

EthnoBot: A Semantic Platform for Ethnobotanical Data Collection and LLM-Aided Plant Name Resolution

Theodora Manika^{2,*}, Michail Tsenos^{2,8}, Christian Ebeling⁴, Vijay Ingalalli⁵,
Antigoni Cheilari^{1,2}, Evanthia Dina^{1,2}, Avni Hajdari⁶, Andrea Pieroni^{6,7},
Dimitrios Gunopulos^{2,3}, Marc Jacobs⁴ and Nektarios Aligiannis^{1,2}

¹Laboratory of Pharmacognosy and Natural Products Chemistry, Faculty of Pharmacy, School of Health Sciences, National and Kapodistrian University of Athens, 15771 Athens, Greece

²Institute for Ethnopharmacological Studies and Phytotherapy, National and Kapodistrian University of Athens, 15771 Athens, Greece

³Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece

⁴Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, 53757, Germany

⁵PangeAI, Pangea Botanica GmbH, Hardenbergstraße 32, 10623 Berlin, Germany

⁶Department of Biology, Faculty of Mathematical and Natural Science, University of Prishtina, Mother Theresa St, Prishtinë, 10000, Republic of Kosovo

⁷University of Gastronomic Sciences, Piazza Vittorio Emanuele II 9, I-12042 Pollenzo, Italy

⁸Dept. of Informatics, Athens University of Economics and Business, 10434 Athens, Greece

Abstract

Ethnobotanical fieldwork is typically recorded in spreadsheets, with multilingual vernacular plant names harmonised against scientific taxonomy only at publication time — a manual, error-prone step that limits reproducibility and reuse. We present EthnoBot, an open semantic platform combining structured field data collection with a two-stage, LLM-assisted plant name resolution pipeline, linking observations to authoritative biodiversity resources (WFO, IPNI, IUCN Red List) as a Darwin Core-aligned RDF graph. On a $N = 200$ Himara reference dataset, using a field-session confirmation simulation, the correct species appeared in the top-5 candidates 50.5% of the time, well ahead of a zero-LLM fuzzy-matching baseline on names in an unseen language, with all final mappings expert-validated. EthnoBot further computes ethnobotanical indices and supports interactive analytics and geospatial visualisation. A live demonstration is available at: <https://ethnobot.ethcstwin.di.uoa.gr:65000>.

Keywords

ethnobotany, semantic web, knowledge graphs, Darwin Core, plant name resolution, large language models, linked data, biodiversity informatics

1. Introduction

Ethnobotanical fieldwork is commonly recorded in spreadsheets, with taxonomic harmonisation deferred to data cleaning and publication — a manual, hard-to-reproduce step. The central challenge is reconciling multilingual vernacular names, which vary across languages and communities, with the stable scientific identifiers needed for linkage to global biodiversity knowledge bases.

We present **EthnoBot**, an open semantic platform that resolves multilingual vernacular names to WFO [1] identifiers and IUCN Red List [2] metadata and exposes the result as queryable linked data, detailed in the sections that follow.

2. Related Work

Existing biodiversity platforms such as GBIF and TRY [3] support specimen, occurrence, or trait management but not multilingual vernacular name reconciliation. The Taxonomic Name Resolution

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

*Corresponding author.

✉ theodoramn@gmail.com (T. Manika)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Service [4] and the World Checklist of Vascular Plants [5] resolve scientific-name synonymy but not vernacular names — the central challenge here. A related line of work, including SPECIES and ORGANISMS [6], TaxoNERD [7], and LLM-distilled NER for endangered species [8], extracts taxonomic entities from already-written text; EthnoBot instead resolves a single, context-free, orally-transmitted vernacular name at data-collection time into a ranked candidate list for on-the-spot expert confirmation, and further integrates LLM-assisted resolution with persistent-identifier enrichment (WFO, IPNI, IUCN) in one collaborative platform exposed as a Darwin Core-aligned RDF graph.

3. System Architecture

EthnoBot consists of three independently deployable components:

Backend API A Python/FastAPI service backed by MySQL, managing the core domain entities and integrating with the IUCN Red List, WFO, and Nominatim¹/OpenStreetMap for enrichment and geocoding. JWT authentication and collaborator roles enforce access control.

Plant Mapper Microservice A dedicated FastAPI service (Section 4) for plant name resolution, decoupled from the backend for independent scaling. A provider abstraction layer supports Google Gemini, Groq-hosted Llama models, and local Ollama [9] deployments, with translation and mapping caches in SQLite or MySQL.

Frontend A React 19/TypeScript application built with Vite, offering a data entry wizard, a spreadsheet view with inline validation, Leaflet-based interactive maps, Recharts analytics dashboards, and a plant species gallery annotated with IUCN conservation status.

4. Plant Name Resolution

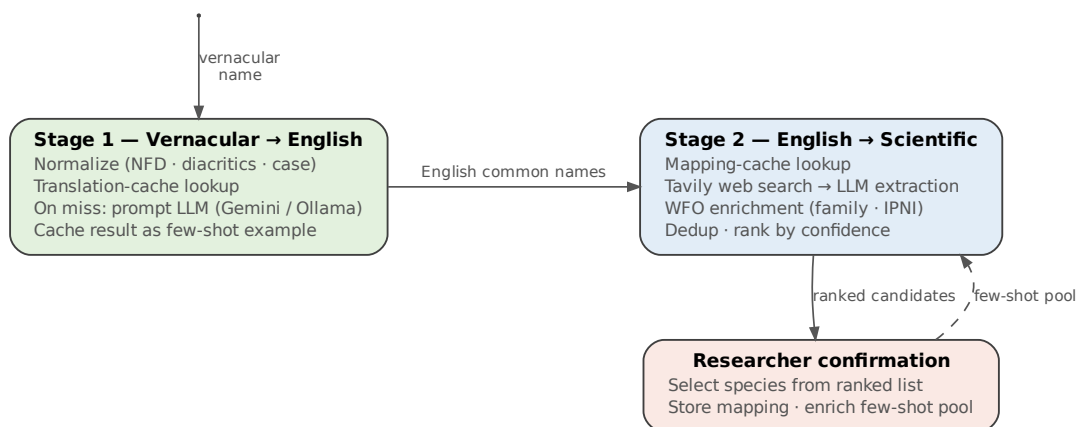


Figure 1: Two-stage pipeline. Stage 1 maps a vernacular name to English common names via LLM or cache; Stage 2 maps English names to ranked scientific candidates via Tavily search and LLM inference, then WFO API enrichment. Confirmed mappings are reused as few-shot examples.

The pipeline (Figure 1) treats name reconciliation as two-step entity linking: a language-neutral English pivot, then authoritative scientific identifiers.

¹<https://nominatim.org/>

Stage 1 – Vernacular to English. Input names are normalised and matched against a persistent cache. On cache misses, Google Gemini [10] generates English common-name candidates from multilingual inputs; confirmed mappings are persisted and reused as context-aware few-shot examples specific to this deployment’s own dialect, following the prompting approach of [11] for low-resource dialect translation, rather than generic cross-language examples. The Stage 1 prompt asks the model to reason etymologically before answering, e.g.:

System: You are a botanical linguist and ethnobotanist, tracing vernacular and folk plant names across languages and dialects to their correct English common names.

Identify the English common name(s) for: "{local_name}"
Reason step by step: language/region -> literal meaning or linguistic roots -> English plant name. For example:
"Dafni" -> Greek dafne (daphne) = laurel -> ["bay laurel", "bay leaf"]
Output a single line: a JSON array of >3 English common names, most likely first.

Stage 2 – English to Scientific. Each candidate English name is submitted to Tavily [12] web search; an LLM extracts ranked scientific candidates with confidence scores, falling back to a structured Gemini prompt on search failure. Candidates are enriched via the WFO API [1] (author-qualified name, family, IPNI ID), deduplicated, and ranked. The researcher’s selection feeds the few-shot pool for subsequent calls.

Both caches also perform a fuzzy pre-check (Jaro–Winkler, threshold 0.92) ahead of the LLM calls, preferring human-confirmed matches, so a dialectal spelling variant of an already-cached name is served from cache instead of re-querying the LLM. Our generalisation analysis (Section 4.1) shows this resolves genuinely novel same-language spellings correctly 46–57% of the time (top-1/5) from cache alone.

4.1. Evaluation

We built a reproducible evaluation harness around a $N = 200$ reference set from the Himara raw fieldwork spreadsheet (stratified by ethnic group, 124 Greek / 76 Albanian), resolved against a disposable, unseeded database. The baseline and generalisation results below are cold (order-independent); the pipeline result further below instead simulates confirmed suggestions accumulating over a field session (order-dependent, a session learning-curve measurement).

Baseline comparison. A zero-LLM fuzzy-matching baseline (Jaro–Winkler [13]) scores 41.5%/50.5%/52.5% (top-1/3/5) under random leave-one-out — already exceeding the pipeline’s originally reported numbers, but inflated by dialectal near-duplicates across neighbouring villages. On a genuinely unseen village it holds up moderately (46.4%/55.9%/57.0%); with *zero* target-language entries in the dictionary, simulating an unseen language, it collapses to 16%/23%/24% — the gap an LLM-assisted approach is intended to close.

Pipeline results. Using the fuzzy-cache pre-check with simulated confirmations (session learning-curve, see above) and Stage 1 few-shot examples drawn from this deployment’s own confirmed translations rather than generic cross-language examples, the pipeline achieved 41.0%/48.0%/50.5% (top-1/3/5) overall; Greek 46.0%/52.4%/54.8%, Albanian 32.9%/40.8%/43.4%. Stratified by serving provider (almost entirely prepaid Gemini; one item’s calls fell back to Groq): Gemini-only 41.2%/48.2%/50.8% ($N = 199$), Groq fallback ($N = 1$, too small to be conclusive).

Of the 97 misses (classified by whether any candidate shares the expected genus), 88.7% are *Stage 1 failures* — wrong English pivot entirely (e.g. *Rodees* rendered as *Rosa canina* rather than *Punica granatum*) — and 11.3% are *Stage 2 failures* — plausible pivot, wrong species (e.g. *Thymari* mapped to *Thymus vulgaris* instead of *T. capitatus*).

Confidence calibration. Pooling every candidate’s self-reported confidence against correctness ($N = 1040$ candidates) gives an Expected Calibration Error of 0.776 and a Brier score of 0.713; candidates with confidence ≥ 0.95 are correct 18.1% of the time versus 12.0% at confidence ≥ 0.5 , indicating

self-reported confidence remains far from reliable in this run despite improving over an earlier iteration. All accepted mappings are in any case reviewed by domain experts; calibration guides *how much* scrutiny a suggestion warrants, not whether to review it.

5. Platform Features

5.1. Structured Data Collection and Collaboration

Research is organised around *Studies*, each containing *Questionnaires* and *Participants* with demographic metadata. Owners share access via invitation tokens; collaborators record observations through a draft → final workflow, after which finalised records may undergo expert review and enrichment. Use categories, plant parts, preparation methods, and administration routes draw on a standardised ethnobotanical terminology, enabling consistent cross-study comparison.

5.2. Knowledge Graph Construction and SPARQL Access

Upon confirmation of a scientific name, the backend asynchronously enriches the record with (i) the WFO identifier and plant family, (ii) the IPNI identifier, and (iii) the IUCN Red List category. Each confirmed observation is materialised as a Darwin Core-aligned RDF description, exposed as a Turtle dump and queryable through the platform's SPARQL endpoint². Observations are minted as `dwc:Occurrence` resources linked to their WFO taxon URI via `dwc:taxonID`; vernacular names are emitted as BCP-47 language-tagged literals (`@el` Greek, `@sq` Albanian):

```
@prefix dwc: <http://rs.tdwg.org/dwc/terms/> .
@prefix wfo: <https://list.worldfloraonline.org/> .
@prefix eb: <https://ethnobot.ethcstwin.di.uoa.gr/ns#> .
@prefix obs: <https://ethnobot.ethcstwin.di.uoa.gr/obs/> .

obs:1003 a dwc:Occurrence ;
  dwc:scientificName "Sideritis raeseri Boiss. & Heldr." ;
  dwc:taxonID wfo:wfo-0000291547 ;
  dwc:family "Lamiaceae" ;
  dwc:vernacularName "caj mali"@sq ;
  eb:iucnRedListCategory "VU" ;
  eb:useCategory "respiratory" .
```

The `eb:` namespace covers ethnobotanical attributes not in Darwin Core (part used, preparation mode, use category, administration route, IUCN category). A representative SPARQL query — retrieving threatened or near-threatened species with their vernacular forms and resolvable WFO URIs, filtering `eb:iucnRedListCategory` against {VU, NT, EN, CR} — runs directly against the endpoint above.

The Himara graph currently holds 1,296 observations spanning 179 species across 11 villages; at roughly 18 triples each, its persistent URIs enable federation with external biodiversity endpoints.

5.3. Analytics, Import, and Export

EthnoBot computes standard ethnobotanical indices — Use Value (UV), Fidelity Level (FL), and Relative Importance (RI) — following established methodologies [14, 15, 16]. Dashboards display plant family, use category, and plant part distributions alongside participant demographics, and a Leaflet choropleth renders observation counts per village. A bulk import pipeline parses Excel workbooks, fuzzy-matches free-text to the standardised terminology, geocodes villages, and inserts validated records in batch; full datasets export as Excel, CSV, or Darwin Core-aligned Turtle.

²<https://ethnobot-backend.ethcstwin.di.uoa.gr:65000/docs/>

6. Demonstration

Use case. Ioanna, an ethnobotanist surveying a Himara village, records a participant naming a local remedy, *dafni*, in Greek. Rather than leaving the scientific name for post-hoc cleanup, she enters the vernacular term; EthnoBot proposes ranked candidates and she confirms *Laurus nobilis* L., which is immediately enriched with its WFO and IPNI identifiers and Vulnerable IUCN status and added to the knowledge graph — turning a raw field note into a linked, queryable record on the spot.

The demonstration uses the bilingual Greek-Albanian Himara dataset. We demonstrate: (1) real-time name suggestions during data entry; (2) AI-assisted resolution of Greek and Albanian vernacular names to WFO-validated identifiers with automatic IUCN enrichment; (3) Turtle export and SPARQL querying of confirmed observations; (4) analytics dashboards computing UV, FL, and RI; and (5) interactive geospatial visualisation across Himara villages. Visitors who register begin with an empty study and enter their own observations through the same workflow. The live demo is publicly available at <https://ethnobot.ethcstwin.di.uoa.gr:65000>.

6.1. User Evaluation

We conducted a preliminary usability study with $N = 10$ participants (ethnobotany researchers and computer science practitioners) using a think-aloud protocol and the System Usability Scale [17], obtaining a mean score of 76.5 ($SD = 11.2$), indicating good perceived usability. Study conditions differed by participant group: ethnobotanist participants used the platform during or immediately after real Himara fieldwork sessions, while computer-science practitioners evaluated it in desk-based sessions without field conditions. This heterogeneity limits direct comparability between the two groups’ scores and is a limitation of the present study rather than a controlled between-group design. Scores were lower among non-technical participants, and the main concern across both groups was uncertainty about when LLM-generated candidates could be trusted without expert confirmation — a concern the calibration analysis in Section 4.1 is intended to help address.

7. Conclusion and Future Work

EthnoBot shows how structured ethnobotanical field data, LLM-assisted entity linking, and biodiversity knowledge bases together produce research-ready datasets linked to global knowledge graphs via persistent WFO, IPNI, and IUCN identifiers. Current limitations include the residual risk of LLM hallucination in name suggestion and the scalability of the RDFLib backend. Expert confirmation remains mandatory not only as a safeguard against such model error but because vernacular names are themselves intrinsically ambiguous within folk taxonomies: informants in our own dataset used the same local name, *Radikia*, for both *Borago* sp. and *Taraxacum officinale* — two unrelated wild greens from different families that a system resolving names in isolation, however accurate, cannot disambiguate without the collector’s first-hand knowledge of which plant was actually meant. Planned work includes migration from RDFLib to a dedicated triple-store (e.g., Apache Fuseki or GraphDB) for federated querying and reasoning, a formal RDFS/OWL definition of the `eb:` namespace together with SHACL/ShEx validation shapes to make the ethnobotanical extension machine-checkable, and a mobile-first offline field entry mode.

Acknowledgments

The authors thank the communities of the Himara region for their participation in the ethnobotanical surveys underlying the demonstration dataset. This work was funded under the European HORIZON-WIDERA-2023-Twinning Bottom-Up project “ETHCSTWIN: Reinforcing the Scientific Excellence of the Institute for Ethnopharmacological Studies and Phytotherapy (IESP) through Twinning with Computer Science (CS)”, id: 101159928.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: Drafting content, Improve writing style, Content enhancement. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] World Flora Online Consortium, World flora online, <https://wfoplantlist.org>, 2024. Accessed: April 2026.
- [2] IUCN, The IUCN red list of threatened species, <https://www.iucnredlist.org>, 2024. Version 2024-1. Accessed: April 2026.
- [3] J. Kattge, G. Bönisch, S. Díaz, et al., TRY plant trait database — enhanced coverage and open access, *Global Change Biology* 26 (2020) 119–188. doi:10.1111/gcb.14904.
- [4] B. Boyle, N. Hopkins, Z. Lu, J. A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, M. L. Narro, W. H. Piel, S. J. McKay, S. Lowry, C. Freeland, R. K. Peet, B. J. Enquist, The taxonomic name resolution service: an online tool for automated standardization of plant names, *BMC Bioinformatics* 14 (2013) 16. doi:10.1186/1471-2105-14-16.
- [5] R. Govaerts, E. Nic Lughadha, N. Black, R. Turner, A. Paton, The World Checklist of Vascular Plants, a continuously updated resource for exploring global plant diversity, *Scientific Data* 8 (2021) 215. doi:10.1038/s41597-021-00997-6.
- [6] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, L. J. Jensen, The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text, *PLOS ONE* 8 (2013) e65390. doi:10.1371/journal.pone.0065390.
- [7] N. Le Guillarme, W. Thuiller, TaxoNERD: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature, *Methods in Ecology and Evolution* 13 (2022) 625–641. doi:10.1111/2041-210X.13778.
- [8] J. Atuhurra, S. C. Dujohn, H. Kamigaito, H. Shindo, T. Watanabe, Distilling named entity recognition models for endangered species from large language models, <https://arxiv.org/abs/2403.15430>, 2024. ArXiv:2403.15430.
- [9] Ollama, Ollama: Get up and running with large language models locally, <https://ollama.com>, 2024.
- [10] Google DeepMind, Gemini: A family of highly capable multimodal models, <https://arxiv.org/abs/2312.11805>, 2024. ArXiv:2312.11805.
- [11] T. T. Prama, C. M. Danforth, P. S. Dodds, LLMs for low-resource dialect translation using context-aware prompting: A case study on sylheti, <https://arxiv.org/abs/2511.21761>, 2025. ArXiv:2511.21761.
- [12] Tavily, Tavily documentation, 2026. URL: <https://docs.tavily.com>, accessed: 2026-05-26.
- [13] J. Turk, Jellyfish: Python library for approximate and phonetic matching of strings, <https://jamesturk.github.io/jellyfish/>, 2024.
- [14] O. Phillips, A. H. Gentry, Some quantitative methods for analyzing ethnobotanical knowledge, *Journal of Ethnobiology* 14 (1994) 83–109.
- [15] J. Friedman, Z. Yaniv, A. Dafni, D. Palewitch, A preliminary classification of the healing potential of medicinal plants, based on a rational analysis of an ethnopharmacological field survey among Bedouins in the Negev desert, Israel, *Journal of Ethnopharmacology* 16 (1986) 275–287. doi:10.1016/0378-8741(86)90094-0.
- [16] J. Tardío, M. Pardo-de Santayana, Cultural importance indices: a comparative analysis based on the useful wild plants of Southern Cantabria (Northern Spain), *Economic Botany* 62 (2008) 24–39. doi:10.1007/s12231-007-9004-5.
- [17] J. Brooke, SUS: A “quick and dirty” usability scale, Taylor & Francis, London, 1996, pp. 189–194.