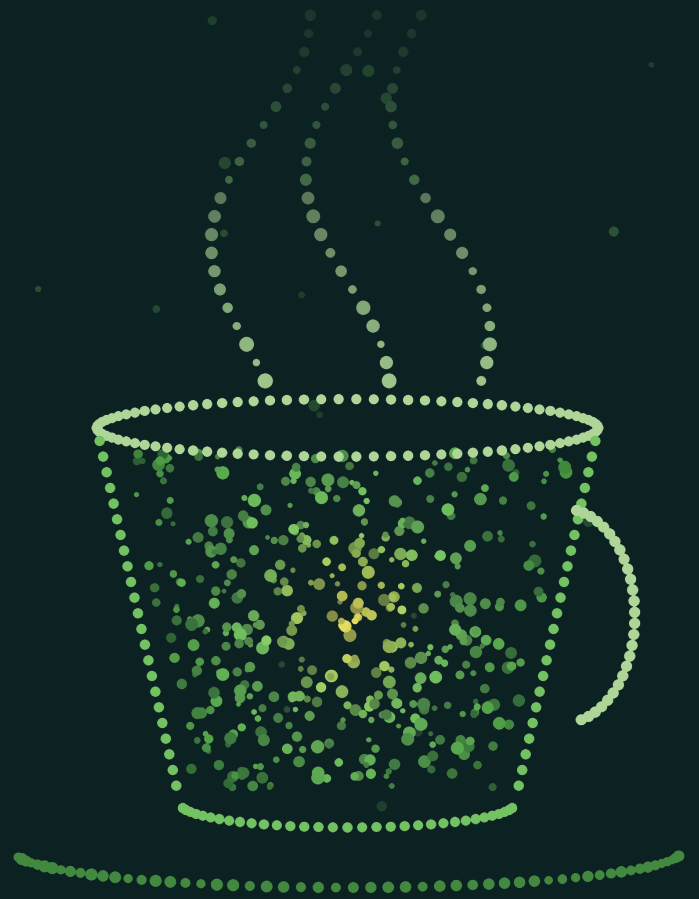
GBIF SWEDEN



SESSION 3 · 13 NOVEMBER 2026

STANDARDS, IDENTIFIERS & METADATA

WHAT GBIF DATA IS MADE OF



Lena Thöle

GBIF Global Biodiversity Information Facility

SBDI Swedish Biodiversity Data Infrastructure

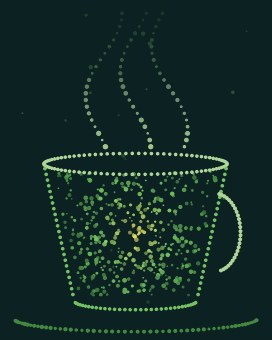
Swedish Museum of Natural History

Vetenskapsrådet Swedish Research Council

Quick Update:

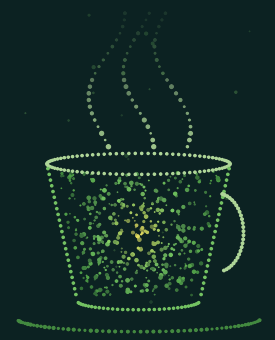
Data Fika #5 will move:
January 8 → January 15

Before we start



- **We are recording. T**he recording and the slides go on gbif.se/data-fika afterwards.
- **Microphones & cameras are off.** Type questions in the chat and just relax.
- **Questions in the chat at any time.** We take them as they come, and again at the end.
- **Please fill out the poll.**

The ten sessions – one journey



1–2

Get the data

What GBIF is, and how to pull a download you can cite

3–4

Understand it

Darwin Core, identifiers, metadata and the GBIF data model

5

Clean it

Coordinates, names, dates, duplicates, encoding

6

Judge it

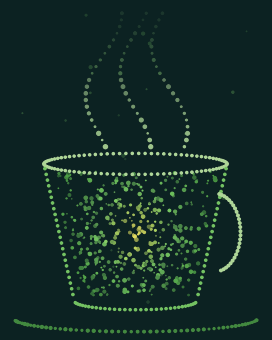
Cleaned is not the same as fit for your question

7–10

Use and share

Licences, FAIR & CARE, publishing, the whole workflow end to end

What does a row say?



Look at your ladybird download from session 2

CHECK THE COLUMN NAMES

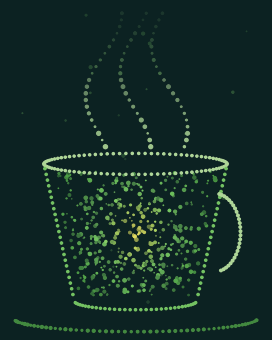
50 columns in the Simple CSV · 230 in the DwC-A

- 2 Dublin Core terms
- **33 Darwin Core terms**
- 15 added by GBIF

```
5896573334 · 38b4c89f-584c-41bb-bd8f-cd1def33e92f
· urn:lsid:artportalen.se:sighting:129617894 ·
Animalia · Arthropoda · Insecta · Coleoptera ·
Coccinellidae · Harmonia · Harmonia axyridis · ·
SPECIES · Harmonia axyridis (Pallas, 1773) ·
Harmonia axyridis · (Pallas, 1773) · SE ·
Botaniska trädgården, Uppsala, Upl · Uppland ·
PRESENT · 2 · b8323864-602a-4a7d-9127-
bb903054e97d · 59.85247 · 17.62604 · 200.0 · ·
· · · · 2025-11-09 · 9 · 11 · 2025 · PNVYQ ·
PNVYQ · HUMAN_OBSERVATION · SLU Artdatabanken ·
Artportalen · 129617894 · · · · CC0_1_0 ·
User11855 · User11855 · · · 2026-10-
03T12:56:13.689Z · ·
```

One record, all 50 values, as your computer sees it · gbifID 5896573334

Today



1

Metadata

The files about the data: who made it, and how

2

Darwin Core

What the columns mean, and what GBIF did to them

3

Values

Agreed lists, ISO formats, and DOI, ORCID, ROR, UUID

4

TDWG

Who decides, and how you can have a say

Zooming in: from the file to its columns to the values.

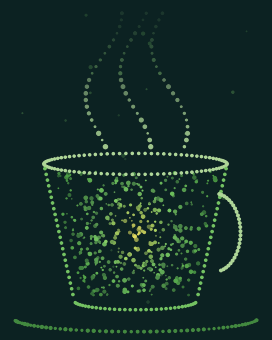
01

Metadata

The data about the data: who made it, how, where, when, and under which licence.



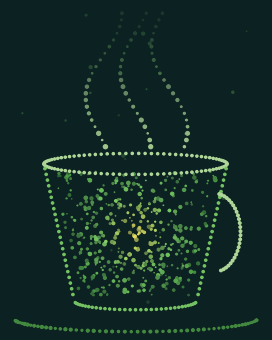
Open the DwC-A zip



`occurrence.txt`	DATA	What GBIF made of the records · 230 columns
`verbatim.txt`	DATA	What the publishers sent · 194 columns
`multimedia.txt`	DATA	About 5,800 photos, each with its own licence
`meta.xml`	METADATA	Which column is which Darwin Core term
`metadata.xml`	METADATA	Your download: the query, the datasets, the counts
`dataset/`	METADATA	One EML file per source dataset: four here
`citations.txt`	METADATA	How to cite each source
`rights.txt`	METADATA	The licence of each source

Five of the eight are about the data. The Simple CSV zip holds one file: the table.

What is metadata?



Data about the data: everything you'd otherwise have to ask the person who made it.

Who?

Who made the data, who publishes them, who to ask

What?

What the dataset contains: title, description, taxa

Where?

The area the data cover

When?

The period they cover, and when they were last updated

How?

How the data were collected: methods, effort, quality checks

On what terms?

The licence, and how to cite it

Written in a standard, so machines can read it too. In GBIF: EML, the Ecological Metadata Language.

Same download, very different metadata



	Artportalen	iNaturalist	Observation.org	NABU|naturgucker
Records in your download	5,053	896	18	6
Licence	CC0	CC BY-NC	CC BY-NC	CC BY
DOI	✓	✓	✓	✓
How the data were collected	–	✓	✓	–
Where (geographic coverage)	✓	✓	✓	–
When (temporal coverage)	–	✓	–	–
Which taxa (taxonomic coverage)	–	–	✓	–
Contact person with an ORCID	✓	–	–	–

Metadata is where fitness for use starts: how were these data collected? (Session 6)

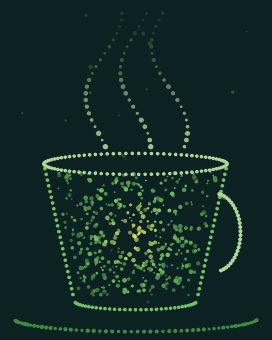
02

Darwin Core Terms

A shared vocabulary: every column has a name, a definition and an address.



Every column name is a term



NAME

`coordinateUncertaintyInMeters`

DEFINITION

A horizontal distance (in meters) from a given dwc:decimalLatitude and dwc:decimalLongitude describing the smallest circle containing the whole of the dcterms:Location. Zero is not a valid value for this term.

ADDRESS (IRI = Internationalized Resource Identifier)

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

CLASS

Location with decimalLatitude, countryCode, locality …

In your download

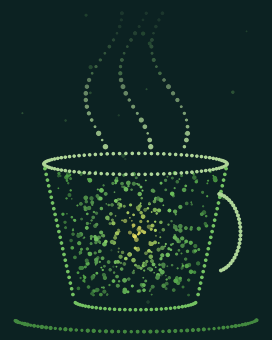
50 m

median, from 1 m to 394 km

197 left empty: unknown, not exact

Look up any column:
dwc.tdwg.org/terms

A simple download, 50 columns, sorted



DARWIN CORE · 33 COLUMNS

Taxon
scientificName · taxonRank · kingdom … genus · infraspecificEpithet

Location
decimalLatitude · decimalLongitude · coordinateUncertaintyInMeters · coordinatePrecision · countryCode · stateProvince · locality

Event
eventDate · year · month · day

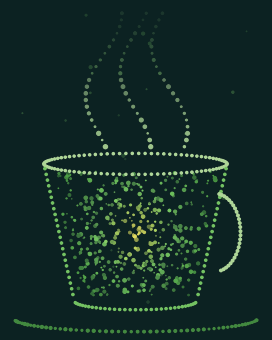
Occurrence
occurrenceID · catalogNumber · recordNumber · recordedBy · individualCount · occurrenceStatus · establishmentMeans

Identification
identifiedBy · dateIdentified · typeStatus

Record
basisOfRecord · institutionCode · collectionCode

Not on dwc.tdwg.org? Then GBIF added it.

Columns that look obvious, and aren’t



`coordinateUncertaintyInMeters`

A radius in metres. Not the same as coordinatePrecision (A decimal representation of the precision of the coordinates given in the dwc:decimalLatitude and dwc:decimalLongitude.”)

`eventDate`

A date or an interval (ISO 8601): 2021-08-24/2021-08-25.

`individualCount`

Empty does neither mean zero nor one:
Absences go in occurrenceStatus.

`license`

Per record, not per file. iNaturalist alone sends three: CC BY-NC (776), CC0 (71) and CC BY (49).

03

Values

What’s in a cell: a word from an agreed list, a format from ISO, or an identifier that points to one thing.



Controlled vocabulary – a list of allowed values



NO SHARED LIST – FREE TEXT

samplingProtocol

observed (806)
light trap (499)
lamp/light trap (389)
sweep-net (66)
sweep netting (6)

Same method, different words: a computer can’t count it

SHARED LIST

basisOfRecord

PreservedSpecimen
FossilSpecimen
LivingSpecimen
MaterialSample
MaterialCitation
HumanObservation
MachineObservation
Occurrence

Every record picks one.

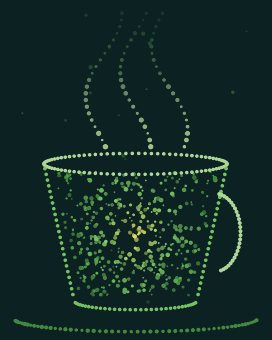
ANOTHER LIST

establishmentMeans

native
nativeEndemic
introduced
nativeReintroduced
vagrant
uncertain
introducedAssistedColonisation

A controlled vocabulary is an agreed list of values, each with a definition and its own IRI.

Borrowed standards: ISO and EPSG



ISO 8601 · DATES

eventDate

2025-11-09

2024-09-09T12:06:02+02:00

All four publishers use it. Year first: no guessing whether 09/11 is September or November.

ISO 3166-1 · COUNTRIES

countryCode

SE NO FI DK IS

All 5,973 records say SE. The free-text column country is empty in 902.

EPSG · COORDINATE SYSTEMS

geodeticDatum

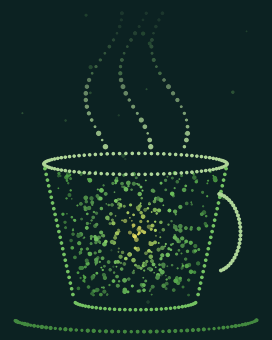
EPSG:4326 (5,949) WGS84 (18)

empty (6)

Same system, two spellings. If empty, GBIF assumes WGS84 and flags it.

Darwin Core doesn't reinvent: dates, countries and coordinate systems use standards that already exist.

Persistent identifiers: who’s who



DATA AND PAPERS

DOI

`10.15468/kllkyl`

- Artportalen on GBIF
- Resolves at doi.org
- Metadata at DataCite

Cite it

PEOPLE

ORCID

`0000-0002-1825-0097`

- recordedByID, identifiedByID
- Dataset metadata
- Log in to GBIF.org

Get credit

ORGANISATIONS

ROR

`ror.org/05k323c76`

- Swedish Museum of Natural History
- GBIF publisher pages
- institutionID of NRM specimens

Credit institutions

THINGS INSIDE GBIF

UUID

`38b4c89f-584c-41bb-bd8f-cd1def33e92f`

- Datasets, publishers, checklists
- Unique, no registry
- Not a link by itself

Join tables

Persistent = unique + resolvable + someone has promised to keep it working.

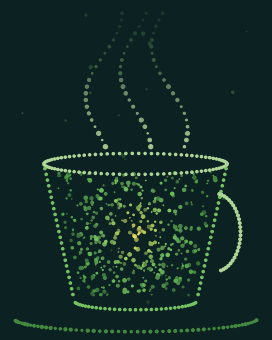
04

TDWG

Who decides what a column means, and how you can have a say.



Who decides?



Biodiversity Information Standards (TDWG)

- A community, not an authority: volunteers since 1985, first as the Taxonomic Databases Working Group
- Darwin Core became a TDWG standard in 2009
- All the work happens in the open, on GitHub
- TDWG conference, basically every year!

The standards in your zip

`occurrence.txt` `verbatim.txt` `meta.xml`

Darwin Core

`verbatim/audubon.txt`

Audiovisual Core, formerly Audubon Core

`license` `rightsHolder`

Dublin Core, from the library world

`metadata.xml` `dataset/*.xml`

EML, Ecological Metadata Language

HAVE YOUR SAY

Standards change, in public



TODAY

1 — **1985** — TDWG starts, as the Taxonomic Databases Working Group

2 — **2009** — Darwin Core ratified

3 — **May 2026** — Conceptual model and Data Package (DwC-DP)

4 — **13 Oct 2026** — July batch of term changes ratified

5 — **16 Nov 2026** — Next Darwin Core public review opens

6 — **4 Dec 2026** — MIDS review closes: what counts as “digitised”?

Anyone can comment, no membership needed: **github.com/tdwg/dwc**

Read the definition, not the column name.

Metadata says how it was made. Darwin Core says what a column means.
Vocabularies and identifiers say what a value means.

YOUR TURN

Look up three columns

Open your session 2 download, or any GBIF download.

Pick the three columns your analysis depends on.

Look each one up at **dwc.tdwg.org/terms**. Does it mean what you assumed?

Want a say? The Darwin Core public review opens 16 November: **github.com/tdwg/dwc**



Questions?

Lena Thöle · lena.thole@nrm.se · gbif@nrm.se

Slides & recording: gbif.se/data-fika



GBIF Global Biodiversity Information Facility

SBDI Swedish Biodiversity Data Infrastructure

Swedish Museum of Natural History

Vetenskapsrådet Swedish Research Council

Next fika

Session 4 · Friday 11 December, 11:00–11:45

The GBIF Data Model

How rows connect: occurrences, events and measurements.

Thank you for the fika – see you next month.

gbif.se/data-fika

Slides, recording and all materials

GBIF Global Biodiversity Information Facility

SBDI Swedish Biodiversity Data Infrastructure

Swedish Museum of Natural History

Vetenskapsrådet Swedish Research Council

Links

- Darwin Core quick reference guide
- Darwin Core on GitHub: issues and public reviews
- TDWG
- MIDS public review, until 4 December
- GBIF data quality recommendations
- GBIF: occurrenceID and identifier stability
- GBIF Metadata Profile (EML)
- ORCID
- ROR
- Bionomia

