UNIVERSITY OF COPENHAGEN

DEPARTMENT OF COMPUTER SCIENCE

Intro to PhD Project: Automated Visual Taxonomic Identification and Clustering of Insects

PhD Student: Roberta Hunt[1]
Supervisors: Kim Steenstrup Pedersen[1], François Lauze[1]
In collaboration with Alexey Solodovnikov[2] and Anders Drud Jordan[2]
[1]Department of Computer Science, University of Copenhagen [2]Natural History Museum of Denmark

Contact Information:
✉ r.hunt@di.ku.dk

Research Goal

Use and develop state-of-the-art image processing techniques to increase the efficiency and effectiveness of phylogenetic research

Introduction

Determining the phylogeny (evolutionary distance) of insect species is currently a manual and time-consuming task undergone by expert entomologists. We hope to find ways to **make phylogenetic research more efficient** using state-of-the-art deep learning techniques to generate heirarchical clusters of species which can then be used to automatically create a phylogenetic tree where the results of the phylogenetic tree are interpretable.

In this project we will focus on the species rich family of rove beetles (*Staphylinidae*) which contains at least 52,000 known species [1].

Dataset

Much of the phylogenetic research completed is done on so called **'pinned-insect' collections** kept at research facilities in Natural History Museums around the world. A single museum can house millions of insect specimens in this manner. See Figure 1 below. This method of housing specimens makes it very easy to collect and compare images of different specimens, since they tend to have a standard view point (dorsal) and pose.

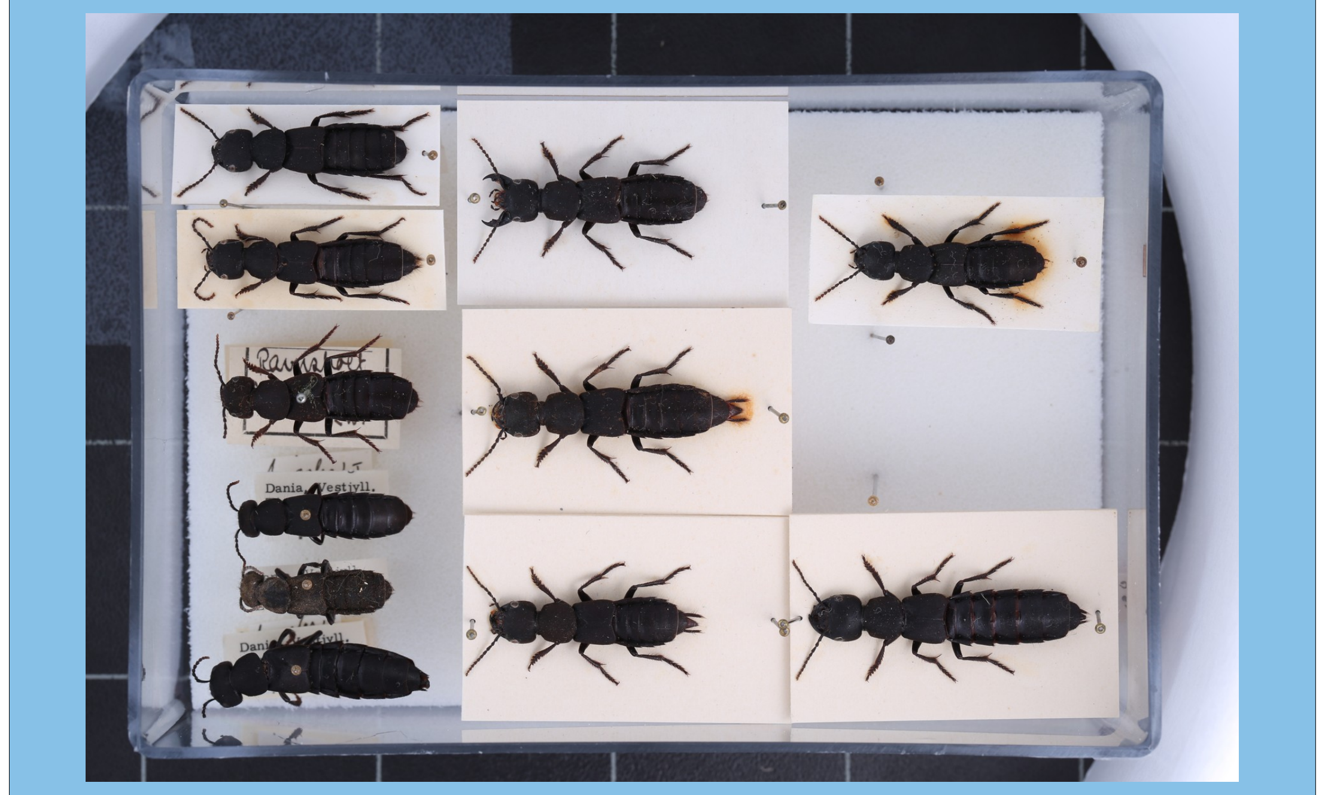


Figure 1: Example image of pinned insect collection from the Natural History Museum of Denmark

Images of over **19,000 specimens** from **218 species** (from 44 genera) of rove beetle has already been collected. An overview of the data distribution is provided in Figure 2 below. This dataset will most likely be supplemented with further images based on expert advice from the entomologists attached to the team as the project progresses.

Research Steps, Unanswered Questions and Potential Directions

1 **Dataset preparation**

Specimen-level cropping, labelling, segmentation and preprocessing

2 **Species-level representation generation:**

Create a representation (embedding, distribution, etc) of each species in a latent space that we can use to calculate the distance between species

Unanswered Questions:
a) How can we train an embedding to specifically encode species-level information and not specimen information.
- Could we use a modified version of self-supervised learning where some percentage of the embedding encodes species level information, and the rest encodes specimen-specific information?

b) Which fields tackle similar problems of creating grouped clusters?
- Fashion? Where many different instances (specimens) and views (poses) of a type of clothing (eg, shirt) exist, which we may want to group together

3 **Generate Interpretable Representations**

Use species level representations to generate representations that biologists can use and understand. This could be a dichotomous tree, or more likely a sketch showing the average or prototypical example of each species.

Unanswered Questions:
a) How can we best generate a prototypical representation of each species?
- Could we use 'deep dreaming' to maximize the classification of each species? (get the most species-like example
- Could we use adverserial networks?
- Could we use autoencoders? Or VAEs?

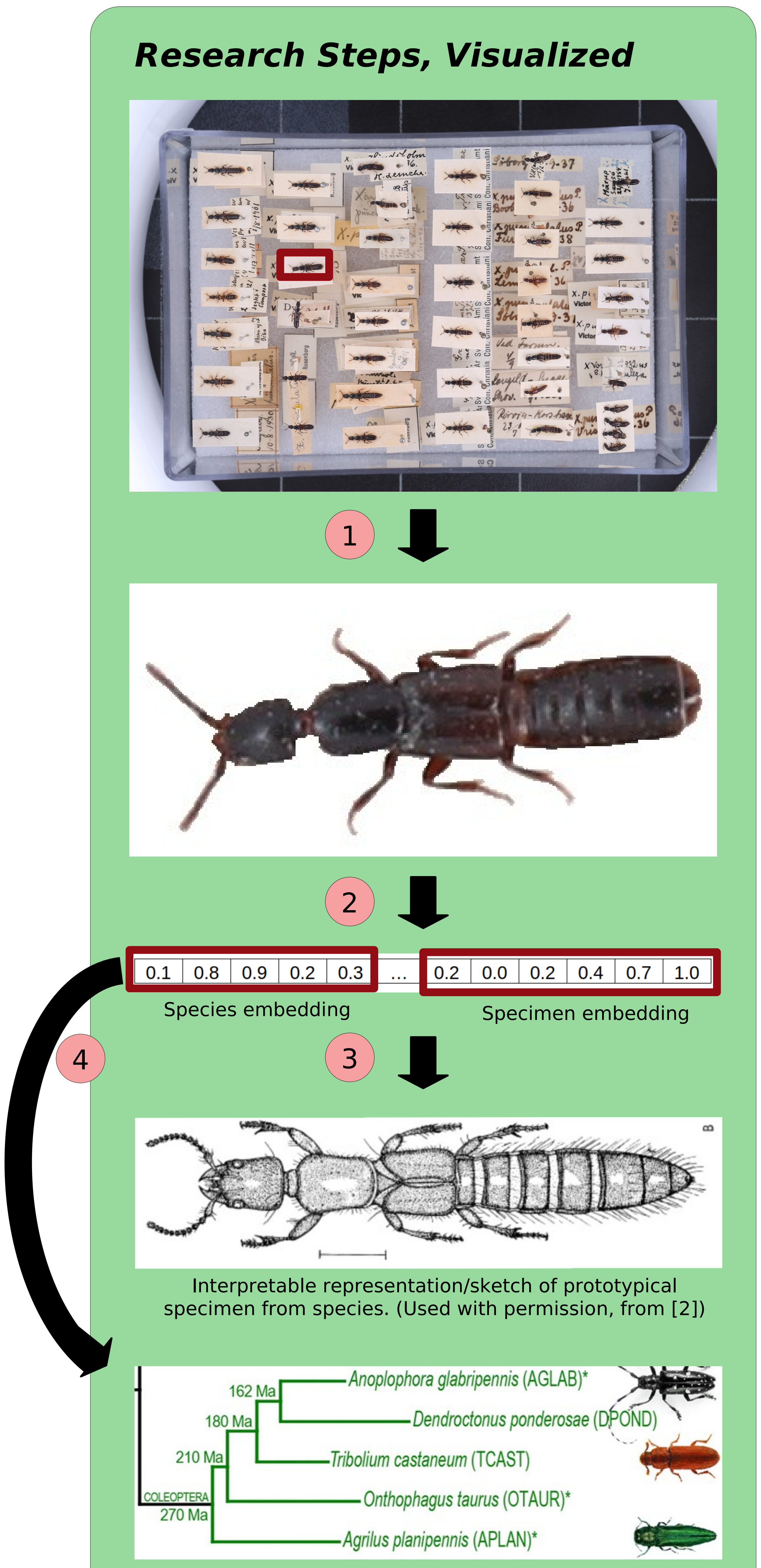
b) Should/Can we make this look like a sketch, similar to the biologists currently use?
- Could we use style transfer or similar?

4 **Generate Phylogenetic Tree**

After latent space generation, group specimens/species into heirarchical clusters

Unanswered Questions:
a) Which heirarchical clustering method will give the most accurate evolutionary representation of the data

Research Steps, Visualized


1
2
0.1
0.8
0.9
0.2
0.3
...
0.2
0.0
0.2
0.4
0.7
1.0
Species embedding
Specimen embedding
4
3
B
162 Ma
180 Ma
210 Ma
COLEOPTERA
270 Ma
Anoplophora glabripennis (AGLAB)*
Dendroctonus ponderosae (DPOND)
Tribolium castaneum (TCAST)
Onthophagus taurus (OTAUR)*
Agrilus planipennis (APLAN)*

Interpretable representation/sketch of prototypical specimen from species. (Used with permission, from [2])

Example simplified phylogenetic tree for order Coleoptera (beetles). Branches in the tree represent evolutionary distance [3]

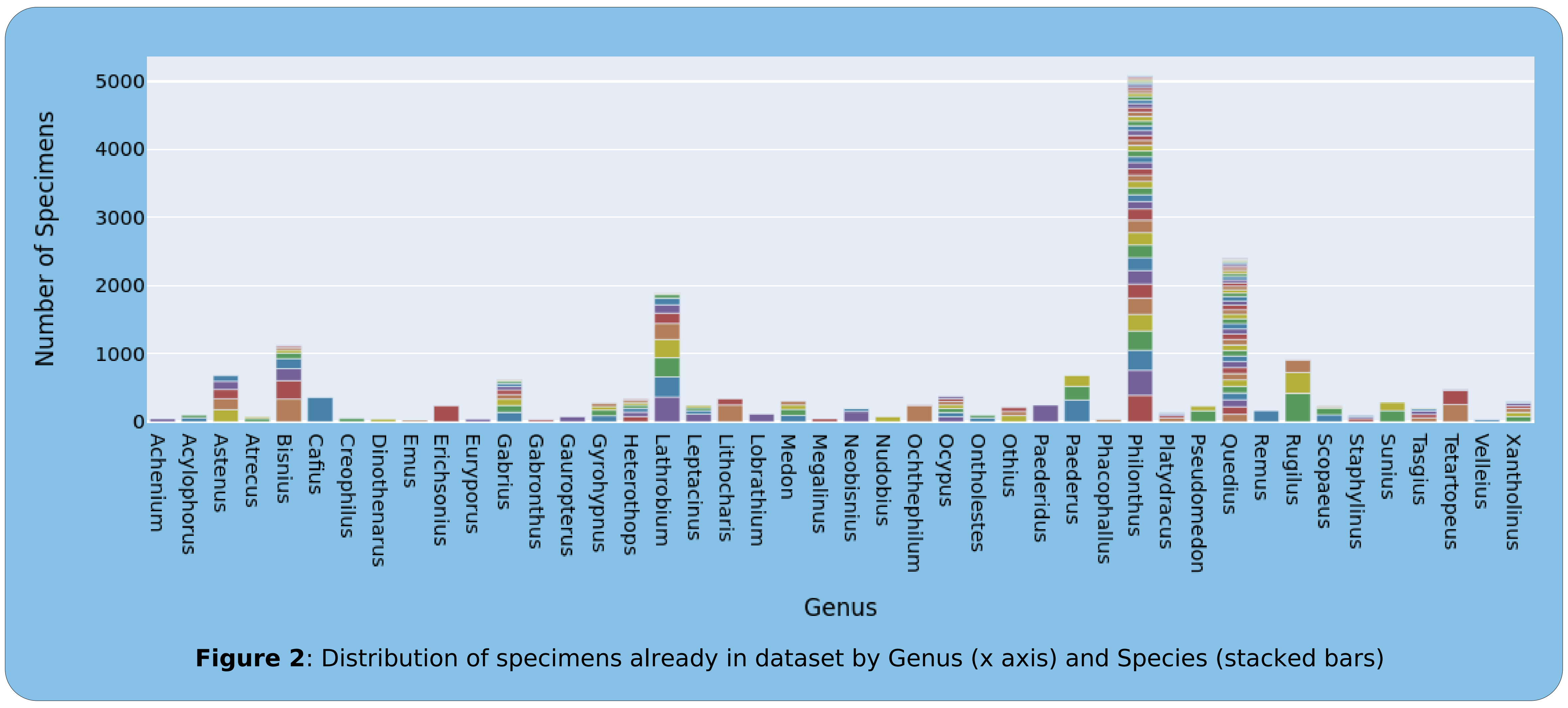


Figure 2: Distribution of specimens already in dataset by Genus (x axis) and Species (stacked bars)

References

[1] Gusarov V.I. (2018) Phylogeny of the Family Staphylinidae Based on Molecular Data: A Review. In: Betz O., Irmler U., Klimaszewski J. (eds) Biology of Rove Beetles (Staphylinidae). Springer, Cham. https://doi.org/10.1007/978-3-319-70257-5_2

[2] Systematic revision of the genera Homalolinus and Ehomalolinus (Coleoptera, Staphylinidae, Xantholinini) - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/A-B-Dorsal-view-A-Atrecus-macrocephalus-Othiini-redrawing-of-Smetana-1982-B_fig3_230444137 [accessed 3 Aug, 2021]

[3] McKenna, D.D., Scully, E.D., Pauchet, Y. et al. Genome of the Asian longhorned beetle (Anoplophora glabripennis), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle–plant interface. Genome Biol 17, 227 (2016). https://doi.org/10.1186/s13059-016-1088-8