EcoParse

Automated Species-Level Data Extraction Tool

Authors: Adam Ulicny, Florencia Grattarola, Gabriel Ortega, Ivo Kadlec, Petr Keil
MOBI lab, Czech University of Life Sciences in Prague, e-mail: ulicny@fld.czu.cz

MOBI Lab

INTRODUCTION:

The Problem:

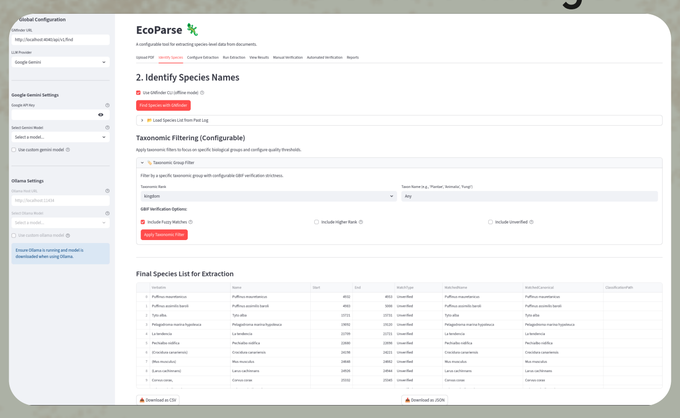
Valuable species-level data are often hidden in various sources.

For large-scale analyses, data must be extracted accurately and manually. This can take many human hours.

In our case:
Thousands of Regional Redlists
RegRed Project

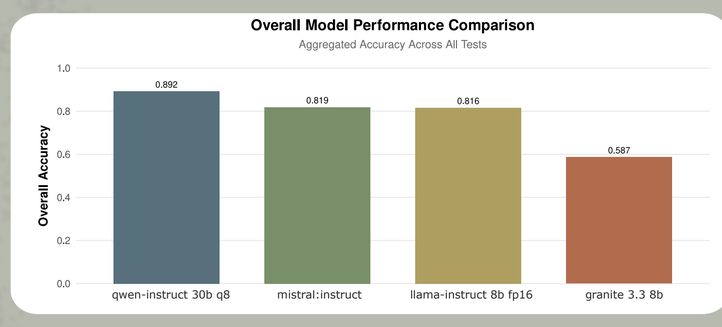
Our Solution:

A pipeline that combines NLP, strategic contextual chunking and existing LLM's into a GUI. Provided as a Docker image.



Accuracy:

We tested against 8 Redlists and compared 4 local models with identical settings between runs. We used a 20GB VRAM GPU



Best performer = Qwen-instruct
~ 89 % mean accuracy

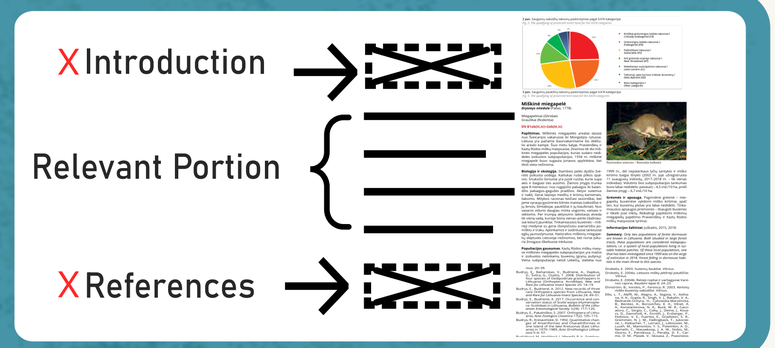
Try it out, contribute:

github.com/AdamUlicny/EcoParse

PIPELINE OVERVIEW:

File Preparation

Trim irrelevant sections.
This step focuses the extraction and reduces all API calls.
Fewer FP's and tokens used.



Species names

GNFinder Locates potential latin names in text.

GBIF API filtering further focuses extraction on relevant taxons.

Initial latin name candidates:

- Actualmente esta zona
- Alnus acuminata
- Ampelion rufaxilla
- Andes orientales
- Andigena hypoglauca
- Andigena laminirostris
- Andigena nigrirostris
- Anisognathus notabilis
- Antidaphne andina
- Ara militaris

Class: Aves

Validation + Taxonomic filtering

Final filtered species list:

- Ampelion rufaxilla
- Andigena hypoglauca
- Andigena laminirostris
- Andigena nigrirostris
- Anisognathus notabilis
- Ara militaris

Examples

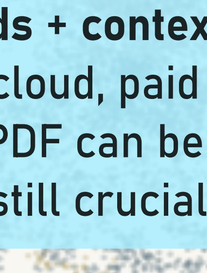
For better results, we guide the LLM by providing examples.

Few-shot prompting technique.

Vuursalamander (*Salamandra salamandra ssp. terrestris*)

Rode Lijst 2023: **ernstig bedreigd**
Rode Lijst 2007: bedreigd
IUCN Nederland 2023: critically endangered

Nederlandse criteria
Zeldzaamheid: de populatiegrootte wordt tussen de 50 en 250 volwassen dieren geschat en de soort plant zich voort in 4 atlasblokken, wat in beide gevallen leidt tot zeldzaamheidsklasse zeer zeldzaam (zzz).
Trend sinds 1950: de verspreiding is afgenomen met 50% wat leidt tot trendklasse sterk afgenomen (tt). De populatiegrootte is afgenomen met 99%, wat leidt tot de zwaardere trendklasse zeer sterk afgenomen (ttt).



Input: "IUCN Nederland 2023: critically endangered"

Output: CR

Explainer: Only IUCN 2023 data

Chunking

For each species, contextual chunks are created.
Text chunks are quicker and cheaper to process.
Image preserves structure at a higher cost (tokens).

Text

<Pardela Chica
Puffinus assimilis baroli
En Peligro; **EN B2ab (i,ii,iii); C2a(ii)**
Autores: Domingo Trujillo y Juan José Ramos
En el archipiélago canario está presente la subespecie P. a. baroli, habiéndose constatado su nidificación sólo en Alegranza, Montaña Clara>

Image

LLM parsing

For each species in the filtered list, a separate API call is created.
This includes: **base prompt + custom data fields + context chunks**
EcoParse supports **Ollama** (local) or **Gemini** (cloud, paid tier).
With some experience, data from a 500 page PDF can be extracted in less than 5 minutes. Manual verification is still crucial for perfect results.

Download the data!

Sample raw output from EcoParse:

Species	Status	Criteria
Gavia immer	VU	D1
Podiceps nigricollis	NT	VU D1
Podiceps cristatus	NF	NF
Bulweria bulwerii	EN	B2ab(ii,iii,iv)
Asio otus	NF	NF
Larus cachinnans	NF	NF
Calonectris diomedea diomedea	EN	A3cde
Bubo bubo	NF	NF
Calonectris diomedea borealis	VU	A3d+4d
Puffinus puffinus	EN	B2ab(ii,iii); C2a(ii)
Puffinus mauretanicus	CR	A3ace+4ace; B2ab(ii,iii,iv,v); E
Balear nidifica	NF	NF
Puffinus assimilis baroli	EN	B2ab(i,ii,iii); C2a(ii)
Pelagodroma marina hypoleuca	VU	NF
Pechialbo nidifica	NF	NF

Funded by the European Union

European Research Council
Established by the European Commission

BEAST Project

Funded by the European Union (ERC, BEAST, 101044740).
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency.
Neither the European Union nor the granting authority can be held responsible for them.