Data audit for technical evaluation of

An updated checklist of Azorean arthropods (Arthropoda)

Downloaded 6 supplementary files from ARPHA platform and
source archive **dwca-checklist_arthropoda_azores-v1.2.zip** on 2022-11-15 from
https://www.gbif.org/dataset/2d91cfd8-0a48-4d80-8128-080e52a1e650

Dr Robert Mesibov (robert.mesibov@gmail.com; https://www.datafix.com.au)
2022-11-15

About this evaluation

Pensoft does a technical evaluation of the dataset (or datasets) referred to in the data paper. If the dataset passes or has only minor problems, the data paper manuscript is referred to reviewers. If the dataset has major problems, a review of the paper is postponed until the dataset has been corrected.

To see what features of a dataset are checked in a technical evaluation, please go to

https://zookeys.pensoft.net/about#DataQualityChecklistandRecommendations

Please note that Pensoft does not check the details of the *content* of a dataset, for example whether the correct author is given for a scientific name, or whether the correct latitude/longitude is given for a locality.

Recommendation. There are quite a few minor data problems that should be fixed before the data paper goes to review.

--

(1) Four of the 6 supplementary files are Microsoft Excel worksheets. While a spreadsheet might be convenient for compiling data, it is not a suitable format for publishing or sharing data, and Excel (.xlsx) is a proprietary format. For optimal re-usability, please convert the 4 Excel files to plain text tables (TSV or CSV) in UTF-8 encoding with plain line endings (linefeed only), not Windows line endings (carriage return + linefeed).

The easiest way to do the conversion is to copy all the active cells in the spreadsheet to the clipboard, then paste into a good-quality text editor (Notepad++ is recommended for Windows users). The result is a tab-separated, plain text file which can be saved as either .tsv or .txt. When saving, use the text editor options to ensure UTF-8 encoding and plain line endings.

See below for additional recommendations about formatting of individual tables.

--

Suppl. material 1: List of species additions by GBIF

(2) There are no identifiers for these records. Since they are all GBIF occurrences, please add a GBIF identifier. For example, the June 2020 observation of *Neomyia cornicina* by Rodrigo Medeiros in GBIF is

> https://www.gbif.org/occurrence/2850563659
>
> (or alternatively for this particular record, https://www.inaturalist.org/observations/50619741).
>
> The identifiers can go in the new field *occurrenceID*. It is important to add identifiers in this particular table because there is more information available in GBIF for each of the 18 records.

(3) Please convert *dateIdentified* from MM/YYYY to ISO 8601 format, YYYY-MM.

--

Suppl. material 2: Complete list of Azorean arthropods

(4) In *Colonization Status*, 4 records have "m" and 3 records have "Migrant". Please be consistent.

(5) In *Colonization Status*, the entry "INDT" for 124 records is not explained in the table caption. Is it "indeterminate", in other words the same as the blank for 291 other records?

(6) Minor issues in the island fields:

> **No. of records | AZ**
> 2304 |
> 10 | Az

103 | AZ

No. of records | FAI
1426 |
1 | ?FAI
988 | FAI
2 | FAI?

No. of records | TER
1077 |
1 | Jardim AH *[unexplained]*
10 | LIFE PV *[unexplained]*
2 | QUA *[unexplained]*
1 | ?TER
1326 | TER

(7) Please convert the no-break space in this entry in *scientificName*: "Coenosia freyi{HERE}freyi Tiensuu, 1945" to a plain space.

(8) Please trim the data items to remove the many trailing spaces in *order*, *family* and *scientificName*, e.g. "Meinertellidae "

--

Suppl. material 3: Darwin Core database - Updated Checklist of Azorean Arthropods (Taxon)

(9) Please convert the no-break space in this entry in *scientificName* and *acceptedNameUsage*: "Coenosia freyi{HERE}freyi Tiensuu, 1945" to a plain space.

(10) The same taxon has been given different *taxonID*s and *acceptedNameUsageID*s:

No. of records | scientificName | taxonID
---
1 | Diplura | 9c33f17c-6e36-45fe-8dc5-c6b75c7f1557
1 | Diplura | df9f0124-e8fe-444d-a491-485566820171
---
1 | Protura | efeba62f-f773-4899-868d-97d0f8a24494
1 | Protura | 27d795fe-3d43-4640-870d-e193cbe02a70
---
1 | Symphyla | 99ef3e57-7e36-4d22-91c1-45de78c413f3
1 | Symphyla | 8ae5fd24-ff99-4762-b15a-b19f6273a88e

No. of records | acceptedNameUsage | taxonID
---
1 | Diplura | 9c33f17c-6e36-45fe-8dc5-c6b75c7f1557
1 | Diplura | df9f0124-e8fe-444d-a491-485566820171
---
1 | Protura | efeba62f-f773-4899-868d-97d0f8a24494
1 | Protura | 27d795fe-3d43-4640-870d-e193cbe02a70
---
1 | Symphyla | 99ef3e57-7e36-4d22-91c1-45de78c413f3
1 | Symphyla | 8ae5fd24-ff99-4762-b15a-b19f6273a88e

No. of records | acceptedNameUsage | acceptedNameUsageID
---
1 | Diplura | 9c33f17c-6e36-45fe-8dc5-c6b75c7f1557
1 | Diplura | df9f0124-e8fe-444d-a491-485566820171
---
1 | Protura | efeba62f-f773-4899-868d-97d0f8a24494
1 | Protura | 27d795fe-3d43-4640-870d-e193cbe02a70
---
1 | Symphyla | 99ef3e57-7e36-4d22-91c1-45de78c413f3
1 | Symphyla | 8ae5fd24-ff99-4762-b15a-b19f6273a88e

(11) Entries in *identificationRemarks* would be clearer as (for example) "Azores Bioportal Code A02503".

(12) Please trim data items to remove trailing spaces.

(13) Please convert *modified* dates to ISO 8601 (YYYY-MM-DD).

--

Suppl. material 4: Darwin Core database - Updated Checklist of Azorean Arthropods (Distribution)

(14) There is a problem with *locationID*:

No. of records | locality | locationID

---
312 | Corvo | COR
2 | Corvo | COR?
---
669 | São Jorge | SJG
1 | São Jorge | SJG?

1043 | Faial | FAI
1 | Faial | FAI?

If the location is uncertain, it would be better to put that information in a new field *locationRemarks*.

(15) I'm not sure I understand how *establishmentMeans* is used in this table:

No. of records | taxonID | establishmentMeans
---
2 | 1b39c3d6-fbb3-4b05-b1d3-4d40c07aac2d |
6 | 1b39c3d6-fbb3-4b05-b1d3-4d40c07aac2d | Macaronesia
---
7 | 175252c8-d054-4d38-a215-d9e85e6386bf | Native
1 | 175252c8-d054-4d38-a215-d9e85e6386bf | Introduced
---
3 | c5168f3a-3a45-4c9a-8f3c-6d6465dd32e3 | Native
1 | c5168f3a-3a45-4c9a-8f3c-6d6465dd32e3 | Introduced
---
1 | 385914af-06f7-4ec8-a032-26363846552f | Native
4 | 385914af-06f7-4ec8-a032-26363846552f | Macaronesia
---
3 | 44aa5dbf-5d82-4b0f-99aa-640bf4c27f49 |
5 | 44aa5dbf-5d82-4b0f-99aa-640bf4c27f49 | Endemic
---
1 | 1fde67a8-0dc4-4fcc-a80e-c2e9cd1d1720 |
4 | 1fde67a8-0dc4-4fcc-a80e-c2e9cd1d1720 | Introduced
---
1 | be020b9c-94b6-45f7-8d88-a13176436aac |
1 | be020b9c-94b6-45f7-8d88-a13176436aac | Native
---
1 | fe2aa72c-6557-4d42-a288-19cf90d93b09 |
1 | fe2aa72c-6557-4d42-a288-19cf90d93b09 | Macaronesia
---
4 | 44807e3d-6204-4daf-8288-d5334bb547ab | Native
1 | 44807e3d-6204-4daf-8288-d5334bb547ab | Introduced

For example:

c5168f3a-3a45-4c9a-8f3c-6d6465dd32e3 | A02088 | Faial | FAI | Native
c5168f3a-3a45-4c9a-8f3c-6d6465dd32e3 | A02088 | Pico | PIC | Native
c5168f3a-3a45-4c9a-8f3c-6d6465dd32e3 | A02088 | São Miguel | SMG | Native
c5168f3a-3a45-4c9a-8f3c-6d6465dd32e3 | A02088 | Terceira | TER | Introduced

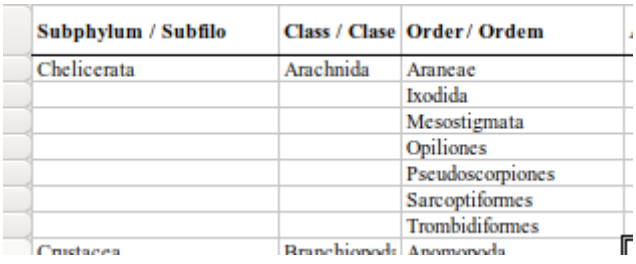
Does this mean *Psychoda albipennis* is introduced to Terceira but native to the other islands?

--

Suppl. material 5: Table of total taxa (species and subspecies) recorded in the updated Azorean arthropods checklist

(16) When converting this Excel file to plain text, please re-organise it to avoid two formatting problems:

(a) Do not leave blank spaces, e.g. in the subphylum and class fields:



Subphylum / Subfilo	Class / Clase	Order / Ordem
Chelicerata	Arachnida	Araneae
		Ixodida
		Mesostigmata
		Opiliones
		Pseudoscorpiones
		Sarcoptiformes
		Trombidiformes
Crustacea	Branchiopoda	Anomopoda

(b) Please combine the 3 sub-tables for clarity. Instead of 3 separate datasets for species, subspecies and combined species and subspecies, format the data as "Total taxa | Species | Subspecies", e.g. Ixodida in the AZ column would be "11 | 9 | 2". A separator other than "|" could be used, e.g. "/", ";" etc.

--

Suppl. material 6: Table of endemic taxa (species and subspecies) recorded in the updated Azorean arthropods checklist

(17) See (16), above.

--

(18) In the GBIF Darwin Core archive, the field called *establishmentMeans* in your distribution supplementary file and in the data paper is called *occurrenceRemarks* in **distribution.txt** (see also above, 15).

(19) In the GBIF Darwin Core archive, the field called *identificationRemarks* in your taxon supplementary file and in the data paper is called *taxonRemarks* in **taxon.txt** (see also above, 11).