Mobilizing Bat1K

through versioned, machine readable and automatically generated data publications.

Jorrit Poelen (UC Santa Barbara, Ronin Institute, GloBI)

2024-10-27

Guiding Questions

How to keep track of Bat1K data corpus?

How to cite specific versions of the Bat1K data corpora?

How to share specific versions of the Bat1K data corpora?

Reuse, reuse, reuse.

- ▶ Darwin Core Archive -> GBIF, GloBI
- ▶ Taxonomic Alignment Tools [1] -> align with specific versions MDD, NCBI Taxonomy
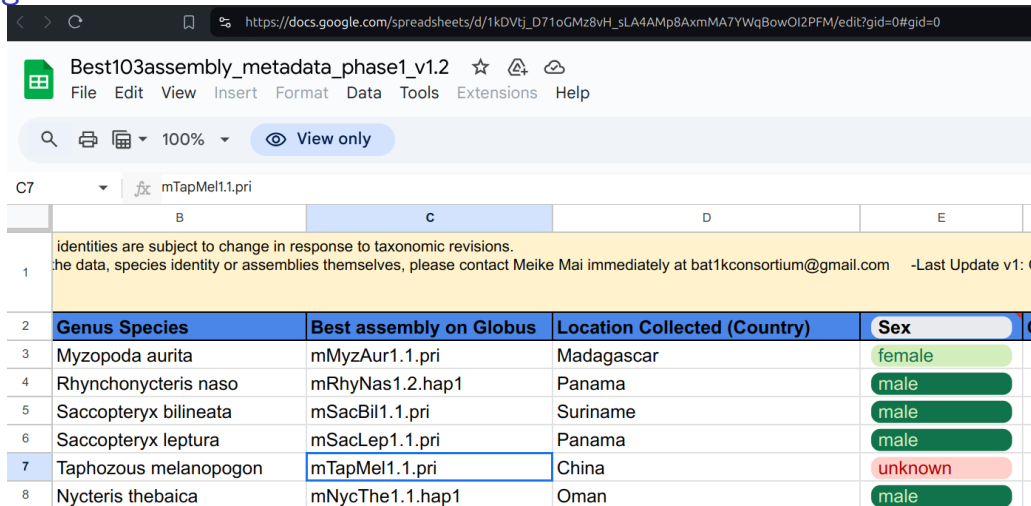- ▶ Signed Data Citations [2]

---

[1] https://github.com/globalbioticinteractions/nomer

[2] Elliott M.J., Poelen, J.H. & Fortes, J.A.B. (2023) Signing data citations enables data verification and citation persistence. *Sci Data*. https://doi.org/10.1038/s41597-023-02230-y
hash://sha256/f849c870565f608899f183ca261365dce9c9f1c5441b1c779e0db49df9c2a19d

At NASBR 2024

Jorrit -> Ariadna -> Sonja -> Meike -> Bat1K data

Google Sheet

Figure 1: Best103assembly_metadata_phase1_v1.2 accessed at
https://docs.google.com/spreadsheets/d/1kDVtj_D71oGMz8vH_sLA4AMp8AxmMA7YWqBowOI2PFM
on 2024-10-27.

Google Sheet -> Versioned, Machine Readable Data Package

```
preston track\
 --message "Bat1K Genome Index"\
https://docs.google.com/spreadsheets/d/1kDVtj_D71oGMz8vH_sLA4AMp8AxmMA7YWqk
 | sha256sum
```

Deriving bat1k.tsv

```
preston cat\
 --remote https://linker.bio,https://softwareheritage.org\
 hash://sha256/710cccc378e6d41e7d2e214bcaf08af76886d9df6e389dc0177c1460fb5
 | grep hasVersion\
 | grep tsv\
 | preston cat\
 | tail -n+2\
 | tee bat1k.tsv
```

Taxonomic Alignment through Nomer: Find (mis-)Alignments

Using Nomer [3]

```
cat bat1k.tsv\
 | nomer append\
 --properties <(echo 'nomer.schema.input=[{"column":0,"type":"externalId"}
 | grep -v HAS_ACCEPTED_NAME\
 | cut -f2\
 | tail -n+2

Lasiurus ega
Hipposideros swinhoii
```

[3]Poelen, J. H. (ed .) . (2024). Nomer Corpus of Taxonomic Resources
hash://sha256/b60c0d25a16ae77b24305782017b1a270b79b5d1746f832650f2027ba536e276
hash://md5/17f1363a277ee0e4ecaf1b91c665e47e (0.27) [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.12695629

Next Steps

- ▶ Versioned Bat1K -> DwC-A - Add meta.xml and eml.xml to describe schema according to Darwin Core Archive to enable:
  - ▶ Indexing by GBIF as bat occurrences through vouchered specimen/genome
  - ▶ Indexing by GloBI as bat<>human interactions evidenced by vouchered specimen/genomes to enable automated data reviews [4]

---

[4]Geiselman, Cullen K. & Sarah Younger. 2020. Bat Eco-Interactions Database. www.batbase.org
https://github.com/globalbioticinteractions/batbase/archive/9c65cfeee1a054f9db8cd8bf6892017fd1b3c840.zip
2024-10-25T22:39:59.352Z 6755e9ff065849a8a7472858e98b62458fab93e4c20006f823e844a3ee77f5f2 see
also https://depot.globalbioticinteractions.org/reviews/globalbioticinteractions/batbase/

Take Aways

- ► track, version and package original data
- ► implement automated data review workflow
- ► implement automated data product workflow

A Review of Biotic Interactions and Taxon Names Found in globalbioticinteractions/batbase

by Nomer and Elton, two naive review bots

review@globalbioticinteractions.org

https://globalbioticinteractions.org/contribute

https://github.com/globalbioticinteractions/batbase/issues

2024-10-26

Abstract

Life on Earth is sustained by complex interactions between organisms and their environment. These biotic interactions can be captured in datasets and published digitally. We present a review process of such an openly accessible digital interactions dataset of known origin, and discuss its outcome. The dataset under review, named globalbioticinteractions/batbase, is 5.24KiB in size and contains 16,491 interaction with 7 unique types of associations (e.g., preysOn) between 601 primary taxa (e.g., Carollia perspicillata) and 3,483 associated taxon (e.g., Lepidoptera). The report includes detailed summaries of interactions data as well as a taxonomic review from multiple catalogs.