs.tdwg.org/dwc/terms/geodeticDatum	
		country	Country where data were collected	loc_country	1	Italy	http://rs.tdwg.org/dwc/terms/countryCode	
		areaType	Type of the area which data refer.	loc_areatyp	1	hunting ground		
		areaSize	Size of this area.	numeric	1	3250		
		areaSizeUnit	Unit of the area	unit_area		hectare		

Time	timeLevel	Level of temporal aggregation of data.	time_level	1	day		
	dayBeginDate	Day of observation or starting day of data collection (if interval).	integer	1	14		
	monthBeginDate	Month of observation or starting month of data collection (if interval).	integer	1	10		
	yearBeginDate	Year of observation or starting year of data collection (if interval).	integer	1	2017		
	dayEndDate	Ending day of data collection (if interval).	integer	1	12		
	monthEndDate	Ending month of data collection (if interval).	integer	1	11		
	YearEndDate	Ending year of data collection. (if interval).	integer	1	2019		

		dataTime	Original representation of the date-time or interval during which an Event occurred. It can be a precise date-time or a range (e.g., hunting season)	string	1	2017/2019	http://rs.tdwg.org/dwc/terms/eventDate	
	Biology	basisOfRecord	The specific nature of the data record.	bio_basis	1	Human observation	http://rs.tdwg.org/dwc/terms/basisOfRecord	
		species	The full scientific name as used by the consortium.	bio_species	1	Capreolus capreolus	http://rs.tdwg.org/dwc/terms/scientificName	
		recordedStatus	A statement about the presence or absence of a Taxon at a Location.	bio_status	1	present	http://rs.tdwg.org/dwc/iri/occurrenceStatus	
		recordType	Type of sightings recorded in the database.	bio_recordtyp	1	alive	http://rs.tdwg.org/dwc/iri/typeStatus	
		sex	The sex of the individual(s) represented in the Occurrence.	bio_sex	1	male	http://rs.tdwg.org/dwc/terms/sex	

		lifeStage	The age class of the individual(s) at the time the Occurrence was recorded	bio_lifeStage	1	juvenile	http://rs.tdwg.org/dwc/terms/lifeStage	
		individualCount	The number of individuals represented present at the time of the Occurrence	integer	1	5	http://rs.tdwg.org/dwc/terms/individualCount	
		individualID	The identification of the individual	string	1	Or73		
	Notes	notes	Notes to the record	string	1		http://rs.tdwg.org/dwc/terms/taxonRemarks	

Measurement or Fact

TABLE	CONCEPT	VARIABLE	DEFINITION	TYPE	N	EXAMPLE	DWC	ISO
NEMOF	Identification	datasetID	The identifier of the dataset as found in the metadata.	string/doi	1		http://rs.tdwg.org/dwc/terms/datasetID	
		measurementID	An identifier for the MeasurementOrFact (information pertaining to measurements, facts, characteristics, or assertions).	string/doi	1		http://rs.tdwg.org/dwc/terms/measurementID	
		parentEventID	An identifier for the broader Event that groups this and potentially other Events.	string/doi	1		http://rs.tdwg.org/dwc/terms/eventID	
		parentOccurrenceID	An identifier for the broader Occurrence that groups this and potentially other Events.	string/doi	1		http://rs.tdwg.org/dwc/terms/occurrenceID	
		parentMeasurementID	An identifier for the broader neMoF that groups this and potentially other Events.	string/doi	1			
	Measurement or Fact	measurementType	The nature of the measurement, fact, characteristic, or assertion.	nemof_type	1		http://rs.tdwg.org/dwc/terms/measurementType	

		measurementValue	The value of the measurement, fact, characteristic, or assertion.	string/numeric	1		http://rs.tdwg.org/dwc/terms/measurementValue	
		measurementUnit	The units associated with the measurementValue.	string/unit_all	1	individual ; kilometer ; individual per square kilometer	http://rs.tdwg.org/dwc/terms/measurementUnit	
	Notes	notes	Comments or notes accompanying the MeasurementOrFact.	string	1		http://rs.tdwg.org/dwc/terms/measurementRemarks	

Appendix B Controlled vocabularies

Lists for metadata

mtd_fwtyp : Type of collection Framework

VALUE	DEFINITION
institutional	data collection is committed or realized by public/private entities for institutional purposes
citizen science	data are collected within a citizen science initiative
research	data are collected by a research team/institution within a targeted research project
hunters	data are collected by committees/associations of hunters within wildlife management activities
other	none of the above options can apply

mtd_role : role

VALUE	DEFINITION
author	creator of the dataset, but not owner or responsible for it
content provider	person contributing to collect data and/or to enter them into the database
owner	formal owner of the dataset or representative of the owner (signing the Data Sharing Agreement with EFSA)
point of contact	reference person collating data collected by other entities
principal investigator	principal investigator of the research which data are from
user	person not contributing to data collection/entering but using the data
other	none of the above options can apply

mtd_access : accessibility

VALUE	DEFINITION
open access	data are freely accessible and usable, e.g. can be downloaded from a public repository
public database on request	data are freely usable but not open access, so they must be requested to the database holder
agreement with the owner	data must be requested and may be used with the permission of the data owner
agreement with the owner and	data must be requested and may be used with the permission of the data owner and

partners	its partners
no access	data are collected but its dissemination is denied

mtd_datasettyp

VALUE	DEFINITION
raw data	data as they are collected on the field, after been cleaned up and verified
summarized data	secondary data, processed from raw or other summarized data with little or no statistics used (e.g. sum or number of records)
results of analysis	processed data obtained after a particular statistical procedure

mtd_upfreq : update frequency

VALUE	DEFINITION
annually	data are updated once per year, in a limited timeframe, e.g. at the end of the hunting season
as needed	data are updated only when needed, e.g. for reporting activity
biannually	data are updated once every two years
continuous	new records are inserted continuously as data are collected
irregular	data are updated occasionally, without any time scheduling
monthly	data are updates once per month
weekly	data are updated once per week
unknown	information on update frequency is not available to data provider

Lists for locations

loc_country : country codes : reference : iso3166-1 alpha2

VALUE	DEFINITION
AD	Andorra
AL	Albania
AM	Armenia
AT	Austria
AZ	Azerbaijan
BA	Bosnia Herzegovina
BE	Belgium
BG	Bulgaria
BY	Belarus
CH	Switzerland
CY	Cyprus
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GB	United Kingdom of Great Britain and Northern Ireland
GE	Georgia
GR	Greece
HR	Croatia
HU	Hungary
IE	Ireland
IT	Italy
LI	Lichtenstein

LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Moldova, Republic of
ME	Montenegro
MK	North Macedonia
MT	Malta
NL	Netherlands
NO	Norway
PL	Poland
PT	Portugal
RO	Romania
RS	Serbia
RU	Russia
SE	Sweden
SI	Slovenia
SK	Slovakia
TR	Turkey
UA	Ukraine
XZ	Kosovo[17]

loc_geoscale : geographic scale (of coverage)

VALUE	DEFINITION
municipality	data cover a whole municipality

[17] This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo declaration of independence.

province	data cover a whole province
county	data cover a whole county
district	data cover a whole administrative district
region	data cover a whole region
country	data cover a whole country
subregional	data cover a portion of a region that does not identify with a specific administrative unit
interregional	data cover two or more regions in a country
subnational	data cover a portion of a country that does not identify with a sum of regions
other	none of the above options can apply

loc_typ : type of delineation of the geographic information

VALUE	DEFINITION
coordinates	coordinates X and Y are provided, either in WKT format, using the XY columns or in external file
polygon	a spatial shape is provided, either in WKT format or in external file
line	a spatial line is provided, either in WKT format or in external file
EEA grid	spatial information is a cell code of the EEA grid
UTM grid	spatial information is a cell code of the UTM grid
NUTS	spatial information is a European NUT code

loc_xytyp :what it the nature of the XY values provided

VALUE	DEFINITION
real	exact coordinates measured
estimated	estimation of the coordinates (for instance, centroid of a municipality)

loc_areatyp : description of the type of area described

VALUE	DEFINITION
administrative unit	administrative area, similar a equal to a NUT level

hunting ground	area where exactly the hunting took place happened
management unit	global area managed
study area	area of the study performed
biogeographical unit	biologically relevant area

loc_uncert : description of the type of area described

VALUE	DEFINITION
0-10m	the given X,Y coordinates have an uncertainty lower than 10m
10-50m	the given X,Y coordinates have an uncertainty between 10m and 50m
50-100m	the given X,Y coordinates have an uncertainty between 50m and 100m
100m-1km	the given X,Y coordinates have an uncertainty between 100m and 1000m
>1km	the given X,Y coordinates have an uncertainty higher than 1000m
unknown	the uncertainty is unknown

Lists for time

time_scale : temporal resolution

VALUE	DEFINITION
year	entries in the dataset refer to annual events, e.g. annual census
month	entries in the dataset refer to monthly events, e.g. hunting bag summarized by month
precise date	entries in the dataset refer to single observations or daily events, e.g. individual sighting
hunting season	entries in the dataset refer to single hunting seasons
other	none of the above options can apply

time_level : level of temporal aggregation of data

VALUE	DEFINITION
day	
month	

year	
interval	begin and end date of the interval has to be provided
hunting season	
pre-birth season	
post-birth season	

Lists for biological records

bio_species : species list of Enetwild

VALUE	DEFINITION (EN)	DEFINITION (FR)
Cervus elaphus	Red deer	Cerf élaphe
Capreolus capreolus	Roe deer	Chevreuil
Dama dama	Fallow deer	Daim
Capra ibex	Alpine	Bouquetin
Capra hispanica	Iberian ibexes	Bouquetin ibérique
Ovis aries	Moufloun	Mouflon
Rupicapra rupicapra	Chamois	Chamois
Rupicapra pyrenaica	Southern chamois	Isard
Alces alces	Moose	Elan
Rangifer tarandus	Reindeer	Renne
Bison bonasus	European Bison	Bison d'Europe
Odocoileus virginianus	White-tailed deer	Cerf de Virginie
Cervus elaphus subspp	Elk	Wapiti
Hydropotes inermis	Chinese water deer	Hydropote
Muntiacus reevesi	Muntjac deer	Muntjac de Reeves
Ovibos moschatus	Musk ox	Boeuf musqué
Ammotragus lervia	Barbary sheep	Mouflon à manchettes
Canis lupus	Wolf	Loup gris
Lynx lynx	Eurasian lynx	Lynx boréal
Meles meles	European badger	Blaireau européen
Procyon lotor	Raccoon	Raton laveur
Nyctereutes procyonoides	Raccoon dog	Chien viverrin
Vulpes vulpes	Red fox	Renard roux
Ursus arctos	Brown bear	Ours brun
Canis aureus	Golden jackal	Chacal doré

bio_status :

VALUE	DEFINITION
present	
absent	
present-stational	present in the whole area
present-inventorial	present somewhere in the area

bio_recordtyp :

VALUE	DEFINITION
alive	the target of observation was observed alive
dead	the target of observation was found dead
indirect sign	indices of the species was observed
other	must be define in notes

bio_sex :

VALUE	DEFINITION
male	the sex of the observed individual(s) is male
female	the sex of the observed individual(s) is female
indetermined	the sex of the observed individual(s) is unknown

bio_lifeStage :

VALUE	DEFINITION
adult	the stage of life of the observed individual(s)is adult/mature
juvenile	the stage of life of the observed individual(s)is juvenile/immature

this list can be extended to fit appropriate stage of live according to the biology of the species.

bio_basisOfRecord : The specific nature of the data record

VALUE	DEFINITION
human observation	the observation was recorded by a human
machine observation	the observation was recorded by an automatic machine
statistical estimation	the record is the result of a statistical estimation

Lists for statistics

samp_process: sampling process: Probability sampling design (PSD) vs (Non probability sampling NPS)

VALUE	DEFINITION
PSD-simple random sampling	samples are randomly selected from the sampling frame with no other consideration
PSD-stratified random sampling	samples are randomly selected from strata, themselves splitting the sampling frame into homogeneous groups
PSD-cluster sampling	samples are randomly selected from clusters. Clusters are mutually homogeneous yet internally heterogeneous groups in the sampling frame
PSD-probability proportional to size sampling	the probability of inclusion of a sample depend on its size value
PSD-systematic sampling	the selection of samples from the sampling frame depends on a fix rule
NPS-self selection	individual answers themselves to a questionnaire based on their own attraction to it.
NPS-judgement sampling	the selection of samples depends on someone estimation/knowledge
NPS-convenience sampling	the selection of samples depends on what is feasible
NPS-quota sampling	the sampled is set up to match pre-defined quota per category
NPS-snowball sampling	
NPS-census	it is estimated that all of the individuals are observed, or all of the samples of a sampling frame are measured
non relevant	the concept of sampling process does not apply to the dataset
unknown	the sampling process is not known

samp_availability : Availability of the sampling frame

VALUE	DEFINITION
open online	the sampling frame can be found online and is accessible and usable

open on request	the sampling frame can be furnished on demand
restricted	the sampling frame is available but its accessibility is restricted (for instance due to EU general data protection regulation)
non available	the sampling frame cannot be furnished
unknown	the sampling frame is unknown

samp_prot : field protocol (apply to enetwild themes)

VALUE	DEFINITION
counting	a number of individuals/contacts detected is recorded on field
inventory	the list of present species of a community on a location is established
presence-only	the presence of a species is recorded but there are no information but absence of data does not inform on the absence of the species
presence-absence	both presence and absence of the species are recorded
individual identification	individuals recorded are uniquely identified (for instance in a CMR protocol, a telemetry protocol, or while ringing birds)
expert knowledge	the record corresponds to the knowledge of someone, based on its own experience and direct and indirect observations.
questionnaire	the record come from the answer to a questionnaire

var_out : (apply to enetwild themes): non restrictive

VALUE	DEFINITION
density	number of individual (or other unit) per surface unit
population size	total number of individual
relative abundance	index of abundance, used for trends
hunting bag	size of hunting bag

size	
presence probability	the analysis provides a probability of presence of the species, which is not a proof of presence, nor a proof of absence
habitat suitability	the analysis provides a suitability value based on habitat characteristics and habitat selection behaviour of the species
detection probability	the analysis provides (usually in combination with other values) a detection probability
occurrence	the dataset correspond to simple occurrence data
count	the dataset correspond to records of count of individuals/contacts or other units

ana_fam : analysis family (apply to enetwild themes): non restrictive

VALUE	DEFINITION
Capture Mark Recapture	
Distance sampling	
Species distribution model	(MaxEnt, Support vector machine, Random forest, Bioclim…)
Habitat suitability	
Regression	
Survey sampling	ex : Horvizt Thompson estimator, hunting bag survey
Camera trapping	
Expert classification	
Multivariate similarity surface	

inference :

VALUE	DEFINITION
Design-based inference	
Design-based inference – Model assisted	
Model-based inference – Frequentist	
Model-based inference – Bayesian	

Hybrid model/design-based inference	
non relevant	
unknown	

ana_uncert : uncertainty Information Type

VALUE	DEFINITION
none	
statistical – measure	
statistical – distribution	
statistical – quantile	the distribution is displayed through quantiles
minimal-maximal values	
ad-hoc	
ad-hoc –advanced statistics	
non relevant	
unknown	

ana_val : Validation method

VALUE	DEFINITION
none	
Cross validation	
Mean square error	
Pearson's correlation	
Area Under the Curve	
Comparison with sightings data	
ad-hoc	
ad-hoc –advanced statistics	
non relevant	
unknown	

ana_quant

VALUE	DEFINITION
p-quantile	none
q-quantile	none

ana_distr

VALUE	DEFINITION
uniform	
normal	gaussian
half-normal	
log-normal	
binomial	
negative binomial	
poisson	
quasi-poisson	
exponential power	
negative exponential	
hypergeometric	
beta	
gamma	
mixture	
non parametric smoother	
hazard-rate	
triangular	used sometimes in distance sampling
other	
other advanced	

List for measurements

nemof_type (this list is expendable according to needs)

MAIN LINK	GROUP	VALUE	ACCEPTED TYPE OF VALUES	ACCEPTED UNIT	DEFINITION
event	context	habitat	string	none	
		weather	string	none	
		surface	numeric	unit_surface	
		pig husbandry presence	yes/no/unknown	none	
		presence of quotas	yes/no/unknown	none	
	effort generic	effort distance	numeric	unit_length	
		effort surface	numeric	unit_surface	
		effort visit	numeric	string	
		effort time	numeric	unit_time	
		total effort	numeric	string	
	effort : specific hunting	effort in dogs	numeric	string	(best : individual)
		effort in hunters	numeric	string	(best : individual)
		effort in baiters	numeric	string	(best : individual)
	sampling	sample weight	numeric	none	
		sample size	numeric	string	
		planned sample size	numeric	string	
occurrence	technical information	distance	numeric	unit_length	
		perpendicular distance	numeric	unit_length	

		angle	numeric	unit_angle	
		weight	numeric	unit_mass	
		cause of death	string	none	
	type of estimation	density	numeric	unit_abundance	
		relative abundance	numeric	none	
		population size	numeric	string	(best : individual)
		hunting bag size	numeric	string	(best : individual)
		detection probability	numeric	string	
		reproductive rate	numeric	string	
		survival	numeric	string	
		sex-ratio	numeric	string	
neMoF 1	linked to type of estimation	interval	numeric|numeric	see estimation	
		distribution	ana_distr	none	
		standard deviation	numeric	see estimation	
		variance	numeric	see estimation[2]	
		median	numeric	see estimation	
		quantile	ana_quant	none	
neMoF 2	linked to interval	confidence level	numeric	see estimation	
	linked to distribution	n	numeric	none	
		p	numeric	none	
		λ	numeric	none	
		M	numeric	none	
		N	numeric	none	
		c	numeric	none	

	p	numeric	none
	μ	numeric	none
	σ	numeric	none
	α	numeric	none
	β	numeric	none
	a	numeric	none
	b	numeric	none
	θ	numeric	none
	k	numeric	none
linked to quantile	x_0.025	numeric	see estimation
	x_0.12	numeric	see estimation
	x_0.215	numeric	see estimation
	x_0.31	numeric	see estimation
	x_0.405	numeric	see estimation
	x_0.5	numeric	see estimation
	x_0.595	numeric	see estimation
	x_0.69	numeric	see estimation
	x_0.785	numeric	see estimation
	x_0.88	numeric	see estimation
	x_0.975	numeric	see estimation

Lists of units

the units should be based as much as possible on the full name of international unit system:
https://physics.nist.gov/cuu/pdf/sp811.pdf
For the database purpose, they must be spelt in English, minuscule, singular and using the full name. there are no space between the prefix and the unit (kilometer not kilo meter). Division of unit is indicated by " per ".
Exposant are indicated in full name before the unit: square, cubic.
For publication in direction to readers, it is however recommended to translate them into their official symbol

unit_all : (this list is expendable according to needs)

GROUP	VALUE	DEFINITION
Surface	square meter	
	square kilometer	
	hectare	
Length	meter	
	kilometer	
Time	second	
	minute	
	hour	
	day	
	month	
	year	
angular	degree	
	radian	
mass	gram	
	kilogram	
abundance	individual	
	individual per kilometer	
	individual per square kilometer	
	group	
	contact	

	couple
Temperature	degree Celsius

Appendix C Implementation of the wildlife monitoring standard

Available at this link