Description-based species identification: creating a database for robust and interpretable models for biodiversity monitoring

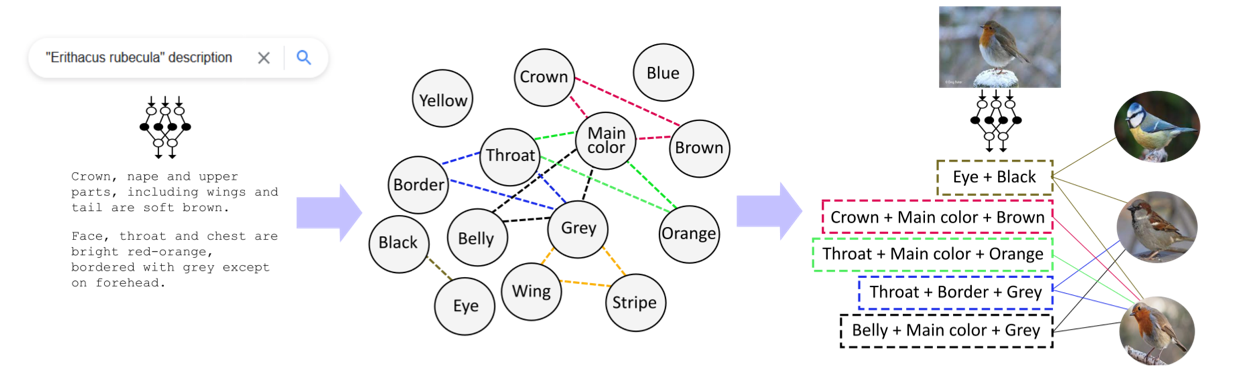


Figure This internship will focus on the left and middle figure. First a large textual database of descriptions is created, and common parts and traits are extracted, resulting in a high-order graph for each species as can be seen in the middle figure. Figure created by Diego Marcos.

Name student: Robert van de Vlasakker

Registration number: 920523897020

Period of Internship: 2023-01-01 – 2023-06-01

Date final report: 2023-06-01

Telephone number student: +3165713120

|  |  |
| --- | --- |
| Supervisor’s Name Internship provider | Diego Marcos |
| Name of Company Institution | Diego Marcos (on behalf of Inria) |
| Address | 2 Pl. Pierre Viala, 34000 Montpellier, France |
| Telephone number | - |
| MGI Supervisor’s Name Internship | Ioannis Anthasiadis |

# Introduction

## Internship organisation background

The Inria centre at Université Côte d'Azur was established in 1983. It has developed along with Université Côte d'Azur on the Sophia Antipolis and Nice sites. It also has a branch in Montpellier, where it is developing alongside Montpellier University.

Their main areas of interest are:

1. Biology-health, including medicine, neuroscience and bioinformatics
2. The links between AI, geometry, heterogeneous data and modelling.
3. Collaborative robotics for open and dynamic environments.

## Context and justification of research

With the current rate, over 50% of the species will be extinct before every being described (Lees & Pimm, 2015); biodiversity loss is understudied problem. Scalable technologies that can automatically classify species are more needed than ever (Van Horn et al., 2018). However, species classification is a very delicate field. We need to be able to track the models reasoning to improve the trustworthiness and the performance on less common species (Carvalho et al., 2019; Van Horn & Perona, 2017). By first building a large dataset that contains species and their description, we can teach a neural network to reason like a taxonomist. This way we can keep the network interpretable, and less common species can benefit from parts and traits shared with more common species.

### Significance of this topic

Although computer vision methods for plant species identification are improving at a fast pace, we cannot be sure if they are doing so in a way that is compatible with current botanical knowledge, resulting in unreliable predictions for data-poor species. In order to make sure that computer vision methods follow a process that more closely resembles what botanists do, we first need to enable them to identify the relevant morphological traits on the image. This requires the construction of a large and comprehensive dataset of morphological traits to train such a system.

In this project we aim at enabling large scale biodiversity monitoring by citizen scientists by developing Explainable Machine Learning methods that reason like a taxonomist. This will help scientists obtain valuable data from rare or undescribed species and make it easier for everyone to become an amateur naturalist, thus raising awareness about biodiversity and the rapid pace at which we are losing it. We will use Natural Language Processing (NLP) to extract relevant traits from a large textual database of species descriptions. We aim at publishing the dataset at the end of the internship.

### Research objective(s) and/or Research questions

How can relevant parts and traits be extracted from scraped sentences from text descriptions and processed in such a way resulting a higher order graph by using latest state of the art language models?

# Proposed Methodology

## Type of scientific procedure(s)

The first step is to scrape large amount of data about species. A trained model that can distinguish description and non-description text is already available. By deploying this model in a web crawler large amounts of text descriptions can be harvested.

The next step is extracting relevant information from these text snippets, this is where language models come into place. Large language models (LLMs) have already made a significant impact on AI research. These models can be used for a wide variety of downstream tasks. Recent studies have shown that natural language processing models that are used to learn linguistic features of unstructured text, also might capture relational knowledge present in the training data (AlKhamissi et al., 2022; Petroni et al., 2019). BERT (Devlin et al., 2019), is such a model that captures unlaying structured knowledge information.

More recent large language models also have to ability to understand context given and give an appropriate response, showing significant improvements over earlier models (Susnjak, 2022). More recent models are ChatGPT and BLOOM. Bloom and ChatGPT are both LLM trained models (*ChatGPT*, 2022; Scao et al., 2022). Bloom is freely available for to download to investigate the performance and its behaviour. Bloom is embedding in the Python Transformer library for easy access. It already shows remarkable with text questions and answering.

## Expected Outputs and Outcomes

A large (relative) clean graph/database with species and their traits.

## Methods / procedures followed

Provide a flowchart identifying major research phases with the required inputs and expected results. Also, relate, if possible, specific software/hardware/materials related to each phase.

# Internship information concerning internship period and provider

The internship will start on in January 2023 and the internship will end in June 2023. The internship will be remote. The working location will be Wageningen University and Research, but the internship provider is Inria. The official internship supervisor will be prof. Diego Marcos of the University of Montpellier. The supervision will mainly be done by Diego Marcos of the University of Montpellier. See the internship agreement for phone numbers, addresses and emails addresses.

# References

AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., & Ghazvininejad, M. (2022). *A Review on Language Models as Knowledge Bases* (arXiv:2204.06031). arXiv. http://arxiv.org/abs/2204.06031

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), Article 8. https://doi.org/10.3390/electronics8080832

*ChatGPT: Optimizing Language Models for Dialogue*. (2022, November 30). OpenAI. https://openai.com/blog/chatgpt/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Lees, A. C., & Pimm, S. L. (2015). Species, extinct before we know them? *Current Biology*, *25*(5), R177–R180. https://doi.org/10.1016/j.cub.2014.12.017

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). *Language Models as Knowledge Bases?* (arXiv:1909.01066). arXiv. http://arxiv.org/abs/1909.01066

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., … Wolf, T. (2022). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model* (arXiv:2211.05100). arXiv. http://arxiv.org/abs/2211.05100

Susnjak, T. (2022). *ChatGPT: The End of Online Exam Integrity?* (arXiv:2212.09292). arXiv. http://arxiv.org/abs/2212.09292

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). *The iNaturalist Species Classification and Detection Dataset* (arXiv:1707.06642). arXiv. https://doi.org/10.48550/arXiv.1707.06642

Van Horn, G., & Perona, P. (2017). The Devil is in the Tails: Fine-grained Classification in the Wild. *ArXiv:1709.01450 [Cs]*. http://arxiv.org/abs/1709.01450

# [if applicable] Appendix…