

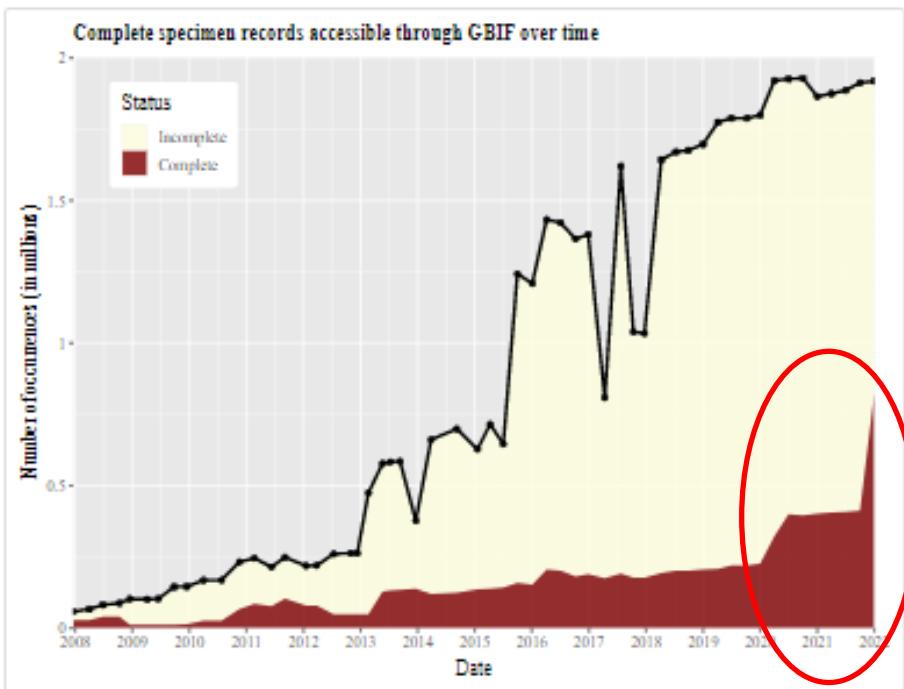
Datakwaliteit en hoe deze te verbeteren met behulp van GBIF

Jeroen Creuwels, Datamanager NLBIF

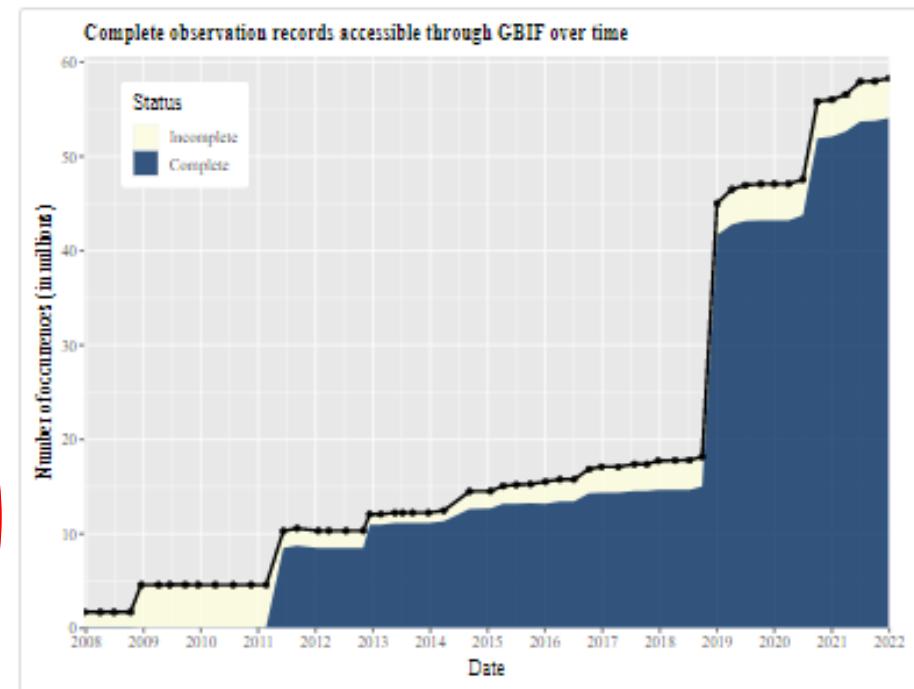
Compleetheid van de NL-data bij GBIF

“Complete = identification at least to **species rank**, valid **coordinates**,
a **full date** of occurrence and a given **basis of record**”

Specimen records



Observation records



Incomplete

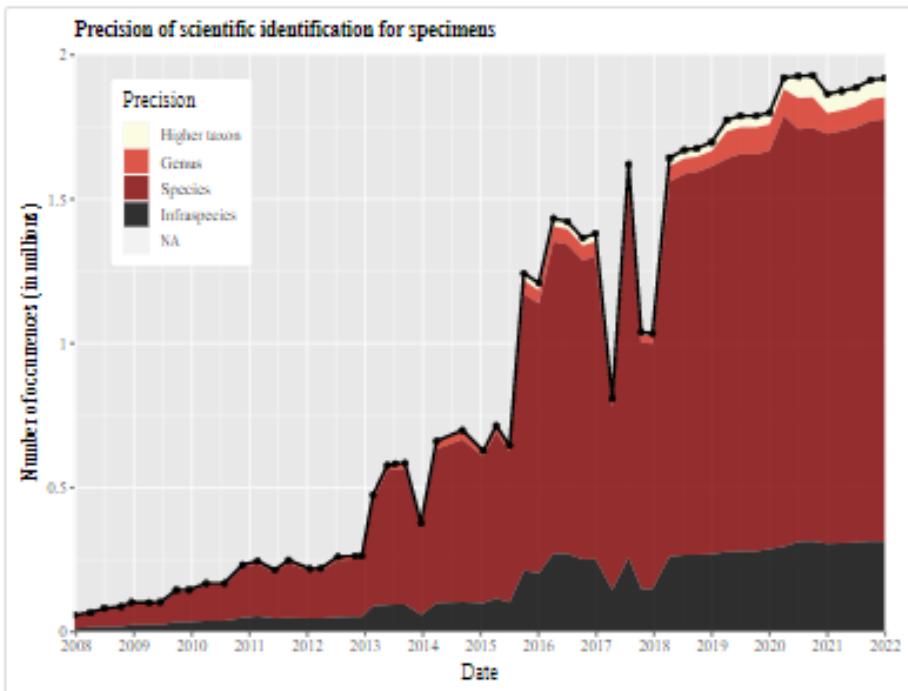
Complete

Incomplete

Complete

Taxonomische precisie

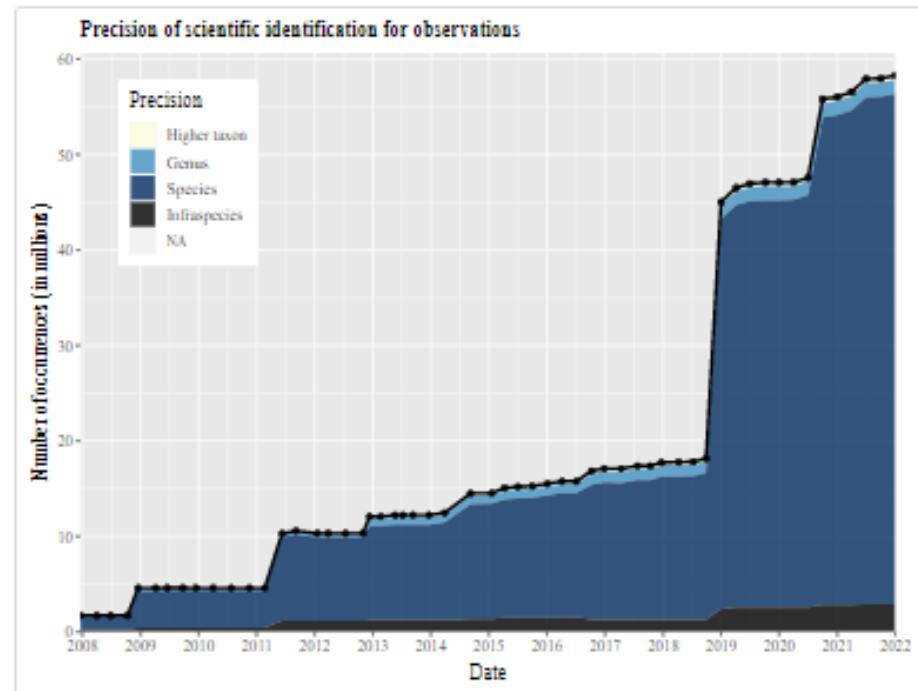
Specimen records



Species

Infraspecies

Observation records

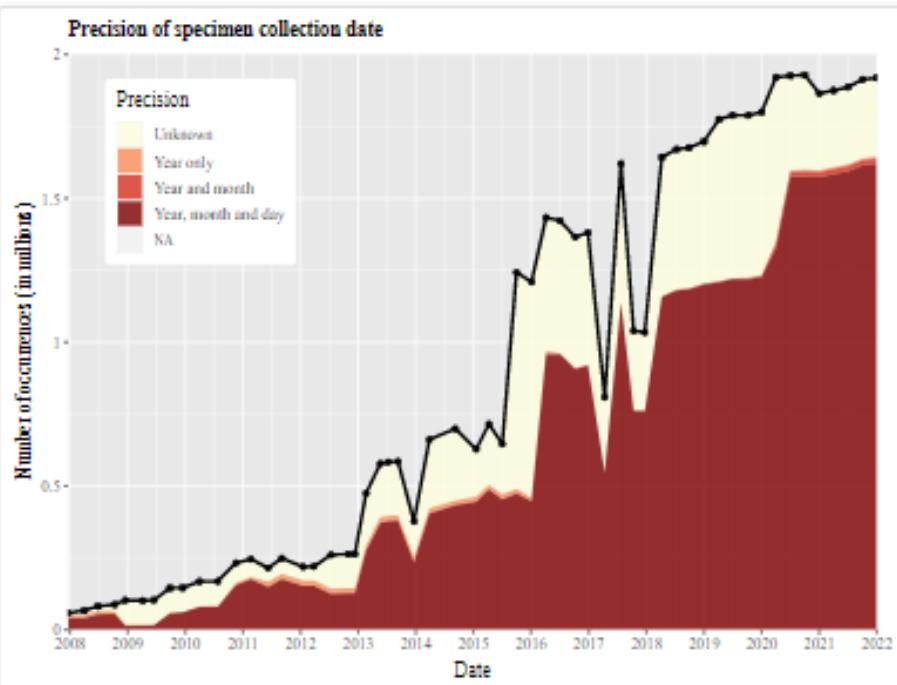


Species

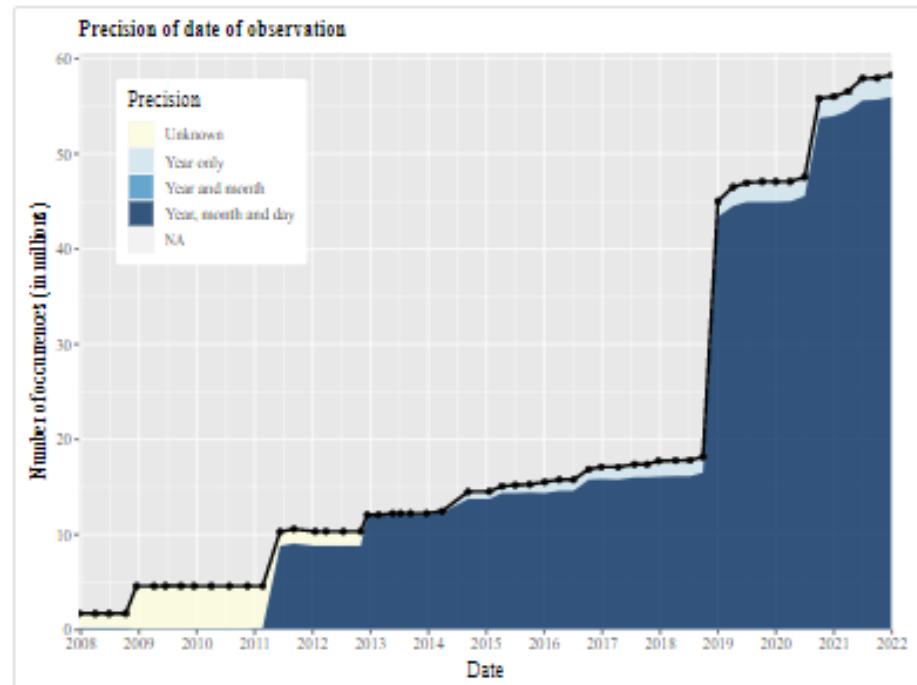
Infraspecies

Temporele precisie

Specimen records



Observation records



Unknown

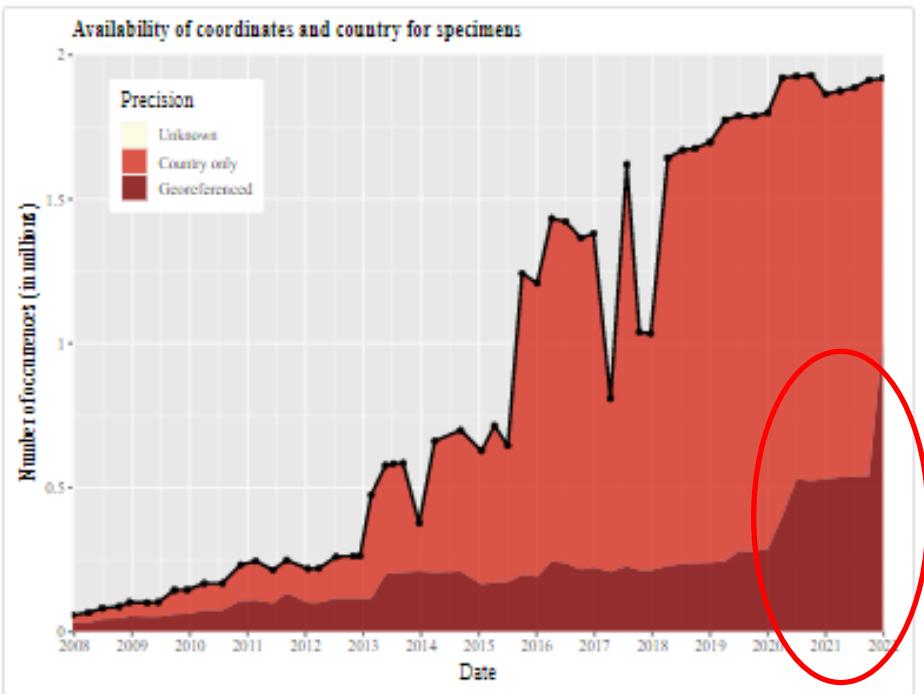
Year + Month + Day

Unknown

Year + Month + Day

Geografische precisie

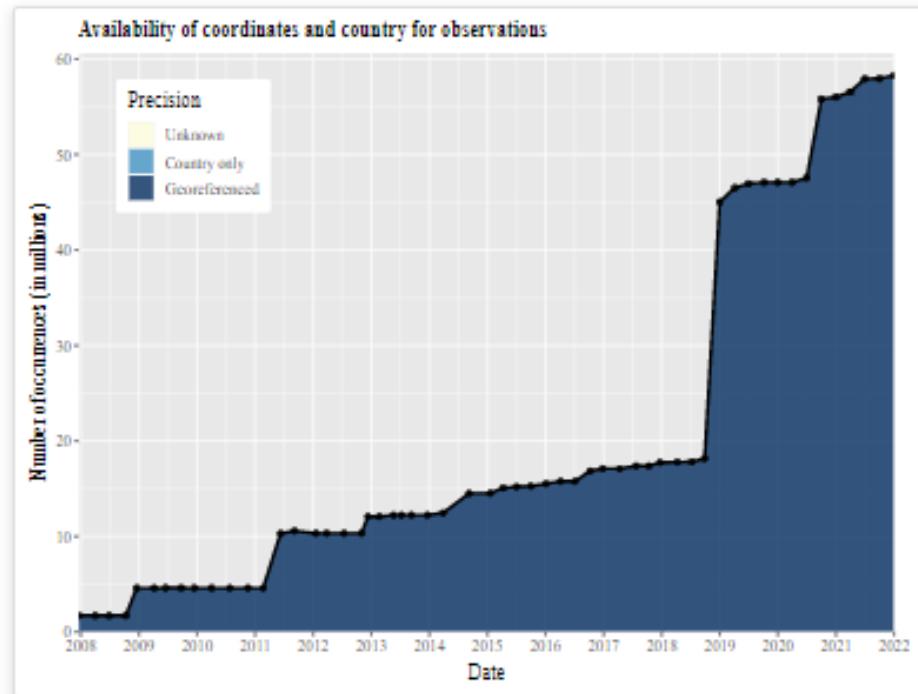
Specimen records



Countries only

Georeferenced

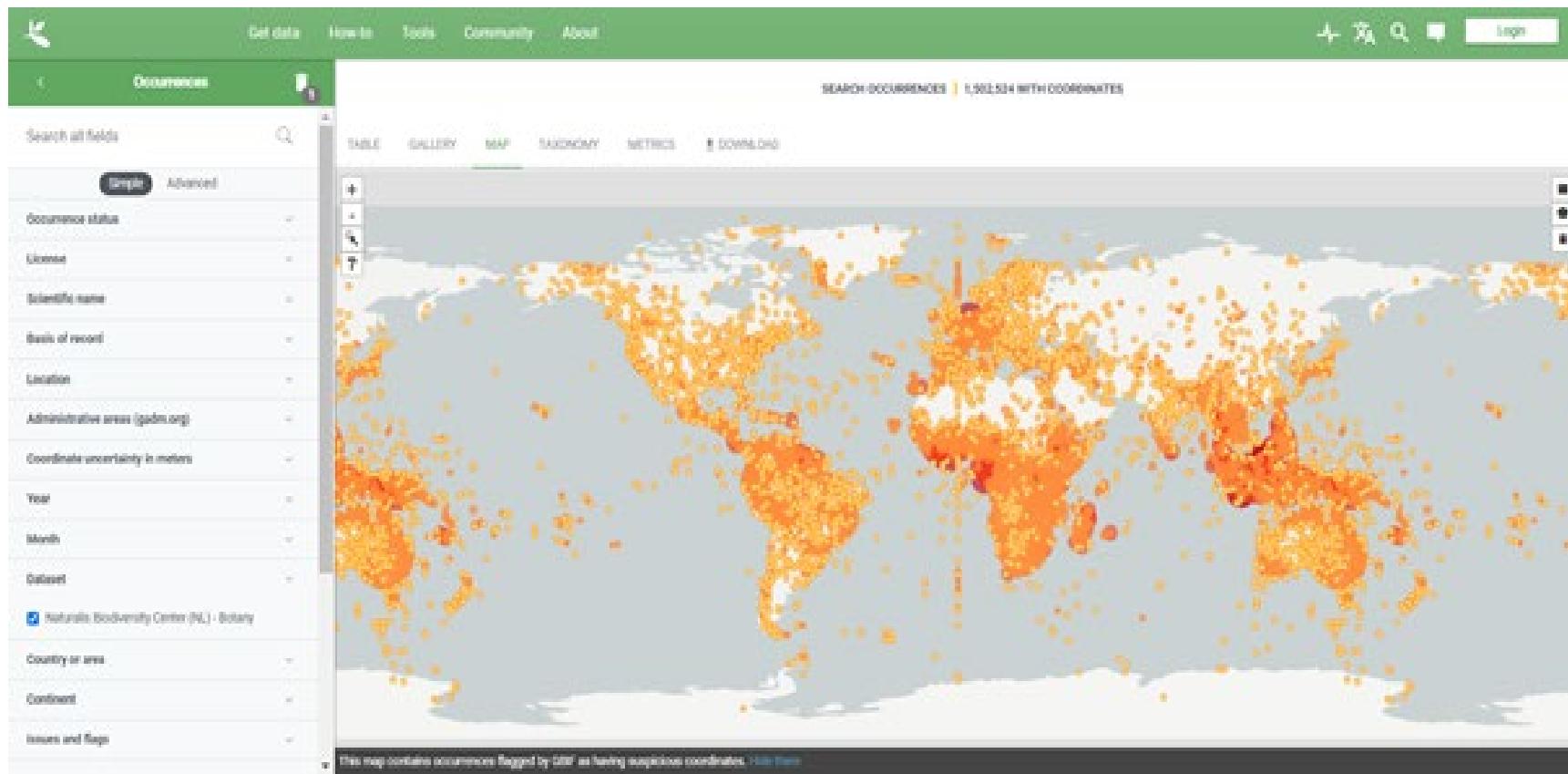
Observation records



Georeferenced

Hoe kun je de datakwaliteit verhogen?

- restricties/aanbevelingen bij invoer
- controle na publicatie bij GBIF (met flags/issues)
- en data opruimen
(in bron-database)



Darwin Core records

- *Minimum quality standards*

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required

Darwin Core records

- *Minimum quality standards*

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required
Examples	PreservedSpecimen , FossilSpecimen , LivingSpecimen , MaterialSample , Event , HumanObservation , MachineObservation , Taxon , Occurrence , MaterialCitation

Darwin Core records

- *Minimum quality standards*

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required

Recommended best practice is to use a date that conforms to **ISO 8601-1:2019**.

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required

Recommended best practice is to use a date that conforms to **ISO 8601-1:2019**.

YYYY-MM-DD

- | | |
|------------|--|
| 1809-02-12 | (some time during 12 February 1809). |
| 1906-06 | (some time in June 1906). |
| 1971 | (some time in the year 1971). |
| 1900/1909 | (some time between the beginning of the year 1900 and the end of the year 1909). |

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required

Recommended best practice is to use a date that conforms to **ISO 8601-1:2019**.

YYYY-MM-DD

1809-02-12 (some time during 12 February 1809).

1906-06 (some time in June 1906).

1971 (some time in the year 1971).

1900/1909 (some time between the beginning of the year 1900 and the end of the year 1909).

2007-11-13/15

2009-02-20T08:40Z

2018-08-29T15:19

2007-03-01T13:00:00Z/2008-05-11T15:30:00Z

Darwin Core records

- *Minimum quality standards*

Term	Status
occurrenceID	Required
basisOfRecord	Required
scientificName	Required
eventDate	Required
countryCode	Strongly recommended
taxonRank	Strongly recommended
kingdom	Strongly recommended
decimalLatitude & decimalLongitude	Strongly recommended
geodeticDatum	Strongly recommended
coordinateUncertaintyInMeters	Strongly recommended
individualCount, organismQuantity & organismQuantityType	Strongly recommended

Fouten, outliers en onnauwkeurigheden opsporen met “flags and issues” bij GBIF

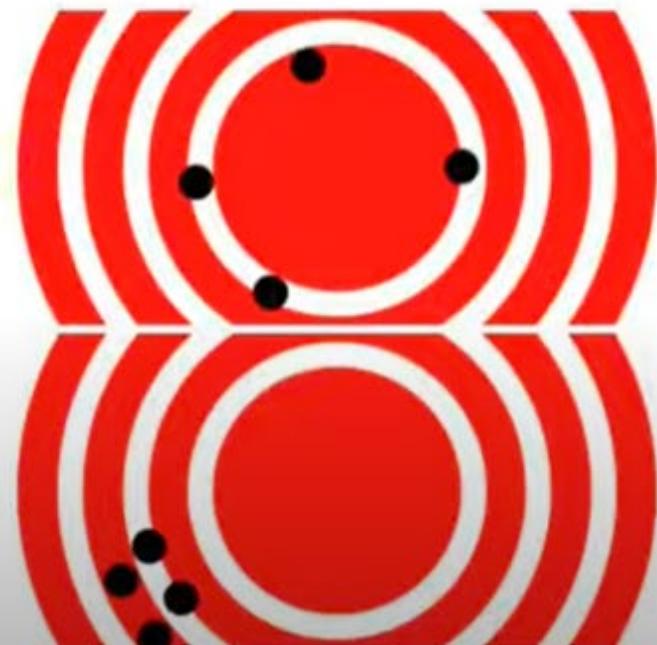
MEASURES OF QUALITY

Correctness (Accuracy)

How close is the recorded value to the actual value?

Consistency (Precision)

How often do you get it right?



Fouten, outliers en onnauwkeurigheden opsporen met “flags and issues” bij GBIF

“

“All data include error – there is no escaping it! It is knowing what the error is that is important, and knowing if the error is within acceptable limits for the purpose to which the data are to be put.”
– Chapman 2005

From: Chapman, A. D. 2005. Principles of Data Quality, version 1.0.

OCCURRENCES PER ISSUES AND FLAGS 10 meest voorkomende flags/issues bij NL-data



NEXT



OCCURRENCES PER ISSUES AND FLAGS

Issues and flags	Count
Occurrence status inferred from individual count	44,321,611

Examples

present , absent

Occurrence

Term	Interpreted	Original	Remarks
Individual count	1	1	
Occurrence ID	18068	18068	
Occurrence remarks	Staart met wervelkolom. Laten liggen.	Staart met wervelkolom. Laten liggen.	
Occurrence status	PRESENT		Occurrence status inferred from individual count
Recorded by	Sabine	Sabine	
Sex		Onbekend	Excluded

OCCURRENCES PER ISSUES

Issues and flags

Occurrence status inferred

Coordinate rounded

WHAT THE NUMBER OF DIGITS IN YOUR COORDINATES MEANS

LAT/LON PRECISION	MEANING
28°N, 80°W	YOU'RE PROBABLY DOING SOMETHING SPACE-RELATED
28.5°N, 80.6°W	YOU'RE POINTING OUT A SPECIFIC CITY
28.52°N, 80.68°W	YOU'RE POINTING OUT A NEIGHBORHOOD
28.523°N, 80.683°W	YOU'RE POINTING OUT A SPECIFIC SUBURBAN CUL-DE-SAC
28.5234°N, 80.6830°W	YOU'RE POINTING TO A PARTICULAR CORNER OF A HOUSE
28.52345°N, 80.68309°W	YOU'RE POINTING TO A SPECIFIC PERSON IN A ROOM, BUT SINCE YOU DIDN'T INCLUDE DATUM INFORMATION, WE CAN'T TELL WHO
28.5234571°N, 80.6830941°W	YOU'RE POINTING TO WALDO ON A PAGE
28.523457182°N, 80.683094159°W	"HEY, CHECK OUT THIS SPECIFIC SAND GRAIN!"
28.523457182818284°N, 80.683094159265358°W	EITHER YOU'RE HANDING OUT RAW FLOATING POINT VARIABLES, OR YOU'VE BUILT A DATABASE TO TRACK INDIVIDUAL ATOMS. IN EITHER CASE, PLEASE STOP.

From: <https://xkcd.com/2170/>

NEXT



OCCURRENCES PER ISSUES AND FLAGS

Issues and flags	Count
Occurrence status inferred from individual count	44,321,611
Coordinate rounded	13,240,937
Institution match none	8,512,722

Record

Term	Interpreted	Original	Remarks
Basis of record	Preserved specimen	PreservedSpecimen	
Collection code	Mammalia	Mammalia	
Institution ID	Naturalis Biodiversity Center	Naturalis Biodiversity Center	Institution match none

Term	Interpreted	Original	Remarks
Basis of record	Preserved specimen	PreservedSpecimen	
Collection code	NMR Collection Specimen collection of the Natural History Museum Rotterdam	NMR Collection	
Collection ID	187a8a55-7de6-4230-8250-7d4ee274004f Specimen collection of the Natural History Museum Rotterdam	187a8a55-7de6-4230-8250-7d4ee274004f	
Institution code	NMR Natural History Museum Rotterdam	NMR	
Institution ID	http://grbio.org/cool/jyde-k516 Natural History Museum Rotterdam	http://grbio.org/cool/jyde-k516	



OCCURRENCES PER ISSUES AND FLAGS

Issues and flags	Count
Occurrence status inferred from individual count	44,321,611
Coordinate rounded	13,240,937
Institution match none	8,512,722
Country derived from coordinates	6,002,934

Location

Term	Interpreted	Original	Remarks
Continent	EUROPE	Europe	
Coordinate uncertainty in metres	111	111	
Country or area	Netherlands		Country derived from coordinates
Country code	NL		Country derived from coordinates
Decimal latitude	51.496	51.496	Country derived from coordinates
Decimal longitude	4.774	4.774	Country derived from coordinates
Geodetic datum	WGS84	WGS84	Country derived from coordinates
Locality	B_HS_val 2_processiepark	B_HS_val 2_processiepark	
Location ID	e254a13c-26e8-483d-b664-4a0f1f4e9995	e254a13c-26e8-483d-b664-4a0f1f4e9995	



OCCURRENCES PER ISSUES AND FLAGS

Issues and flags	Count
Occurrence status inferred from individual count	44,321,611
Coordinate rounded	13,240,937
Institution match none	8,512,722
Country derived from coordinates	6,002,934
Recorded date invalid	1,553,002

Event			
Term	Interpreted	Original	Remarks
Day	25		Recorded date invalid
Month	2		Recorded date invalid
Year	1964		Recorded date invalid
Event date	1964-02-25T18:12:00	1964-02-25T18:12:00Z/1964-02-26T7:33:00Z	Recorded date invalid

Term	Interpreted	Original	Remarks
Year		1972-1995	Recorded date invalid

Term	Interpreted	Original	Remarks
Day	2		Recorded date unlikely
Month	7		Recorded date unlikely
Year	1577		Recorded date unlikely
Event date	1577-07-02		Recorded date unlikely



OCCURRENCES PER ISSUES AND FLAGS

Issues and flags	Count
Occurrence status inferred from individual count	44,321,611
Coordinate rounded	13,240,937
Institution match none	8,512,722
Country derived from coordinates	6,002,934
Recorded date invalid	1,553,002
Footprint WKT invalid	1,434,407
Taxon match higherrank	1,388,311
Individual count invalid	479,672
Geodetic datum assumed WGS84	456,998
Basis of record invalid	369,638

NEXT

Voorbeeld: Taxon match higherrank

Taxon

Term	Interpreted	Original	Remarks
Kingdom	Plantae	Plantae	Taxon match higherrank
Phylum	Tracheophyta		Taxon match higherrank
Class	Magnoliopsida	Magnoliopsidae	Taxon match higherrank
Order	Gentianales	Gentianales	Taxon match higherrank
Family	Rubiaceae	Rubiaceae	Taxon match higherrank
Genus	Psychotria	Psychotria	Taxon match higherrank
Specific epithet		varians	Taxon match higherrank
Higher classification	Plantae Magnoliopsidae Gentianales Rubiaceae	Plantae Magnoliopsidae Gentianales Rubiaceae	
Nomenclatural code	ICN	ICN	
Scientific name	Psychotria L.	Psychotria varians O.Lachenaud	Taxon match higherrank
Scientific name authorship		O.Lachenaud	Taxon match higherrank
Rank	Genus	species	Altered
Taxonomic status	Accepted		Inferred

Voorbeeld: Taxon match higherrank

Taxon

Term	Interpreted	Original	Remarks
Kingdom	Plantae	Plantae	
Phylum	Tracheophyta		Inferred
Class	Magnoliopsida	Magnoliopsidae	Altered
Order	Gentianales	Gentianales	
Family	Rubiaceae	Rubiaceae	
Genus	Psychotria	Psychotria	
Specific epithet	varians	varians	
Generic name	Psychotria		Inferred
Higher classification	Plantae Magnoliopsidae Gentianales Rubiaceae	Plantae Magnoliopsidae Gentianales Rubiaceae	
Nomenclatural code	ICN	ICN	
Scientific name	Psychotria varians O.Lachenaud	Psychotria varians O.Lachenaud	
Scientific name authorship	O.Lachenaud		Excluded
Rank	Species	species	
Taxonomic status	Accepted		Inferred

Taxonomie: “flags and issues”

Wanneer geen exacte match van de wetenschappelijke naam met de GBIF taxonomic backbone wordt gevonden dan komen de volgende ‘flags’ voor:

Flag	Meaning	Example
Taxon match fuzzy	A match with a different spelling was found.	<i>Pelagodes antiquadrarius</i> and <i>Pelagodes antiquadraria</i>
Taxon Match higherrank	No match was found at the same taxonomic rank but one was found for a higher rank.	<i>Hylatomus pileatus</i> and <i>Hylatomus</i>
Taxon match none	No match was found.	<i>Flagellate</i>

Voorbeelden van geografische issues/flags

Country or area	Norfolk Island	Australia	Country derived from coordinates
Country code	NF	AU	Country derived from coordinates
Decimal latitude	-31.559919	-31.5599193572998	Coordinate rounded
Decimal longitude	167.386169	167.386169433594	Coordinate rounded
Decimal latitude	52.904243	6.243582	Presumed swapped coordinate
Decimal longitude	6.243582	52.904243	Presumed swapped coordinate
Geodetic datum	WGS84	WGS84	
Locality	Doldersummerveld	Doldersummerveld	
State province	Drenthe	Drenthe	
Country or area	France	France	
Country code	FR		Inferred
Decimal latitude	49.4372	49.4372	Presumed negated longitude
Decimal longitude	2.6287	-2.6287	Presumed negated longitude

Hoe issues/flags te vinden bij GBIF (gbif.org)

The screenshot shows the GBIF Occurrences search results page. At the top, there are navigation links: Get data, How-to, Tools, Community, About, and a Login button. Below the header, the search results are displayed with a green header bar containing the title "Occurrences". A sidebar on the left contains various filters: Simple (selected), Advanced, Occurrence status (with a warning icon), Licence, Scientific name, Basis of record, Location, Administrative areas (gadm.org), Coordinate uncertainty in metres, Year, Month, Dataset, Country or area, Continent, Issues and flags, and Zero coordinate (with a count of 498). The "Netherlands" filter under "Country or area" is highlighted with a red border. The main content area shows a table of search results with the following columns: TABLE, GALLERY, MAP, TAXONOMY, METRICS, and DOWNLOAD. The results list various bird species and their details:

Scientific name	Country or area	Coordinates	Month & year	Basis of record	Dataset
<i>Cyanistes caeruleus</i> (Linnaeus, 1758)	Netherlands	52.3N, 4.6E	2022 January	Human observation	Xeno-car
<i>Streptopelia decaocto</i> (Frivaldszky, 1838)	Netherlands	52.3N, 4.6E	2022 January	Human observation	Xeno-car
<i>Phocoena phocoena</i> (Linnaeus, 1758)	Netherlands		2022 January	Human observation	walvisstr
<i>Phocoena phocoena</i> (Linnaeus, 1758)	Netherlands	52.7N, 4.6E	2022 January	Human observation	walvisstr
<i>Ardea alba</i> Linnaeus, 1758	Netherlands	52.6N, 6.9E	2022 January	Human observation	naturguc
<i>Mergus merganser</i> Linnaeus, 1758	Netherlands	52.6N, 6.9E	2022 January	Human observation	naturguc
<i>Podiceps nigricollis</i> C.L.Brehm, 1831	Netherlands	51.3N, 5.7E	2022 January	Human observation	naturguc
<i>Tachybaptus ruficollis</i> (Pallas, 1764)	Netherlands	52.7N, 6.9E	2022 January	Human observation	naturguc
<i>Ardea cinerea</i> Linnaeus, 1758	Netherlands	52.6N, 6.9E	2022 January	Human observation	naturguc
<i>Calidris maritima</i> (Brünnich, 1764)	Netherlands	51.8N, 3.8E	2022 January	Human observation	iNaturalist
<i>Parasteatoda tepidariorum</i> (C.L.Koch, 1841)	Netherlands	52.9N, 4.8E	2022 January	Human observation	iNaturalist
<i>Branta canadensis</i> (Linnaeus, 1758)	Netherlands	51.9N, 4.7E	2022 January	Human observation	iNaturalist

Occurrences		
	Occurrences	
Country or area		 1
<input checked="" type="checkbox"/> Netherlands		
Continent		
Issues and flags		
<input type="checkbox"/> Zero coordinate	498	
<input type="checkbox"/> Coordinate out of range	17,282	
<input type="checkbox"/> Coordinate invalid	140	
<input type="checkbox"/> Coordinate rounded	8,793,869	
<input type="checkbox"/> Geodetic datum invalid	13,563	
<input type="checkbox"/> Geodetic datum assumed WGS84	511,187	
<input type="checkbox"/> Coordinate reprojected	2,231	
<input type="checkbox"/> Coordinate reprojection failed	0	
<input type="checkbox"/> Coordinate reprojection suspicious	439	
<input type="checkbox"/> Coordinate accuracy invalid	0	
<input type="checkbox"/> Coordinate precision invalid	8,522	
<input type="checkbox"/> Coordinate uncertainty metres invalid	36,113	
<input type="checkbox"/> Coordinate precision uncertainty mismatch	0	
<input type="checkbox"/> Footprint SRS invalid	55	
<input type="checkbox"/> Footprint WKT invalid	117,713	
<input type="checkbox"/> Country coordinate mismatch	15,029	
<input type="checkbox"/> Coordinate precision uncertainty mismatch	0	
<input type="checkbox"/> Footprint SRS invalid	55	
<input type="checkbox"/> Footprint WKT invalid	117,713	
<input type="checkbox"/> Country coordinate mismatch	15,029	
<input type="checkbox"/> Country mismatch	2	
<input type="checkbox"/> Country invalid	31,711	
<input type="checkbox"/> Country derived from coordinates	5,560,958	
<input type="checkbox"/> Continent country mismatch	0	
<input type="checkbox"/> Continent invalid	2,114	
<input type="checkbox"/> Continent derived from coordinates	0	
<input type="checkbox"/> Presumed swapped coordinate	231	
<input type="checkbox"/> Presumed negated longitude	190	
<input type="checkbox"/> Presumed negated latitude	55	
<input type="checkbox"/> Recorded date mismatch	5	
<input type="checkbox"/> Recorded date invalid	159,649	
<input type="checkbox"/> Recorded date unlikely	894	
<input type="checkbox"/> Taxon match fuzzy	161,867	
<input type="checkbox"/> Taxon match higherrank	712,614	
<input type="checkbox"/> Taxon match aggregate	0	
<input type="checkbox"/> Taxon match none	112,526	
<input type="checkbox"/> Depth not metric	0	
<input type="checkbox"/> Depth unlikely	1	
<input type="checkbox"/> Depth min/max swapped	1	