

Reclaiming Biodiversity Knowledge: Mapping Indigenous and Common Names in Latin American Biodiversity Texts

Examples:

- <https://www.biodiversitylibrary.org/page/45264025#page/179/mode/1up>
- <https://www.biodiversitylibrary.org/item/335397#page/1/mode/1up>
- https://upload.wikimedia.org/wikipedia/commons/5/5e/Diccionario_de_botanica_brasileira%3B_ou%2C_Compendio_dos_vegetaes_do_Brasil%2C_tanto_indigenas_como_acclimados%3B_%28IA_diccionariodebot00alme%29.pdf
- https://commons.wikimedia.org/wiki/Category:Les_bois_indig%C3%A8nes_de_S%C3%A3o_Paulo
- <https://www.biodiversitylibrary.org/item/176782#page/33/mode/1up>

Project Vision:

This project aims to **decolonize biodiversity narratives** by repurposing naturalist texts, identifying, extracting, and integrating **indigenous and culturally significant common names** of species from historical biodiversity texts—particularly from the **Biodiversity Heritage Library (BHL)**—into Wikimedia projects (Wikisource, Wikidata, and Wikipedia). By focusing on **Latin America**, we address regions where biodiversity knowledge is both rich and historically marginalized in global knowledge systems.

Core Objectives:

1. Identify Historical Texts Rich in Indigenous and Common Names

1. Develop methods (AI, heuristics, linguistic markers) to detect Commons-compatible high-potential texts in BHL and similar repositories.
2. Prioritize texts from colonial periods, ethnobotanical studies, missionary accounts, and early scientific expeditions known for referencing indigenous knowledge.
3. Provide a curated set of high-value works for Wikipedia, Wikidata and Wikisource in relation to indigenous languages
4. **Develop Tech Tools to Empower Volunteers**
5. **Consult with volunteers in the BHL-Wiki Working Group**
6. **Consult with key indigenous stakeholders, including**
https://pt.wikipedia.org/wiki/Altaci_Kokama and others
 - Help of <https://www.linkedin.com/in/carlo-brescia-74169a/> ?
 - Maybe get to Krenak and Selvagem?
7. Build lightweight, tools for volunteers to engage in unlocking the **indigenous names** in historical biodiversity texts (BHL and beyond).

- Create **gap analysis dashboards** to help volunteers prioritize underrepresented species, regions, and languages.
 - Integrate **AI-assisted workflows** to support volunteers in data validation and decision-making.
 - 8. **Curate High-Potential Texts for Indigenous Knowledge Extraction**
 - Develop **automated methods** to identify historical texts likely to contain indigenous names.
 - Engage volunteers in a **curated review process**, focusing on texts from South America's diverse ecological and linguistic regions.
 - 9. **Foster a Community of Practice**
 - Provide training, documentation, and **mentorship programs** to support volunteers in using the tools effectively.
 - Collaborate with **Wikimedia affiliates**, indigenous language communities, and biodiversity researchers.
-

Key Innovations

- **AI-Assisted Indigenous Name Detection**
 - Leverage **Natural Language Processing (NLP)** to detect indigenous and common names in digitized texts.
 - Incorporate **multilingual support**, focusing on South American indigenous languages (e.g., Quechua, Guarani, Aymara).
 - **Decolonial Knowledge Gap Dashboard**
 - A web-based dashboard that helps volunteers **visualize content gaps** in biodiversity data across Wikidata, Wikipedia, and Commons.
 - Integrates with **SPARQL endpoints** to provide real-time insights on where indigenous knowledge is underrepresented.
 - **Volunteer-Centric Workflow Design**
 - Build lightweight tools for **batch editing**, **crowdsourced validation**, and **automated data reconciliation** with Wikidata.
 - Include **“impact tracking” features** to help volunteers see the results of their contributions.
-

Project Phases & Deliverables

Phase 1: Discovery & Community Co-Design (Months 1–2)

- **Activities:**
 - Conduct interviews/workshops with Wikimedia volunteers, biodiversity experts, and indigenous language advocates.
 - Define technical requirements based on community feedback.

- **Deliverables:**
 - User stories and design prototypes for tools and dashboards.
 - Curated list of historical texts likely to contain indigenous names.

Phase 2: Tool Development & Text Analysis (Months 3–5)

- **Activities:**
 - Develop and deploy:
 - **Indigenous Name Detector:** AI-powered NLP tool for identifying potential indigenous/common names in BHL texts.
 - **Decolonial Knowledge Gap Dashboard:** Visualizes biodiversity content gaps in Wikimedia projects.
 - Pilot tools with volunteer communities and iterate based on feedback.
- **Deliverables:**
 - MVP of detection tools and dashboard.
 - Annotated datasets of indigenous names for volunteer review.

Phase 3: Community Activation & Knowledge Integration (Months 6–8)

- **Activities:**
 - Organize **online workshops** to train volunteers in using tools.
 - Facilitate **edit-a-thons** and **content campaigns** to integrate indigenous names into Wikidata and Wikipedia.
- **Deliverables:**
 - 3–4 community workshops with documented best practices.
 - Final report and toolkit for ongoing volunteer-driven efforts.

Key Technologies & Tools

- **NLP & AI:**
 - **SpaCy** + custom NER models for indigenous name detection.
 - **Tesseract OCR** for historical text digitization.
 - **Pre-trained multilingual models** (e.g., Hugging Face transformers) for language processing.
- **Data Visualization:**
 - **D3.js** or **Plotly** for interactive dashboards.
 - **SPARQL endpoints** integrated with Wikidata for real-time data insights.
- **Volunteer Tools:**
 - Browser-based tools for **batch editing** and **data reconciliation**.
 - **Impact tracker widgets** to show volunteers the effect of their contributions.