Two customized queries to search for Puerto Rican plants by common name, using the Google Programmable Search Engine and the Wikidata Query Service

Mazen El Makkouk

Prof. Carlos Suarez

CINF 6807

12/12/2022

Overview

My project was to create a custom search engine for Puerto Rican plants that would be part of a subject guide for botany at a Puerto Rican university. There were two things that I wanted it to do: provide access to information for a general audience, and be an entry point to a rich variety of information, relevant to botany proper as well as to conservation and cultural (ethnobotanical) knowledge.

Because of this desire to serve a general audience, my first objective for the search engine was that it would work by entering only the common name of a plant.

I was interested in learning how to make use of two resources towards this end: the Google customizable search engine, and Wikidata.

Google programmable search engine

As a sample of names of plants to test my search engine and develop it, I started with a few plants from *Frutas Olvidadas de Puerto Rico* (Mari Mut, 2014). I wanted plants that were relatively common (not rare or endangered or even endemic) but which would be of special cultural interest for Puerto Ricans—something you could find relatively easily growing wild, and which people might have eaten or heard about. To test my searches and develop them, I used the common names of these plants, which were the following:

Uva de playa (Coccoloba uvifera)

Algarrobo (Hymenea courabil)

Jacana (Pouteria multiflora)

Jagua (Genipa americana)

Maricao (Byrsonima spicata)

Maya (Bromelia pinguin)

I used Wikidata and/or Wikipedia to find the unique identifiers for each plant, which were links to their entries in international and authoritative plant databases, such as *Plants of the World Online* (POWO) and *Tropicos*. The Google programmable search engine allows you to limit searches to up to ten websites. However, using the addresses of 10 of these plant databases and running a search for a common name would turn up no results for a search using a common name. I discovered that these databases, at least the biggest ones such as POWO and Tropicos, did not support a common name search, even directly from their websites.

I needed to find other websites. To search the whole internet yet limit the search results and help keep them relevant, I set my search engine to the whole internet again, but limited the result to pages about plants, using Knowledge Graph entities from Schema.org, a feature available in the Programmable search engine. Because many Taino names of plants in Puerto Rico can also refer to features like rivers and towns, and names of African origin can refer to animals in Africa, this was a very helpful feature.

The next step was a matter of trial and error, where I looked at promising websites in terms of quality. I looked for at least one of these criteria, but preferably more than one, in the results found:

1. Does the site present accurate taxonomic information, and other basic descriptive and contextual information such as the qualities of the plant (including pictures) and its habitat, range, etc. (including maps), plus references to herbaria, etc.

2. Does the site present pertinent ecological information, such as place in the ecosystem (eg. what animals does it support) and its current conservation status (endangered, threatened, etc). To improve the engine against this criterion, I tested the engine was against a further test sample of plants drawn from Rare and Endangered Plants of Puerto Rico (Woodbury, 1975), which were the following:

Cobana Negra (Stahlia monosperma)

Plumeria portoricensis (Aleli cimarron)

Te (Ilex cookie)

Cotorilla (Heliotropium guanicanse)

Pinon (Tillandsia lineatispica)

Higo chumbo (Harrisia portoricensis)

3. Does the site present pertinent ethnobotanical information, historical and/or current? To develop a good list of sites, I used a further list of plants used by the Tainos, where I prepared a list of plants based on a project to educate the public about plants in Taino culture prepared for the Caguas Botanical Garden (Caras et al, 2010). The sample was the following:

Ausubo/balata (Manilkara bidentata)

Sarobei/algodon (Gossypium hirsutum)

Chambibe (Sapindus Saponaria)

Achiote/Bija (Bixa ordeña)

When I had found a list of 10 good sites based on these criteria, I limited the search again to only these sites, keeping as well the limit to results only about plants.

Wikidata

When I started working on this project, I knew nothing about how to query Wikidata. The project impelled me to start learning how to write a SPARQL query: I though it would be a good impetus to my learning that I needed the results. My first results were disappointing, however. A search for items with the property "endemic to" and the value "Puerto Rico," turned up only three results (none of which was a plant, and one of which was the Chupacabra. Run that query here.). Further digging showed that not many plants have a geographic value on Wikidata, not even better known endemic ones. (I tried the edelweiss as a test, which only grows in the Alps, and a type of cedar that grows only in Cyprus—neither had a geographical value on Wikidata).

What Wikidata was strong on was identifiers. I got the idea of searching for all items with a POWO identifier, and then filtering that result in a way relevant to Puerto Rico. Searching for all items with a POWO identifier worked, but the search timed out because the results were too large (there is a 60 second limit). However, when I combined that query with an additional filter, to show only those results that had a POWO identifier *and* a common name in Spanish, I got a manageable number: 5000+. I considered trying geographical filters, but as the existing data did not support it, I thought that a list of all plants with a Spanish common name would be a good start.

I decided to include this query as a complementary resource to the programmable search engine. The Wikidata query has features that make it appropriate to a subject guide. First, the results are live, meaning that the query runs every time the page is opened, and the results reflect the state of the data in Wikidata at that moment. If there is an improvement in the data, it will reflect it, but otherwise it will show to anyone interested in contributing to Wikidata what data is missing. Second, the data is presentable as a table, which I used in my case to show both common and scientific names, with hyperlinks to each item in Wikipedia, and it is a searchable table. Third, the query is editable and can be copied to be used elsewhere and adapted by anyone interested in doing so, for example to search all values in another database, or for common names in other languages (these would be the most simple substitution edits).

Presenting the query boxes

To facilitate use of the Google programmable search engine, I shared an annotated list of the 10 sites the engine was programmed to search. This would give users an idea of what results to expect, but it also serves indirectly as a list of internet resources which can be accessed directly.

To facilitate use of the Wikidata query, I included a brief explanation of how it works.

To encourage the use of both for someone with a limited knowledge of Puerto Rican plants, I included a list of pictures with names of Puerto Rican plants that could be entered in the search boxes, including what type of interest they might have: edible wild fruits, plants with

ethnobotanical significance, and threatened plants. To make sure that the plants would appear in the Wikidata search, I edited the Wikidata pages for the plants suggested to make sure they included the common names provided. Cheating, I know, in terms of my project, but an improvement to the representation of Puerto Rican plants on Wikidata.

Link to subject guide page

My page can be accessed here.

References

Mari Mut José A. (2014). Frutas olvidadas de Puerto Rico. Ediciones Digitales. Retrieved November 13 2022 from http://biblioteca.uprrp.edu/BIB-COL/cpr/Ediciones Digitales PDFs/Frutas olvidadas de Puerto Rico.pdf.

Woodbury, R. O. (1975). *Rare and endangered plants of Puerto Rico: a committee report*. US Department of Agriculture, Soil Conservation Service. https://www.govinfo.gov/content/pkg/CZIC-qk86-p9-w66-1975/pdf/CZIC-qk86-p9-w66-1975.pdf

Caras, A., Travis, B. A., Lapinel, E. A., Tsai, M., & Farren, T. E. (2010). *Interpretive Programming at the Caguas Botanical and Cultural Garden in Puerto Rico* (Doctoral dissertation, Worcester Polytechnic Institute). https://digital.wpi.edu/downloads/tb09j621h