The compilation vector-pathogen dataset provided by Dr. David Simons et al has the potential to be of great value to multiple studies if shared on the GBIF. The authors have done a good job summarizing and standardizing the data as a whole. To assess how best to map these data to the existing GBIF data model using currently accepted Darwin Core standards, I selected 4 datasets from the broader compilation and went through a mapping exercise. I have summarized the steps needed (below) to make this mapping possible. I don’t think it should take too much effort to map the vector data; updating data fields to map the pathogen data will take a bit more effort, however. Additional information about each of the terms can be found using the Darwin Core Quick Reference Guide: <https://dwc.tdwg.org/terms/>.

I hope my work is clear and reproducible. However, Dr. Simons or the other authors are free to contact me ([kingenloff@gbif.org](mailto:kingenloff@gbif.org)) should they have any questions or concerns regarding mapping of the data.

Great job on an excellent compilation!

Cheers,

Kate Ingenloff, PhD

Programme Officer for Science Support, GBIF

NOTE: It's important to note that with the existing data model and data standards, quite of bit of the data that is currently parsed into unique fields will need to be consolidated. I highly recommend retaining a copy of this highly parsed dataset as sometime in the future, we will improve the data model and implement data standards that allow for increased granularity in data reporting. At which time, we could consider re-publishing the compilation dataset using the updated terms and structure.

#-==============================================================================#

Data mapping recommendation

Core: DarwinCore Event

|  |
| --- |
| **🡪** denotes the DarwinCore field an existing term can be mapped to  Terms in **BLUE** are additional fields that should be added |

# 'Study' tab - Event

* **title** **🡪** **datasetName**
* **unique\_id** **🡪** **parentEventID**
* consolidate the following as **samplingProtocol: aim\_1, aim\_2, aim\_3, aim\_detail, metric, trap\_types, trapping\_method, repeated\_visit, diversity measurement, trapping\_effort,** and **pathogen**
* add the field **bibliographicCitation** and provide the full citation to the published study. This should include consolidation of the fields: **link, year\_publication, title, journal\_name,** and **first\_author**
* double check information listed under **license**. Some of this information is indeed licensing information (e.g., CC BY 4.0) but other information should be moved to a difference field, **accessRights** (e.g., all rights reserved).

# 'Study' tab - Occurrence extension

* **speciation** **🡪** **identificationRemarks**

# 'trapping' tab - Occurrence extension

* **unique\_id** **🡪** should be the same **parentEventID** from the study tab
* add a unique **eventID** for each sampling Event in a study. This can be whatever you want, but within the dataset, each sampling event must have its own unique ID. For simplicity in illustrating this, note that I took the parentEventID and added a simple code for the sampling Event location. Example. *ad\_2001\_guinea\_c* indicates sampling of the coast region in parentEventID *ad\_2001\_guinea*)
* add a unique **occurrenceID** for each vector occurrence recorded at a sampling Event. This can be whatever you want, but within the dataset, each occurrence must have its own unique ID. Example. *ad\_2001\_guinea\_c\_crociduraspp* indicates occurrences of Crocidura spp recorded during sampling of the coast region in parentEventID *ad\_2001\_guinea*)
* **year\_trapping** **🡪** **year**
* **month\_trapping** **🡪** **month**
* **country** **🡪** **country**
* **region** **🡪** **stateProvince**
* Note: It may be good to review how well this information is standardized. Some entries appear to be habitat rather than formal regions.
* **town\_village** **🡪** **locality**
* **latitude\_DMS\_N** **🡪** **verbatimLatitude**
* **longitude\_DMS\_W** **🡪** **verbatimLongitude**
* indicate the **verbatimCoordinateSystem** in a new field
* **latitude\_D\_N** **🡪** **decimalLatitude**
* **longitude\_D\_E** **🡪** **decimalLongitude**
* indicate the **geodeticDatum** in a new field
* **UTM\_coordinates** **🡪** **verbatimCoordinates**
* **habitat** **🡪** **habitat**
* **intensity\_use** **🡪** **locationRemarks**
* **genus** **🡪** **genus**
* **species** **🡪** **genericName**
* **number** **🡪** **individualCount**
* add the field **associatedTaxa**. For each occurrence that had a positive identification of one or more pathogens (the third value in obis:measurementValue > 0), list them here using a pipe separator between positively identified taxa (e.g., 'Lassa mammarenavirus IgG Ab | Lassa mammarenavirus Ag')
* consolidate ***trap\_nights, capture\_rate, trap\_night\_unit,*** and ***study\_nights*** under the field **samplingEffort**

# 'pathogen' tab - obis (extended measurement or facts) extension

#### Note: The current data model does not do well with linked data like this, however there are two ways we can add the pathogen data: (1) We include the pathogen information as an extended measurement or fact of the occurrence record it's associated with - this allows us to retain for presence and absence information directly with each record. (2) We can create a unique occurrence for each pathogen tested and then manually link the pathogen presence or absence occurrence to the vector record via their unique occurrence IDs. This second method is nice, because now the pathogen is searchable as its own occurrence, but ensuring the records remain linked requires a lot more work than the first option. In the interest of time and energy, I have proceeded with mapping using the first method.

* add a unique **measurementID** for each pathogen tested. This can be whatever you want, but within the dataset, each occurrence must have its own unique ID. So, because ad\_2001\_guinea\_c\_mastomysspp was tested for two separate pathogens, there will be two separate measurementIDs. Example. *ad\_2001\_guinea\_c\_mastomysspp\_IgG\_Ab* indicates the test for lassa\_mammarenavirus\_IgG\_Ab in the Mastomy spp occurrence sampled in the coastal areas in parentEventID *ad\_2001\_guinea*)
* Note: Records that are a consolidation of multiple or all occurrences for the event should not be included (see lines 3, 16, and 17 of the pathogen tab).
* add the field **measurementValue**. Consolidate the fields ***path\_#, path\_#\_tested,*** and ***(method)\_#\_positive***. Separate values using a pipe separator (e.g., lassa\_mammarenavirus\_IgG\_Ab | 78 | 3)
* add the field **measurementUnit**. Specify the units for the measurements provided in measurementValue separating values using a pipe separator (e.g., pathogen | # pathogen(?) tested | # positive - ab ag)
* add the field **measurementMethod**. This is where you denote the method used to test for the specified pathogen (e.g., pcr, histology)
* add the field **measurementRemarks**. Add any additional information about the pathogen test here.