Detecting Specimen Citations in Scientific Literature

Sara Lafia &
Andrea K. Thomer
(University of Michigan)

Challenge: How can we characterize the *impact* and *reach* of natural history collections?

Case study: University of Michigan's Museum of Zoology (UMMZ)

Approach: manual annotation, machine learning, metrics

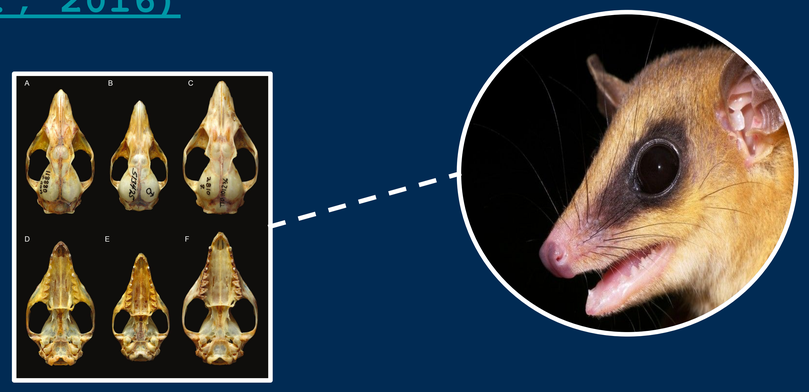
1. Curate a bibliography of literature citing UMMZ
2. Label specimen citations from selected papers
3. Train a custom named entity recognition model
4. Apply model and rule-based matching

"The additional sequence data presented herein reinforce those conclusions: we obtained 1,063 nucleotides of Cytb sequence from S. brasiliensis **UFPE 1740** (GenBank no. MH115201) and 608 nucleotides from S. andinus **UMMZ 77075** (GenBank no. MH460962)." (Ruedas et al., 2019)

Tapeti "cottontail rabbits"

Examples of specimen citation contexts extracted from papers

"The naked (scaly) caudal skin is completely dark from base to tip in some specimens (e.g., **AMNH 61382**, **USNM 513425**), but the tail of the holotype has pale mottling near the tip, and the tail of one specimen (**UMMZ 176563**) is almost half white." (Díaz-Nieto et al., 2016)



Marmosa "mouse opossums"

By the numbers...

- **794** papers published between **1910 - 2022**
- Top outlets include *Journal of Mammalogy* (**202** papers), *Proceedings of the Biological Society of Washington* (**21** papers), *American Museum Novitates* (**18** papers)
- **7%** (76 papers) collected have supplementary materials that need manual review

Next steps: promote citation guidelines (e.g., following physical samples) and extend citation network

Acknowledgements: Thank you to our collaborators: Cody Thompson[1], Ellen Cassidy[1], Faye Polasek[2] , and Katherine Polasek[2] (University of Michigan - UMMZ[1] and LSA[2]).

LSA MUSEUM OF ZOOLOGY
UNIVERSITY OF MICHIGAN

SCAN ME