Finding Data Gaps in the GBIF Backbone Taxonomy

John Thomas Waller [‡]

‡ GBIF, Copenhagen, Denmark

Corresponding author: John Thomas Waller (jhnwllr@gmail.com)

Abstract

When publishers supply GBIF (Global Biodiversity Information Facility) with a dwc:scientificName, this name is sometimes *not* found in the GBIF taxonomic backbone. The backbone is needed to organize occurrences on GBIF. In these cases, the occurrence records get a data quality flag called taxon match higher rank. This means that GBIF was only able to match the name to a higher rank. Matching is a process whereby a name supplied by the publisher is compared to a name in the already existing in the GBIF backbone taxonomy.

At GBIF, we would always like to match the name supplied by the publisher to the lowest rank possible, so that when a user comes to GBIF looking for a certain name, they will have access to the largest amount of occurrence data possible.

The main goals of this project were:

1. Identify the types of issues that prevent matching occurrences to the backbone that come in with an identification at species level (or below) to backbone names at that same rank.
2. Identify the responsible actors (GBIF processing, occurrence record curators, missing checklist) who are best placed to help improve the name.

In Fig. 1, I divide unique names from occurrences supplied to GBIF from publishers that have received the taxon match higher rank flag. Here we see that GBIF is probably missing many names from Coleoptera (Beetles) and Lepidoptera (Butterflies/Moths).

Publishers to GBIF sometimes do not provide enough information in the dwc:scientificName for GBIF to choose between names in the backbone Fig. 2. If a publisher only supplied GBIF with "*Glocianus punctiger*" we would not be able to determine between the two choices, and it would get moved to the higher rank (genus *Glocianus*).

Publishers also supply GBIF with a variety of what I call unmatchable names, which are names that are impossible to match to the GBIF backbone. Sometimes these names are

acceptable names, but still missing from the backbone, like missing hybrids or OTUs (Operational Taxonomic Units). Other names are simply bad names that we can't expect to fix. Some examples below:

Table 1

It is often hard to tell if a missing name is a real data gap. To check, I randomly sampled five possibly missing names from each group from Fig. 1 to check if I could manually locate a source outside GBIF with the name.

Around 50% (44 of 86) of the possibly missing names appear to be genuinely missing from the GBIF backbone. We can therefore conservatively assume that there are thousands of missing names in the GBIF backbone. Keep in mind, however, that many missing names are missing synonyms—that is, they are not unique taxon concepts. Taking half of 50% (25%), we can make a conservative minimum missing names Table 2.

As a data publisher, there are a few things that can be done to improve name matching to the GBIF backbone.

- Run your dataset through the data validator
- Match your names to the GBIF backbone before publishing using species lookup or rgbif
- Add authorship if appropriate
- Fill known higher-taxonomy
- Try to avoid working name placeholders for the dwc:scientificName
- Do not put identification qualifiers in the dwc:scientificName field but rather use the dwc:identificationQualifier field.

Keywords

taxonomic backbone, scientific name, data quality

Presenting author

John Thomas Waller

Presented at

TDWG 2022

Conflicts of interest

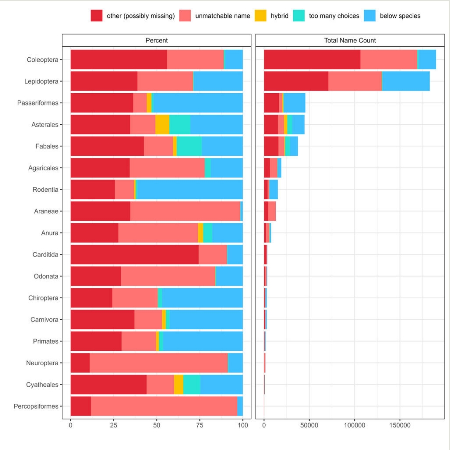
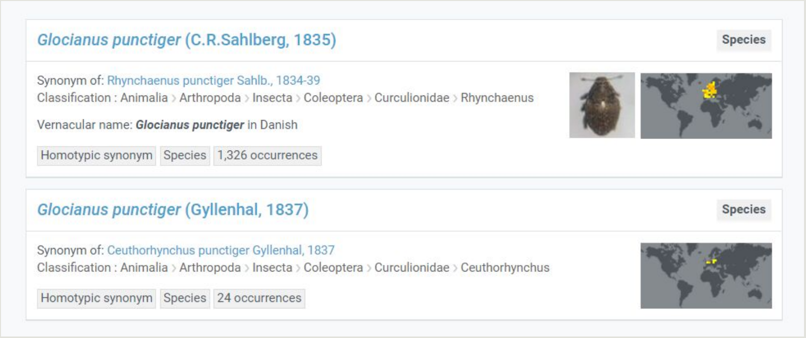


Figure 1.

Unique names from occurrences supplied to GBIF from publishers that have received the taxon match higher rank flag.

- **other**: means that my alogrithm could not find a good reason for this name not matching. This could be a misspelling or the name could be missing from the GBIF backbone. These are names that might reflect data gaps.
- **unmatchable name**: is a catch-all group for poorly formatted or unmatchable names. (see Table 1).
- **hybrid** (hybrid formula): means the name refers to a hybrid. We expect poor checklist coverage for hybrid names.
- **below species**: means a name at a taxonomic rank below the species level could not be matched at that level. Usually we expect less checklist coverage for subspecies and varieties.
- **too many choices**: GBIF has two or more names with different authorship (homonyms), but the publisher does not provide authorship and/or higher taxonomy, so the name cannot be matched unambiguously.



Glocianus punctiger (C.R.Sahlberg, 1835) Species

Synonym of: Rhynchaenus punctiger Sahlb., 1834-39
Classification : Animalia › Arthropoda › Insecta › Coleoptera › Curculionidae › Rhynchaenus

Vernacular name: ***Glocianus punctiger*** in Danish

Homotypic synonym | Species | 1,326 occurrences

Glocianus punctiger (Gyllenhal, 1837) Species

Synonym of: Ceuthorhynchus punctiger Gyllenhal, 1837
Classification : Animalia › Arthropoda › Insecta › Coleoptera › Curculionidae › Ceuthorhynchus

Homotypic synonym | Species | 24 occurrences

Figure 2.

Too many choices. Authorship is needed to decide between these entries in the backbone.

Table 1.

Unmatchable (or hard to match) names.

name not matched	**reason**
Mystery mystery	bad name
Sonus naturalis	bad name
Bambusoideae spec.	subfamily name
Coleoptera indet.	order name
Astarte juv.	genus name with life stage
Gen. sp.	bad name
Astarte sp. BIOUG14667-B01	family with id
Phoneutria depilata (Strand 1909) sp. reval.	species name with remark
Anisoptera Unknown Dragonfly Species	infra-order name with remarks
Zygoptera	suborder name
Philodromus Philodromus albidus / rufus	doubtful identification (alternative)
Certhia brachydactyla/Certhia familiaris	doubtful identification (alternative)
Corvus corone x C. cornix	hybrid
BOLD:ADV7315	OTU (Operational Taxonomic Unit)
BOLD:ADX5419	OTU

Table 2.

Conservative minimum missing names. Based on conservative judgment, 25% of potentially missing names are genuinely absent from the GBIF backbone. Download a full table of possibly missing names from the groups above here.

group	friendly name	min estimated missing names
Coleoptera	Beetles	26,600
Lepidoptera	Butterflies	17,700
Passeriformes	Bird order	4,200
Fabales	Plant order	4,100
Asterales	Plant order	4,000
Agaricales	Mushrooms	1,600
Araneae	Spiders	1,200
Rodentia	Rodents	1,100
Carditida	Bivalves	700
Anura	Frogs	600
Carnivora	Carnivores	300
Odonata	Dragonflies	300
Chiroptera	Bats	200
Cyatheales	Ferns	100
Primates	Primates	100
Neuroptera	Insect order	<100
Percopsiformes	Fish order	<100