**Databases (EN.600.315) Final Project**

**Sack’s Snake Sssearcher**

**Isaac Nemzer | inemzer1@jhu.edu**

Sack’s Snake Sssearcher is an online database of all of the snake species of the world. I defined snakes as being a member of the suborder Serpentes—this basically means that the few legless lizard species are excluded as they are not truly snakes. The database includes information on when and by whom the species was discovered by, more detailed taxonomy, if it is venomous and/or live bearing, what country(ies) and continent(s) it lives in, the human population of these countries, and more.

Please head to *https://snake-sssearcher.herokuapp.com* to view the site.

**Platform**

As stated above, the php files and the database itself is hosted on heroku. The database is the JawsDB flavor of MySQL.

**Data Sources and Loading**

The main source of data for Sack’s Snake Searcher is from http://reptile-database.org/. They allow people to download a tab delineated file of all of the species in their database. However, much of the data is in very inconsistent formats so I wrote python scripts (in directory /py-scripts) to parse the data. I also crawled Wikidepia for data about continents, countries, and population and more importantly about venomous and live bearing species. I then wrote some more python scripts to parse that data as well. To load the data, I used the MySQL command LOAD DATA LOCAL INFILE ‘data.tab’ into TABLE table1 FIELDS TERMINATED BY ‘\t’ LINES TERMINTED BY ‘\n’; Since I had been switching between Bash on Windows and Cmd, sometimes ‘\r\n’ was used for line termination. You can see these SQL scripts in the /sql-scripts directory.

**User Guide**

To use Sack’s Snake Sssearcher, simply navigate to *https://snake-sssearcher.herokuapp.com*. You will be prompted with a lot of fields for entry. Things to note are that entering a common name will return all species with that as part of their common name, not just exact matches. Same goes for higher taxa and discoverer. Venomous and live bearing (versus egg laying) are booleans. Note that the population refers to the human population of the country, meaning if you searched for North American countries with population greater than 40,000,000, it would include Mexico and USA. For counting, “Countries with each Species” will group by the species and count the number of countries that it lives in. “Species in each Country” will group by country and count the number of species that live in each particular country. Also, this may be obvious but you cannot sort by the count without having specified something to count. Sort will default to ascending.

**Specialization**

The most specialized component of this project was the original data retrieval, whether that was scrubbing the data from reptile-database or scouring Wikipedia for additional information.

**Possibilities for Improvement**

In future iterations I would like to add more information on a snake’s venom, i.e. its potency and how much is injected per bite. I also would like to add photos of the species, but crawlers for this were difficult to write as some of the species are very obscure with few available images on the web.

**Output**

Here are a few suggested queries for which I will explain:

1. *Common name: Rainbow; Country: Brazil; Year discovered: Before 1900; Sort by: Species, Ascending*

This will return a list of species in alphabetical order by their scientific name that live in Brazil, were discovered before the year 1900, and include the word “rainbow” in their common name.

1. *Continent: North America; Venomous: Yes; Count the: Species in each Country; Sort by: Count, Descending*

What this returns is a list of countries in North America with their populations and ordered by the number of venomous snake species that live within its borders. At the top should be Mexico with 122,273,000 people and 81 venomous species

1. *Continent: Europe; With population: Less than 10000000; Venomous: No; Live bearing: No; Sort by Country, Ascending and Year discovered, Descending*

This will return a list of all non-venomous and egg laying that live in European nations with less than 10,000,000 people. It will sort them alphabetically by country and then by the year the snake was discovered, descending.

**Relational Table Specification**

CREATE TABLE IF NOT EXISTS snake(

genus\_species VARCHAR(100) NOT NULL, #‘Epicrates cenchria’

high\_taxa VARCHAR(150), # ’Boidae, Henophidia, Serpentes’

genus VARCHAR(50) NOT NULL, #‘Epicrates’

species VARCHAR(50) NOT NULL, # ‘cenchria’

author VARCHAR(100), # ‘LINNAEUS’

year INT, # 1758

common\_name VARCHAR(256), # ‘Rainbow Boa’

venomous VARCHAR(5), # ‘No’

live\_bearing VARCHAR(5), # ‘No’

PRIMARY KEY (genus\_species)

);

CREATE TABLE IF NOT EXISTS country(

name VARCHAR(256) NOT NULL, # ‘Brazil’

population INT, # 206789000

continent VARCHAR(256) NOT NULL # ‘South America’

);

CREATE TABLE IF NOT EXISTS lives\_in(

genus\_species VARCHAR(100), # ‘Epicrates cenchria’

country VARCHAR(256) # ‘Brazil’

);

**SQL Code (in a PHP File)**

I have removed lots of code, so the file does not actually look like this. This is just some SQL parts.

// joining the tables

$joined = True;

$qry = " snake S INNER JOIN lives\_in L";

$qry .= " ON S.genus\_species=L.genus\_species";

$qry .= " INNER JOIN country C ON L.country=C.name";

// for searching on continent

if ($cont = $\_POST['cont']) {

if ($where) {

$qry .= " AND C.continent='" . $cont . "'";

} else {

$qry .= " WHERE C.continent='" . $cont . "'";

$where = True;

}

$heads = array\_diff($heads, array("C.continent"));

}

// for searching on population

if ($pop = $\_POST['pop']) {

if ($comp = $\_POST['comp']) {

$adder = "C.population" . $comp . $pop;

if ($where) {

$qry .= " AND " . $adder;

} else {

$qry .= " WHERE " . $adder;

$where = True;

}

if ($comp == "=") {

$heads = array\_diff($heads, array("C.population"));

}

}

} else {

// only joining snake and lives\_in

$heads = array("S.genus\_species", "S.common\_name", "L.country",

"S.author", "S.year", "S.venomous", "S.live\_bearing");

$qry = " snake S INNER JOIN lives\_in L";

$qry .= " ON S.genus\_species=L.genus\_species";

}

// if searching on higher taxa

if ($htaxa = $\_POST['htaxa']) {

if ($where) {

$qry .= " AND S.high\_taxa LIKE '%" . $htaxa . "%'";

} else {

$qry .= " WHERE S.high\_taxa LIKE '%" . $htaxa . "%'";

$where = True;

}

// add higher taxa to the headers

array\_unshift($heads, "high\_taxa");

}

// if searching on year discovered

if ($yeardis = $\_POST['yeardis']) {

// tframe is > | < | =

if ($tframe = $\_POST['tframe']) {

$adder = "S.year" . $tframe . $yeardis;

if ($where) {

$qry .= " AND " . $adder;

} else {

$qry .= " WHERE " . $adder;

$where = True;

}

if ($tframe == "=") {

$heads = array\_diff($heads, array("S.year"));

}

}

}

// if searching on venomous

if ($venom = $\_POST['venom']) {

$adder = "S.venomous" . $venom;

if ($where) {

$qry .= " AND " . $adder;

} else {

$qry .= " WHERE " . $adder;

$where = True;

}

}

// if grouping and counting

if ($groupby = $\_POST['group']) {

if ($groupby == "S.genus\_species") {

$endqry = " GROUP BY S.genus\_species ";

// moving those items to the front

$heads = array\_diff($heads, array("L.country", "L.population", "S.genus\_species", "S.common\_name"));

array\_unshift($heads, "S.genus\_species", "S.common\_name", "COUNT(\*)");

}

if ($groupby == "L.country") {

$endqry = " GROUP BY L.country ";

// these are the only headers necessary

$heads = array("L.country", "C.population", "COUNT(\*)");

}

} else {

$endqry = "";

}

if ($sort1 = $\_POST['sort1']) {

// check that not sorting on count without specifiying something to count

if (!($sort1 == "COUNT(\*)" && !$\_POST['group'])) {

// first order by

$endqry .= " ORDER BY " . $sort1;

if ($sort1dir = $\_POST['sort1dir']) {

$endqry .= " " . $sort1dir;

}

// second orcer by

if ($sort2 = $\_POST['sort2']) {

$endqry .= ", " . $sort2;

if ($sort2dir = $\_POST['sort2dir']) {

$endqry .= " " . $sort2dir;

}

// moving those headers to the front

if ($sort2 == "C.population") {

$heads = array\_diff($heads, array("L.country", "C.population"));

array\_unshift($heads, "L.country", "C.population");

} else if ($sort2 == "COUNT(\*)") {

} else {

$heads = array\_diff($heads, array($sort2));

array\_unshift($heads, $sort2);

}

}

// moving those headers to the front

if ($sort1 == "C.population") {

$heads = array\_diff($heads, array("L.country", "C.population"));

array\_unshift($heads, "L.country", "C.population");

} else if ($sort1 == "COUNT(\*)") {

} else {

$heads = array\_diff($heads, array($sort1));

array\_unshift($heads, $sort1);

}

}

}

// expanding header array

$qrybeg = "SELECT " . implode(", ", $heads) . " FROM";

$allqry = $qrybeg . $qry . $endqry;

**Other Programs and Code**

Take a look in the /py-scripts directory for most of the scripts that I wrote to parse the rough data retrieved from reptile-database. Many of the data intermediary step on their way to how they are today are stored in /extra and the data actually imported into the tables that are used now are in the /data directory.