Species occurrence location data in GBIF (Global Biodiversity Information Facility) accessible through either a [web interface](https://www.gbif.org/occurrence/search?q=) or through the [GBIF API](https://www.gbif.org/developer/summary) via [Python](https://pygbif.readthedocs.io/en/v0.2.0/) or [R](https://ropensci.org/tutorials/rgbif_tutorial/), contains numerous fields for record-level information.

There may be as few as [50 fields](https://tools.gbif.org/dwca-validator/extension.do?id=dwc:Occurrence) or less (but as many as 170) for each record, many of which are not filled out for most species. These fields follow the Darwin Core standards developed by TDWG ([Taxonomic Databases Working Group](https://www.tdwg.org/)).

In collating source data, GBIF and their data providers use Darwin Core standards. To assure data quality, GBIF uses established criteria whose requirements are described in detail [here](https://www.gbif.org/data-quality-requirements-occurrences).

A complete reference guide of terms for each Darwin Core occurrence or event is available [here](https://dwc.tdwg.org/terms/).

GBIF Important Fields

These fields (or columns) are ones most likely to be useful in filtering data for terrestrial vertebrates (amphibians, birds, mammals, reptiles) in North America. They will not necessarily be appropriate for all taxa groups under all time spans or all geographies. Note also that not all of these fields will exist for all species and for many records they will be blank. GBIF in general follows the Darwin Core standard reference terms, however, many fields are unique to GBIF and there is no metadata (at least I couldn’t find any) that describes field schema protocols or standardizations.

**‡** Indicates a Darwin Core reference term

**\*** Indicates a field NOT in Darwin Core reference terms

**basisOfRecord‡**

This denotes the nature of the collection and can be one of the following:

* Preserved specimen
* Fossil specimen
* Living specimen
* Material sample
* Human observation
* Machine observation
* Event
* Taxon
* Occurrence
* Unknown

**coordinateUncertaintyInMeters‡**

If a record has latitude and longitude coordinates (which is mandatory for modeling and/or model assessment purposes), this field will contain the assumed distance from those coordinates in which the species was recorded within.

From Darwin Core references: *“The horizontal distance (in meters) from the given decimalLatitude and decimalLongitude describing the smallest circle containing the whole of the Location. Leave the value empty if the uncertainty is unknown, cannot be estimated, or is not applicable (because there are no coordinates). Zero is not a valid value for this term.”*

**countryCode‡**

Obviously this is the standardized code of the country in which the occurrence location is recorded. It will always be set to ‘US’ for our purposes. Darwin Core recommended best practices abbreviation is the ISO 3166-1-alpha-2 country code.

**decimalLatitude‡**

From the Darwin Core references: *“The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.”*

**decimalLongitude‡**

From the Darwin Core references: *“The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.”*

**eventRemarks‡ / locality‡ / locationRemarks‡ / occurrenceRemarks‡**

These three fields contain textual descriptive information about the occurrence as related by observers. It could potentially be useful for a detailed examination of individual records.

**geodeticDatum‡**

This should always be ‘WGS84’

From the Darwin Core references: *“The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based.”*

**gbifID\***

A supposedly unique numerical identifier for a single location record. Whether this numeric code is a persistently global unique identifier is unknown. This *is not* a Darwin Core standard reference term. See **occurrenceID** below.

**individualCount‡**

The number of individual organisms represented at the time or the recorded occurrence.

**infraspecificEpithet‡**

The “lowest” or “terminal” name of the species’ scientific name – i.e. the subspecies name. From Darwin Core references: “*The name of the lowest or terminal infraspecific epithet of the scientificName, excluding any rank designation.”* This could be used to identify occurrences that only refer to a subspecies without having to parse the scientific name field (see **species**).

**issues\***

Records in this column are a combination of standardized elements in a bracketed list format identifying potential problems with location coordinates, time period, and/or taxonomy. This *is not* a Darwin Core standard reference term. Such “issues” include:

* COORDINATE\_REPROJECTED
* COORDINATE\_ROUNDED
* COORDINATE\_UNCERTAINTY\_METERS\_INVALID
* GEODETIC\_DATUM\_ASSUMED\_WGS84
* GEODETIC\_DATUM\_INVALID
* PRESUMED\_SWAPPED\_COORDINATE
* RECORDED\_DATE\_MISMATCH
* TAXON\_MATCH\_FUZZY

**institutionCode‡**

This appears to be a fairly robust alphabetic code for data origins such as museum acronyms (e.g. NCSM for North Carolina Museum of Natural Sciences) or citizen science aggregators such as iNaturalist. It may need to be combined with **institutionID** just to make sure all possible information about a record source is identified. From Darwin Core references: “*The name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record.”*

**lifeStage‡**

From Darwin Core references: “*The age class or life stage of the biological individual(s) at the time the Occurrence was recorded.”* Some examples include:

* ADULT
* JUVENILE

**month‡**

Ordinal month in which the event occurred. Could be used for determining seasonal status of the recorded occurrence.

**occurrenceID‡**

A unique code for the individual occurrence record. From GBIF terms: *“A unique identifier for the occurrence, allowing the same occurrence to be recognized across dataset versions as well as through data downloads and use (see* [Darwin Core Terms: A quick reference guide](http://rs.tdwg.org/dwc/terms/#occurrenceID)*)  
Ideally, the occurrenceID is a persistent global unique identifier. As a minimum requirement, it has to be unique within the published dataset. It allows to recognize the same occurrence over time when the dataset indexing is refreshed; it links additional data like images; and it makes it possible to cite records e.g. in usage reports or in publications. This means that the occurrenceID needs to reliably stay with the occurrence at source, and to consistently refer to the same occurrence in published datasets and any underlying source data.”*

This is an alphanumeric string often taking the form of a URL source string such as

“https://www.inaturalist.org/observations/19800810”

or other formats such as “187DAF22-74AE-4D6D-A190-BEB8DED8424A”. It may be possible to use the **gbifID** in place of this field to identify unique occurrence records.

**species\***

The latin binomial (i.e. scientific name) for the given taxon without authorship and date text information included in the field as is the case in the **scientificName‡** field. This *is not* a Darwin

Core standard reference term.

**speciesKey\***

A unique numeric code for a given species concept. It appears to be consistent across taxonomic changes. For example, *Clethrionomys gapperi* changed to *Myodes gapperi* in 2005, but the speciesKey is 5706760 for records under both scientific names. This *is not* a Darwin Core standard reference term.

**stateProvince‡**

US state text of occurrence location. Darwin Core describes this as: *“The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the Location occurs.”*

**taxonRank‡**

From Darwin Core description: *“The taxonomic rank of the most specific name in the scientificName.”*  Could refer to a species or subspecies. Possible values include:

* SPECIES
* SUBSPECIES

**taxonomicStatus‡**

From Darwin Core description: *“The status of the use of the scientificName as a label for a taxon. Requires taxonomic opinion to define the scope of a taxon. Rules of priority then are used to define the taxonomic status of the nomenclature contained in that scope, combined with the experts opinion. It must be linked to a specific taxonomic reference that defines the concept.”* Possible values include:

* ACCEPTED
* SYNONYM

**vernacularName‡**

The common (or vernacular) name. Could potentially be more than one common name because of taxonomic changes or inclusion of subspecies taxa.

**year\***

*“The four-digit year in which the Event occurred, according to the Common Era Calendar.”* This *is not* a Darwin Core standard reference term.

**GAP Species Occurrence Record Summary**

As of approximately 12:00 EST, 8 January 2019 the 1590 of GAP modeled full species (using GAP scientific names) have 253,717,720 GBIF occurrences in the United States containing coordinate data with no “geospatial issues”.

Here are some summary stats:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **n Full Species** | **n Occurrences (US & have coordinates)** | **mean**  **# of occ./species** | **std dev.** | **median**  **# of occ./species** | **min** | **max** |
| 1590 | 253717720 | 159571 | 556774 | 2703 | 0 | 6230463 |

This data was gathered using the Python script *GBIFRecordFieldsCount.py*. It uses the pyGBIF API package to assemble information on species occurrences in GBIF. Using the **occurrences** module and its **search** function, it’s possible to obtain a record count under given search criteria. The search criteria used for the 1590 GAP species included occurrences that were 1) only recorded in the United States, 2) must have latitude and longitude coordinates, 3) with no “geospatial issues”. Using this search, a Python dictionary is returned for each species. One of the keys in that dictionary is a ‘count’. For 10 species, the search returned a dictionary that did not contain a ‘count’ key and therefore a record count was not enumerated. It was possible to use the GBIF web search to find the number of occurrences for these 10 species. Those numbers were incorporated in the above summary stats. The 10 species were:

Accipiter cooperii, Actitis macularius, Erethizon dorsata, Oreortyx pictus, Pantherophis vulpine, Poecile hudsonica, Puma yagouaroundi, Rhinella marinus, Spermophilus saturates, Taricha sierrea

*NOTE: the script took approximately 90 minutes to run on the list of 1590 species*.

Because the only way to summarize and quantify the availability of certain parameters within species records requires “downloading” all the records for a given species, and that download can only occur 300 records at a time, it was necessary to take a sample of species with a relatively fewer number of records. Note also that not all occurrence record datasets have the same number of GBIF fields (with or without Darwin Core standard reference terms). It also appears that each individual occurrence can have a different set of fields for the record. At a minimum, the record must have four of the Darwin Core standards including ***occurrenceID***, ***basisOfRecord***, ***scientificName*** and ***eventDate***. These fields are required, however the following are “strongly recommended”: ***countryCode***, ***taxonRank***, ***kingdom***, ***decimalLatitude*** & ***decimalLongitude***, ***geodeticDatum***, ***coordinateUncertaintyInMeters***, ***individualCount***/ ***organismQuantity***/***organismType***. Here are some summary stats for species with <30000 records with coordinates in the US:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **n Full Species**  **(< 30000 records)** | **n Occurrences**  **(US & have coordinates)** | **mean**  **# of occ./species (± sd)** | **median**  **# of occ./species** | **mean**  **# of fields in species occurrence dataset** |
| 1154 | 3939839 | 3414 (5423) | 1122 | 127 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Records Statistic** | | | | |  |
| **Parameter** | mean # recs. WITH parameter | mean # recs. WITHOUT parameter | mean % WITH | mean % WITHOUT | mean of parameter mean | median of parameter mean |
| Coordinate Uncertainty | 1267 | 2146 | 53.1 | 46.9 | 9822 | 6383 |
| Locality Textual Descriptions | 3721 | -- | -- | -- | -- | -- |
| Geodetic Datum Information | 3414 | 0 | -- | -- | -- | -- |
| Month Recorded | 3287 | 127 | -- | -- | -- | -- |
| Year Recorded | 3307 | 107 | -- | -- | -- | -- |
| Scientific Name | 3414 | 0 | -- | -- | -- | -- |
| Vernacular (common) Name | 190 | 3224 | -- | -- | -- | -- |