Formatting records for the NBN Atlas: An introduction to Darwin Core

SOPHIA RATCLIFFE

NBN Trust Technical & Data Partner Support Officer

REUBEN ROBERTS

NBN Systems Developer

NBN Conference 2018 Knowledge Transfer Session

Session Aims

- What is Darwin Core (DwC)?
- How is DwC used in the NBN Atlas
- Can we (NBN) use DwC better?
- What can we contribute to DwC?
- (Improvements to NBN Atlas pages)

What is DwC?

Darwin Core is the data standard for publishing and integrating biodiversity information

Library of terms aimed at to providing common naming conventions and data structure

Primarily based on **taxa** and their **occurrence**

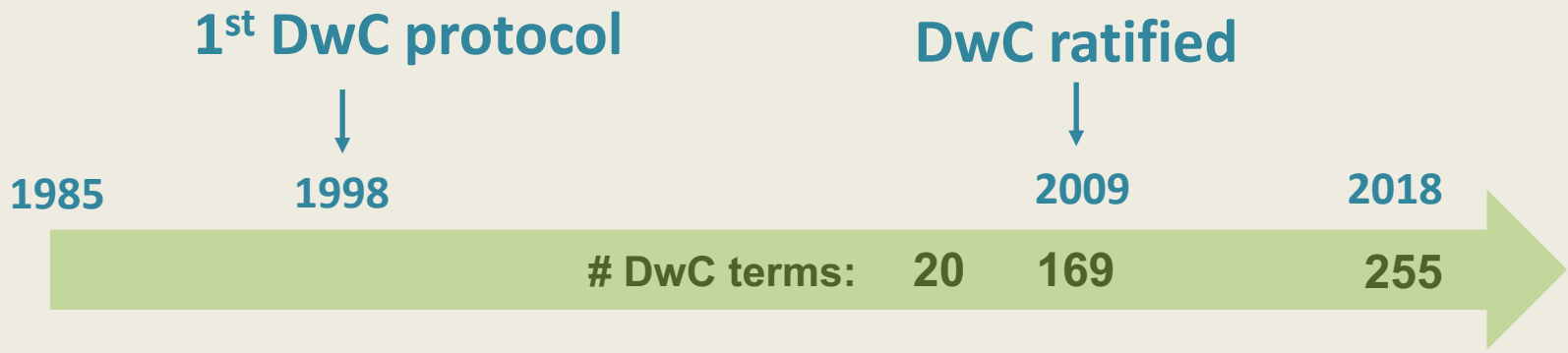
Adapted from: http://rs.tdwg.org/dwc/

Wieczorek et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715

Taxonomic Databases Working Group

Biodiversity Information Standards

TDWG



Who (else) uses DwC?

MOL
MAP OF LIFE

GBIF
Global Biodiversity Information Facility

VertNet

iDigBio
Integrated Digitized Biocollections

DASSH
The archive for marine species and habitats data

DwC reference guide

6

http://rs.tdwg.org/dwc/terms/

TDWG Home Terms Guides Namespace policy

Darwin Core quick reference guide

This page provides a list of all currently recommended terms of the Darwin Core standard. Categories such as `Occurrence` or `Event` correspond to Darwin Core classes which group other terms. Convenient files of these terms and their full history can be found in the Darwin Core repository.

Record-level

type | modified | language | license | rightsHolder | accessRights | bibliographicCitation | references | institutionID | collectionID | datasetID | institutionCode | collectionCode | datasetName | ownerInstitutionCode | basisOfRecord | informationWithheld | dataGeneralizations | dynamicProperties

type	Property
Identifier	http://purl.org/dc/terms/type
Definition	The nature or genre of the resource.
Comments	Must be populated with a value from the DCMI type vocabulary (http://dublincore.org/documents/2010/10/11/dcmi-type-vocabulary/).
Examples	`StillImage` , `MovingImage` , `Sound` , `PhysicalObject` , `Event` , `Text`

Record-level
Occurrence
Organism
MaterialSample
Event
Location
GeologicalContext
Identification
Taxon
MeasurementOrFact
ResourceRelationship
UseWithIRI

LivingSpecimen
PreservedSpecimen
FossilSpecimen
HumanObservation
MachineObservation

DwC classes and terms

Record-level terms	Institution, collection, nature of record, licence, rightsholder
Occurrence	Occurrence ID, recorder, individual count, quantity (and quantity type), sex, life stage, behaviour, status (presence/absence)
Organism	Organism scope (colony, nest, clump), organism remarks
Event	Date, sampling protocol and methods, field notes
Location	Latitude and longitude coordinates, geodetic datum, location name and remarks
Identification	Verification status, identifier
Taxon	Taxon ID (UKSI taxon version key), scientific name, vernacular name

DwC term example

occurrenceID	Property
Identifier	http://rs.tdwg.org/dwc/terms/occurrenceID
Definition	An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique.
Comments	Recommended best practice is to use a persistent, globally unique identifier.
Examples	`http://arctos.database.museum/guid/MSB:Mamm:233627`, `000866d2-c177-4648-a200-ead4007051b9`, `urn:catalog:UWBM:Bird:89776`

http://rs.tdwg.org/dwc/terms/

DwC example

What does it mean in terms of the data?

HBRG Insects Dataset

Highland Biological Recording Group

occurrenceID	taxonID	locality	geodeticD	gridReferer	coordinat	recordedBy	eventDate	identificatic	occurren	individual	license	basisOfReco	rightsHold
CI00016600000S8Q	NBNSYS0000009604	Inverlair	OSGB	NN3379	1000	Unknown-to-	1918	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S8R	NBNSYS0000009844	Strathpeffer	OSGB	NH478575	100	Murdo Macd	1989-04-01	Accepted	present	1	CC-BY	HumanObse	Highland B
CI00016600000S8S	NBNSYS0000009836	Strathpeffer	OSGB	NH478575	100	Murdo Macd	1989-04-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S8T	NBNSYS0000009836	Spinningdale	OSGB	NH684897	100	Donald Beat	1999-04-01	Accepted	present	5	CC-BY	HumanObse	Highland B
CI00016600000S8U	NBNSYS0000009844	Scorguie	OSGB	NH646459	100	Jimmy McKe	1999-04-01	Accepted	present	1	CC-BY	HumanObse	Highland B
CI00016600000S8V	NBNSYS0000009838	Old Aberdeen	OSGB	NJ90	10000	Annie Lamb	1995-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S8W	NBNSYS0000009840	Ruthven, Huntly	OSGB	NJ518472	100	Annie Lamb	1995-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S8X	NBNSYS0000009840	Acre Lea	OSGB	NH845811	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S8Y	NBNSYS0000009840	Erracht, Loch Ey	OSGB	NH843802	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S8Z	NBNSYS0000009840	Geanies House	OSGB	NH895793	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S90	NBNSYS0000009840	Knocknahar, Inve	OSGB	NH855816	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S91	NBNSYS0000009840	near Cadboll	OSGB	NH871779	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S92	NBNSYS0000009840	near Easter Rari	OSGB	NH847747	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S93	NBNSYS0000009840	near Inver	OSGB	NH878828	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S94	NBNSYS0000009840	Nr Hill of Fearn	OSGB	NH832785	100	Bob Swann	1999-05-01	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S95	NBNSYS0000009840	Ardersier	OSGB	NH777561	100	Jimmy McKe	1999-05-01	Accepted	present	1	CC-BY	HumanObse	Highland B
CI00016600000S96	NBNSYS0000009842	Achnahaird mac	OSGB	NC015138	100	Murdo Macd	1999-05-01	Accepted	present	1	CC-BY	HumanObse	Highland B
CI00016600000S97	NBNSYS0000009836	Altandubh	OSGB	NB980130	100	Murdo Macd	2000-05-01	Accepted	present	1	CC-BY	HumanObse	Highland B
CI00016600000S98	NBNSYS0000009842	Culag, Carr Brae	OSGB	NG886253	100	Brian Neath	2000-05-01/	Accepted	present		CC-BY	HumanObse	Highland B
CI00016600000S99	NBNSYS0000009852	Seagreen Restr	OSGB	NG750280	100	Brian Neath	01/05/2001	Accepted	present	1	CC-BY	HumanObse	Highland B

DwC Extensions

- Simple multimedia
- Literature references
- Minimum Information about any (x) Sequence (MIxS)

Who manages DwC?

Darwin Core Maintenance Group

https://www.tdwg.org/standards/dwc/maintenance/

- Issues submitted to a Github site:

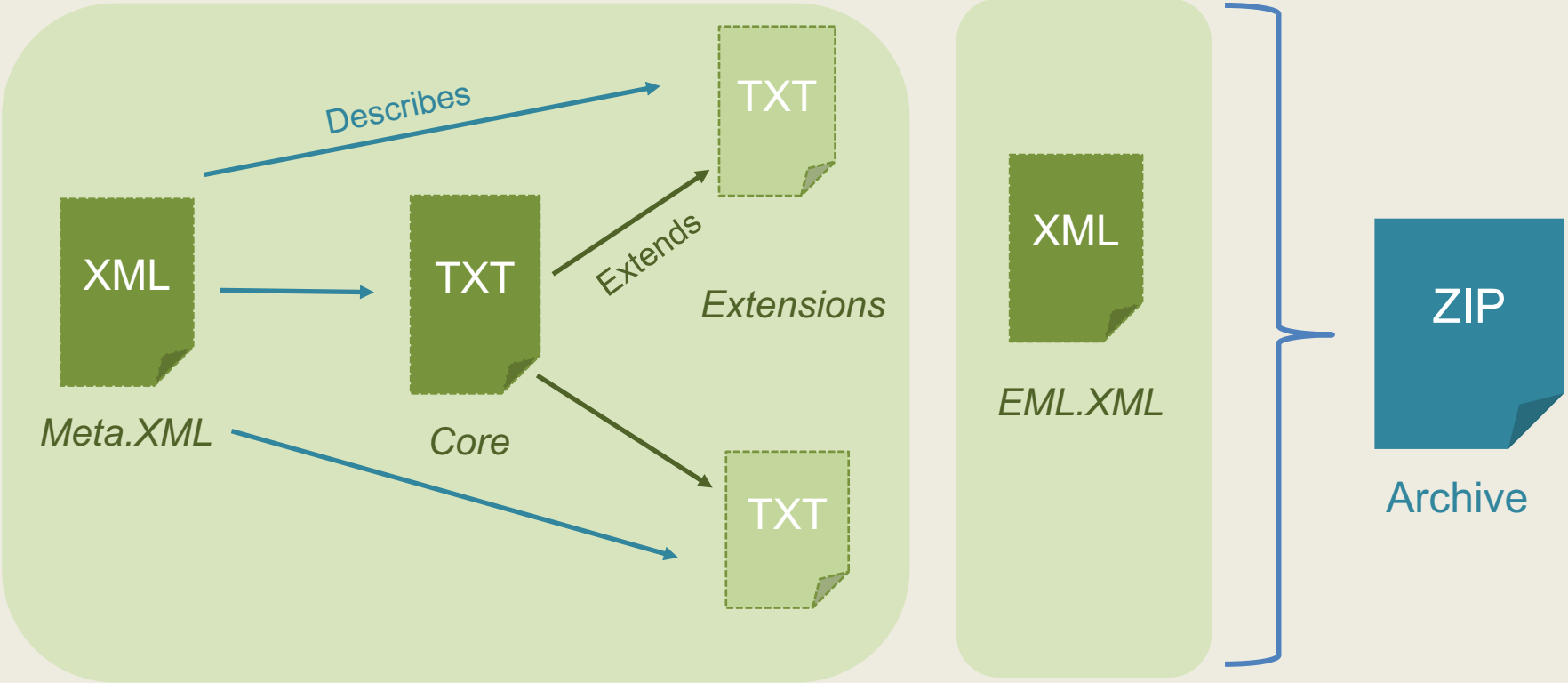
  https://github.com/tdwg/dwc/issues

- 30-day public review

- review by TDWG's Technical Architecture Group

DwC Archives

Sharing data



http://tools.gbif.org/dwca-assistant/

DwC on the NBN Atlas

- Taxon information (species dictionary)
    - updated 6-12 monthly
- Occurrence records
    - monthly processing run (1st weekend of each month)

Species dictionary

UK Species Inventory Access DB (NHM, London)

Taxon identifier
Scientific names
Vernacular names
Rank
Status (accepted/synonym)

Establishment means (native/non-native)
Establishment status
Realm (terrestrial, marine, freshwater)

Darwin Core TAXON

Occurrence records

Accepted formats

- DwCA (iRecord, RBGE)
- NBN Atlas formatted spreadsheets
- NBN Exchange format (Recorder 6, Marine Recorder)
- Unformatted spreadsheets

Occurrence records terms

- Core
- Desirable
- Non-DwC
- Other

Core terms

- occurrenceID
- basisOfRecord
- license
- rightsholder
- institutionCode
- occurrenceStatus (present / absent)
- identificationVerificationStatus

basisOfRecord

basisOfRecord	Property
Identifier	http://rs.tdwg.org/dwc/terms/basisOfRecord
Definition	The specific nature of the data record.
Comments	Recommended best practice is to use the standard label of one of the Darwin Core classes.
Examples	`PreservedSpecimen`, `FossilSpecimen`, `LivingSpecimen`, `MaterialSample`, `Event`, `HumanObservation`, `MachineObservation`, `Taxon`, `Occurrence`

identificationVerificationStatus

identificationVerificationStatus	Property
Identifier	http://rs.tdwg.org/dwc/terms/identificationVerificationStatus
Definition	A categorical indicator of the extent to which the taxonomic identification has been verified to be correct.
Comments	Recommended best practice is to use a controlled vocabulary such as that used in HISPID and ABCD.
Examples	`0` ("unverified" in HISPID/ABCD).

- Accepted
- Accepted - considered correct
- Accepted - correct
- Unconfirmed
- Unconfirmed - plausible
- Unconfirmed - not reviewed

Core terms cont.

- taxonID or scientificName or vernacularName
- eventDate
- gridReference / decimalLatitude & decimalLongitude
- geodeticDatum
- coordinateUncertaintyInMeters
- locality

- recordedBy
- identifiedBy

eventDate

- eventDate (YYYY-MM-DD) (ISO 8601)
  - 1998-03-28
  - 1998-03-28/05-31
  - 1998-03
  - 1998-03/05
  - 1998
  - 1998/2002
- day, month, year (single fields)
  - preferred method for single day events and partial dates (?)

eventDate cont.

- verbatimEventDate
  - “spring 1998”
- datePrecision (non-DwC)
- endDate (non-DwC)
  - endDate day, month, year

Core terms cont.

- taxonID or scientificName or vernacularName
- eventDate
- gridReference / decimalLatitude & decimalLongitude
- geodeticDatum – default WGS84
- coordinateUncertaintyInMeters
- locality

- recordedBy
- identifiedBy

- datasetName

non-DwC terms

- verifier
- organismStatus (alive/dead)

Desirable terms

- individualCount
- organismQuantity
- organismQuantityType

- organismScope
- sex
- lifeStage

individualCount

- 3% records have individual count (~5m)
- 29,000 different values

Examples:

“1 Adult”, “Frequent”, “1 Male”, “#NAME?”, “0.25”,

“2 Adult Male; 1 Juvenile Female”, “Many”

organismQuantity

organismQuantity	Property
Identifier	http://rs.tdwg.org/dwc/terms/organismQuantity
Definition	A number or enumeration value for the quantity of organisms.
Comments	A dwc:organismQuantity must have a corresponding dwc:organismQuantityType.
Examples	`27` (organismQuantity) with `individuals` (organismQuantityType). `12.5` (organismQuantity) with `%biomass` (organismQuantityType). `r` (organismQuantity) with `BraunBlanquetScale` (organismQuantityType).

organismQuantity

- 540,000 records with organismQuantity
- 2,000 different values

Examples:

“Many”, “Several”, ”sev.”, ”Present”

“Occasional” or “O” (organismQuantityType: DAFOR)

”50” (organismQuantityType: % cover)

Desirable terms

- individualCount
- organismQuantity
- organismQuantityType

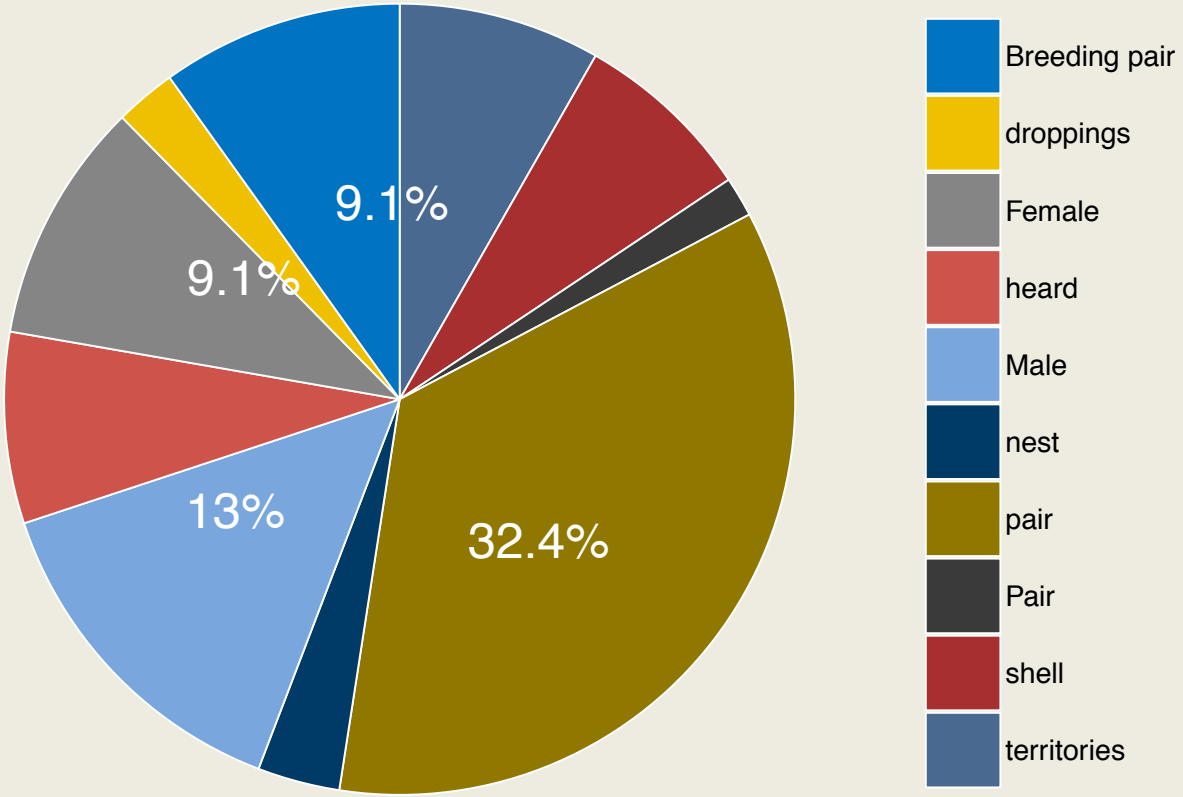
- organismScope
- sex
- lifeStage

organismScope

organismScope	Property
Identifier	http://rs.tdwg.org/dwc/terms/organismScope
Definition	A description of the kind of Organism instance. Can be used to indicate whether the Organism instance represents a discrete organism or if it represents a particular type of aggregation.
Comments	Recommended best practice is to use a controlled vocabulary. This term is not intended to be used to specify a type of taxon. To describe the kind of dwc:Organism using a URI object in RDF, use rdf:type (http://www.w3.org/1999/02/22-rdf-syntax-ns#type) instead.
Examples	`multicellular organism`, `virus`, `clone`, `pack`, `colony`

organismScope

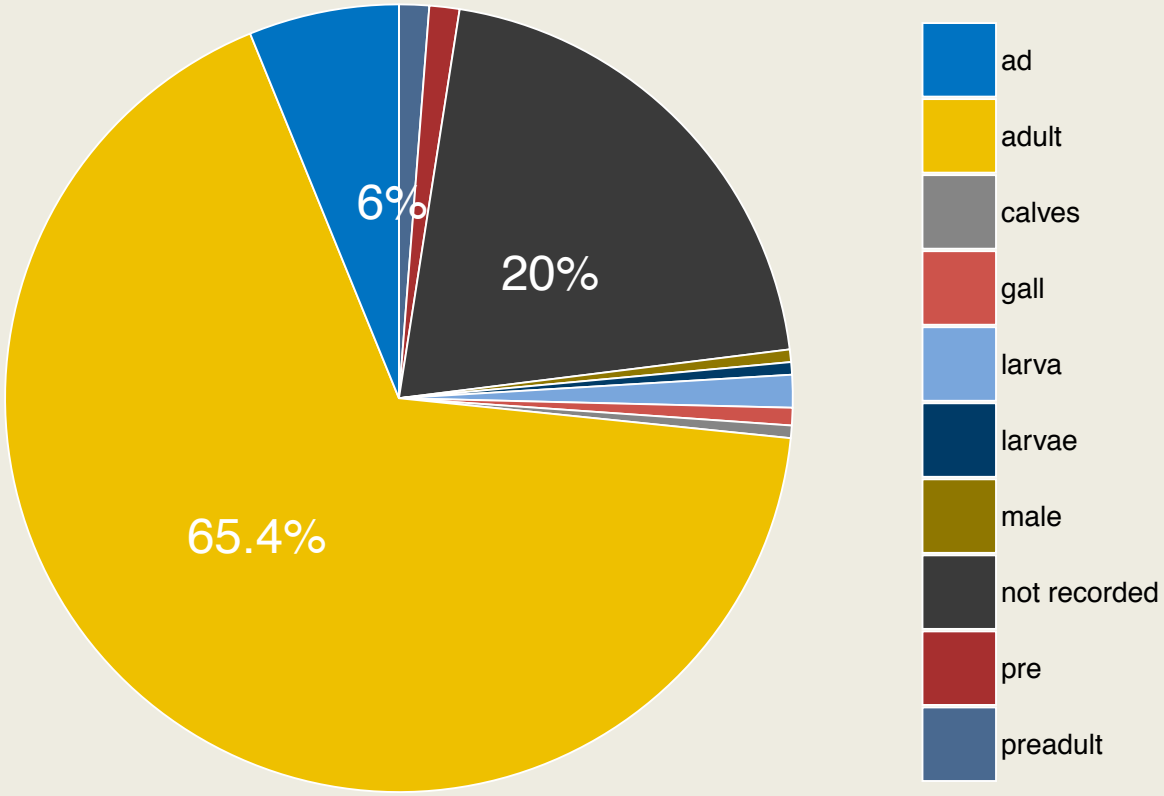
5,227 records with organismScope



Other examples: sett, spraint, tracks, nest, burrow, eggs

lifeStage

473,631 records with lifeStage



Other examples: immature, nymph, young, dead, chick

Comment (remarks) fields

- occurrenceRemarks
- organismRemarks
- eventRemarks
- locationRemarks
- identificationRemarks

Other terms

Event

- eventID
- samplingProtocol
- sampleSizeValue
- sampleSizeUnit
- samplingEffort

Other terms cont.

Record-level

- bibliographicCitation
- references
- informationWithheld
- dataGeneralizations
- dynamicProperties

dynamicProperties

A list of additional measurements, facts, characteristics, or assertions about the record. Meant to provide a mechanism for structured content.

dynamicProperties

A list of additional measurements, facts, characteristics, or assertions about the record. Meant to provide a mechanism for structured content.

National Dormouse Database (NDD)

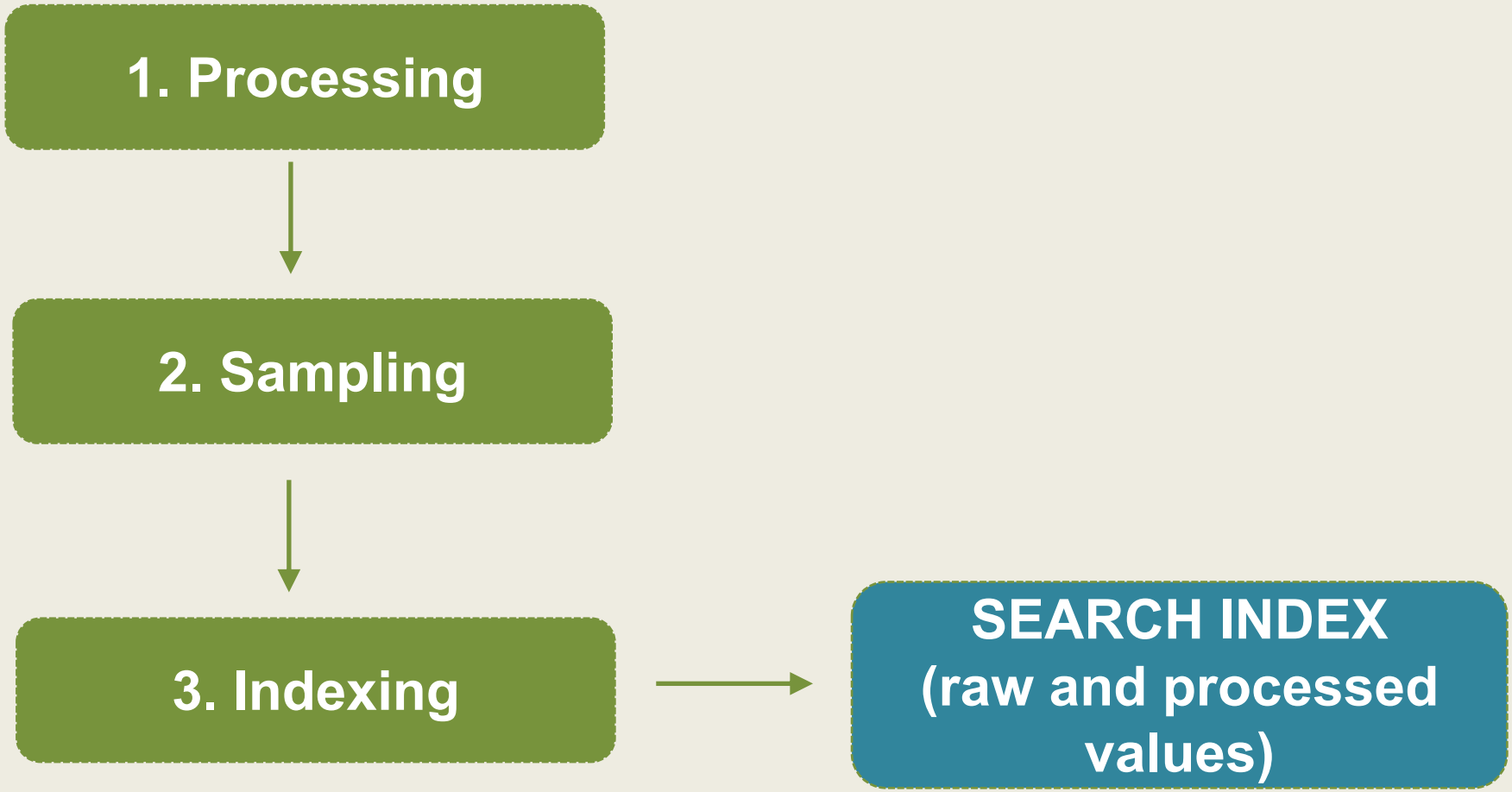
NDMPsite: Yes
RecordType: Live specimen
RecordTypeReliability: Good

people's trust for endangered species

Dynamic properties	{'NDMPsite': 'Yes', 'RecordType': 'Live specimen', 'RecordTypeReliability': 'Good'}

Data processing

Data processing



1. Processing

- Name matching routine
- OSGR <> Latitude/longitude coordinates
- Dates
- Species list membership
- Sensitive species

Sensitive species

Geospatial

Country	United Kingdom
State or territory	England
Latitude	50.97
Longitude	-1.99
Geodetic datum	EPSG:4326
Coordinate precision	1km
Coordinate uncertainty in metres	1000
Coordinates generalised	Location in England is already generalised to 0.01 degrees. Sensitive in ENGLAND, Name: England, Zone: COUNTRY [Sensitive, Natural England]
Information withheld	No geospatial information has been witheld during processing. If this is a sensitive record, this will be due to the geospatial information being sufficiently generalised for this taxon.
Georeference verification status	Accepted
Grid reference	SU0019

1. Processing cont.

- Data quality checks
  - recordHasIssues
  - recordIssues

Data quality checks

Test name	Result
Geodetic datum assumed WGS84	Warning
Country inferred from coordinates	Warning
Basis of record not supplied	Passed
Basis of record badly formed	Passed
Invalid collection date	Passed
Incomplete collection date	Passed
First of the month	Passed
Collector name unparseable	Passed
Data are generalised	Passed
Name not supplied	Passed
Name not recognised	Passed
Invalid scientific name	Passed

2. Sampling

- Boundaries
- Habitats
- Environmental layers

2. Sampling

Additional political boundaries information

Administrative	
Areas Of Outstanding Natural Beauty (England)	Cranborne Chase & West Wiltshire Downs
Local Environmental Records Centre Boundaries of the UK	Dorset Environmental Records Centre
Local Environmental Records Centre Boundaries of the UK	Dorset Environmental Records Centre
Landscape	
National Character Areas (England)	Dorset Downs and Cranborne Chase
Ordnance Survey Grids	
Ordnance Survey Grids - 100km	SU
Ordnance Survey Grids - 10km	SU01
Ordnance Survey Grids - 50km	SUSW
Political	
Countries with sea areas	England
UK countries	England
World base layer	United Kingdom
Vice Counties	
Watsonian Vice Counties	Dorset

3. Indexing

- SOLR
- Occurrence record fields:
  - https://records-ws.nbnatlas.org/index/fields
- Only possible to search / filter / facet indexed fields
- Can add fields to the index (e.g. lifeStage)

Worked examples

- Recorder 6 dataset (Highlands Biological Records Centre)
- CEDaR Northern Ireland Seal Survey

Help

- NBN Atlas documentation site

  https://docs.nbnatlas.org/share-species-occurrence-records-with-the-nbn-atlas/

- Darwin Core quick reference guide
  - https://dwc.tdwg.org/terms/
- Darwin Core Archive Assistant (GBIF)
  - http://tools.gbif.org/dwca-assistant/
- Darwin Core Archive Validator (GBIF)
  - https://tools.gbif.org/dwca-validator/

Can we use DwC better?

Controlled vocabularies:

- lifeStage
- sex
- organismScope

What can we contribute back?

- organismStatus
- verifier

Improvements

Improvements to the presentation of records in the NBN Atlas:

1. Occurrence records page
2. Data resource metadata page
3. Advanced records search

Improvements

- https://github.com/nbnuk/nbnatlas-issues