Traits for Efficient Navigation and Search in Natural History Collections

Elie M. Saliba[‡], Eric Chenin[§], Régine Vignes Lebbe[‡]

‡ Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, Muséum national d'Histoire naturelle, CNRS, EPHE, Université des Antilles, Paris, France
§ UMMISCO, Institut de Recherche pour le Développement, France Nord, Bondy, France

Corresponding author: Elie M. Saliba (elie.saliba@mnhn.fr)

Abstract

The application of AI methods is increasingly fueling research in biodiversity. One of the objectives of the French national e-COL+ project is to enable collections to benefit from the innovative contributions of image recognition and text mining.

The preceding e-ReColNat project aimed to centralize all the images and data from natural history collections on a single platform (Pérez and Pignal 2013). Despite this abundance of collection-related visual media, the options available for exploring them are currently limited to the usual metadata, such as the name of the species, or the place and date of collection. AI methods offer the promise of better usability (see Ariouat et al. 2023) by extracting characteristics linked to specimens and taxa, known as traits.

To go further, it is essential to identify some potential traits that AI models can be trained to recognize. To this end, scientists and curators with expertise in different taxa and conservation techniques were consulted. The taxonomic knowledge of the interviewees covers botany, zoology and paleontology. Their expertise encompasses different types of collections, such as fossils, thin sections, herbarium sheets, alcohol-preserved and dry specimens (Table 1).

Some of the traits mentioned are specific to individual specimens, including visible polymorphic morpho-anatomical characteristics, such as the shape of a leaf. Another possible category of traits is related to the specific preservation state of the specimen, such as early traces of pyrite rot (see Larkin 2011) in fossils. The last main category of traits at the specimen level focuses on the presence or absence of elements or organs such as traces of soil, flowers or seeds on a plant, as a way to filter relevant specimens for given studies. These traits can be efficiently extracted using computer vision models, which are trained using corpora assembled by experts.

Other traits can be deduced from species-level descriptions. These include broader characteristics than those mentioned above, such as invisible morpho-anatomy at the level of the specimen, such as the potential size of a tree. The ecology, phenology,

spatial distribution and relationships with humans were also cited. Natural language processing (NLP) artificial intelligence techniques are used to extract these traits (Sahraoui et al. 2022). There is a synergy between the two AI approaches: taxon-level traits identified through text mining can also be used to train computer vision models, improving their ability to recognize these traits in images. This link between traits and species makes it possible to automatically annotate corpora on a large scale.

The main issue that emerged during the interviews was the vocabulary. As an example, the notions of ‘toothed’ or ‘denticulate’ to describe a leaf margin are difficult to strictly differentiate. Moreover, some collections at the Muséum national d'Histoire naturelle (MNHN) need an upstream improvement of their current metadata (missing or weak taxonomic identification, database populating in progress), before AI-derived data can be implemented effectively.

In conclusion, by systematically identifying and extracting traits relevant to navigation and search from a vast array of images, the e-Col+ project enhances the usability of French collections. Collaboration between scientists, curators and AI experts ensures the robustness and usefulness of the project's outcomes, paving the way for innovative research and application.

Keywords

artificial intelligence, training set, curation, e-COL+

Presenting author

Elie M. Saliba

Presented at

SPNHC-TDWG 2024

Acknowledgements

The authors are grateful to D. Brabant, A. Kerner, A. Ohler, E. Pérez, P. Provini, I. Rouget, T. Bourgoin, F. Jabbour, M. Pignal, P. Pruvost, G. Rouhan (Muséum National d’Histoire Naturelle); C. Loup (Université de Montpellier); and N. Bailly (University of British Columbia) for answering our questions.

Funding program

This work was funded by the e-COL+ PIA (21-ESRE-0053).

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Ariouat H, Sklab Y, Pignal M, Vignes Lebbe R, Zucker J, Prifti E, Chenin E (2023) Extracting Masks from Herbarium Specimen Images Based on Object Detection and Image Segmentation Techniques. Biodiversity Information Science and Standards 7: e112161. https://doi.org/10.3897/biss.7.112161
- Larkin NR (2011) Pyrite Decay: cause and effect, prevention and cure. NatSCA News 21: 35-43.
- Pérez E, Pignal M (2013) Numériser et promouvoir les collections d'histoire naturelle . Bulletin des bibliothèques de France (BBF) 5: 27-30. URL: https://bbf.enssib.fr/consulter/bbf-2013-05-0027-006
- Sahraoui M, Pignal M, Vignes Lebbe R, Guigue V (2022) NEARSIDE: Structured kNowledge Extraction frAmework from Specles DEscriptions. Biodiversity Information Science and Standards 6: e94297. https://doi.org/10.3897/biss.6.94297

Table 1.

Table 1: Taxa and corresponding categories covered by the interviews for the e-Col+ trait project

Taxon	**Type of collection**
Botany	
Angiosperma	Herbarium sheets
Filicophyta	Herbarium sheets
Paleozoology	
Archaeocyatha	Thin sections, fossils
Ammonita	Fossils
Zoology	
Mammalia	Skeletons
Pisces	Skeletons, alcohol-preserved specimens
Amphibia	Alcohol-preserved specimens
Aves	Skeletons, dry specimens