1/ Data collection

Initial dataset
AntWeb: 62,301 records
GABI: 93,557 records

With GPS

No GPS

Geolocalized occurrences
AntWeb: 59,981 records
GABI: 45,593 records

Other occurrences
AntWeb: 2,320 records
GABI: 47,964 records

2a/ Flags for inspection

CoordinateCleaner flags
Non-terrestrial records
Country **centroids**
Country mismatches
Biodiversity **institutions**
Distance outliers (1000km)

N = 4,143 (3.9%)

Custom flags
All **adm** mismatches
Bioregion outliers

N = 15,244 (14.4%)

2b/ Retrieve GPS coordinates

Screen **Gazetteers** for localities
If absent:
Find smallest **valid adm**
Draw **random coordinates**
Record area as **uncertainty**

N = 50,144 (99.7%)

Geolocalized occurrences
With uncertainty
AntWeb: 61,096 records
GABI: 88,445 records

3/ Filter duplicates

Spatial thinning:
Apply 0.1° grid cell
Keep **unique pairs**
taxa-coordinates

N = 111,413 (74.5%)

4/ Optional filters

Visual checking
of taxa maps
Remove entries with
uncertainty > X km^2

5/ Data analyses

Curated occurrences
With uncertainty
AntWeb: 20,333 records
GABI: 17,795 records