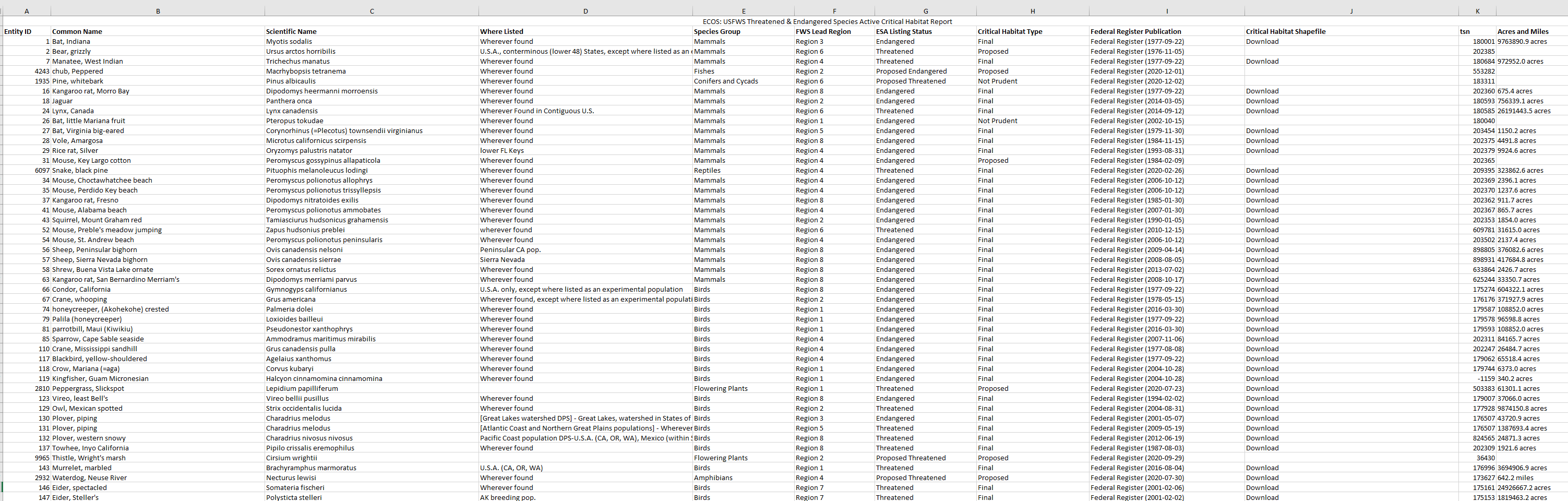
Things accomplished:

* 1. Picked out my MAIN final data source, which will be the ECOS data file, inclusive with a sole table that describes some elements of endangered species.
  2. <https://ecos.fws.gov/ecp/report/table/critical-habitat.html> <-- MAIN data source
  3. <https://data.worldbank.org/indicator/EN.MAM.THRD.NO> <-- Can pull some general information out of here.
  4. <https://data.world/datasets/endangered-species> <-- Another potential data site.
  5. It's going to be difficult to integrate multiple data sources as of now, unless I figure out a way to generalize my database schema and account for populating tables with different data sources…
  + This sounds feasible (Update bullet point #1)
  + I can make my own schema as opposed to basing one off of my data. Using the various variables I have across data, I can choose to populate certain variables or not.
  + Observed and analyzed the data to be inserted into SQL.
  + I have realized I might need to do something with metadata if I'm to properly create a schema with this data.



This is what the data looks like in Excel. The following columns are part of this record…

* 1. Entity ID (This is the Primary Key to be used across tables, as it uniquely identifies each species)
  2. Common Name (Comma separated string that tells you the name of the Animal as it would be used in a normal discussion or without any scientific complexities)
  3. Scientific Name (The proper name of the animal/species)
  4. Where Listed (I believe the origin of where it is considered endangered, or was first found to be endangered)
  5. Species Group (Mammals, Birds, or Reptiles)
  6. FWS Lead Region (As listed on the official site,

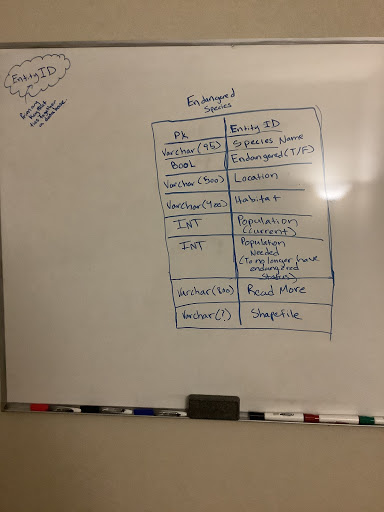
"Lead region for the species. Click the hyperlink to access the homepage of the FWS region. From there you have access to contact information for the region."

*From <*[*https://ecos.fws.gov/ecp/report/table/critical-habitat.html*](https://ecos.fws.gov/ecp/report/table/critical-habitat.html)*>*

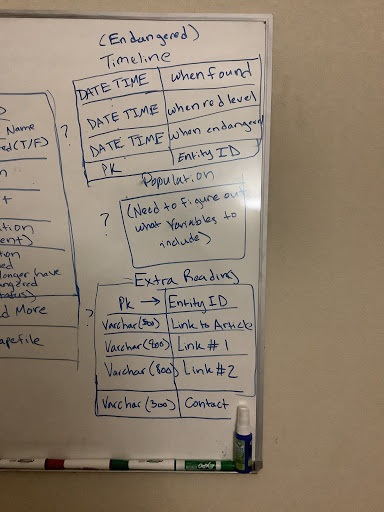
* 1. ESA Listing Status (Whether said species is Endangered or not. What level of Endangerment, etc.)
  2. Critical Habitat Type (This variable says Proposed/Final, I am not sure what it represents…)
  3. Federal Register Publication (Critical Habitat Designation… again, I'm not sure what this represents but might be good to investigate further.)
  4. Critical Habitat Shapefile (This is metadata which I believe includes a LOT more information on animal species)
  5. TSN (Shows up on the excel data file but not on the official website. Am not sure what this variable represents)
  6. Acres and Miles (How big is the habitat in which the endangerment is currently active (?) This is another shapefile correspondence)

Based on all of this information, I need to figure out what parts of this data are valuable and what aren't. For a small scale project, are hexagonal 3D shapefiles of a habitat reaaalllyy necessary? Questions need to be asked- and answered- all by me, the designer, the one finding the data, and the one asking myself to make a easily understandable, thorough database. I need to pick key elements across data files which I think can tie in the overarching purpose of this DB.

So, I got to work, brainstorming ways I can create some kind of data model with what I have now.

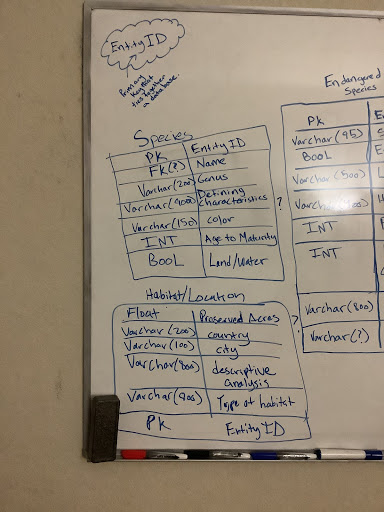


This represents a main table, the first table one would check when they open the database, and this would be a very simple, maybe 8-10 column table telling you things like a basic name of the species, where they fall on an endangered scale, what location they're in, what population numbers look like, habitat, etc.



Now these three tables focus on three different elements.

* 1. Timeline of Endangerment
  2. When was this species first discovered?
  3. When did this hit a level of endangerment? (i.e when did Scientists fear it was quickly moving up the scale?)
  4. How about an in between? When did scientists FIRST fear it moving to endangered status?
  5. All of this will be connected via the Entity ID (PK)
  6. Population
  + Because this is so vague, I need to figure out exactly what to put in this table.
  + Maybe something about current population numbers? Something about when population = x (where x is the number needed for it to escape being endangered)
  + Extra's Table! (TBD)
  + For interested individuals, I should provide a link to the site upon which the data was retrieved from, maybe a contact number, some links to check out for various species types (this might be too ambitious though).



Now these two tables focus on some other different elements. These are the really important ones.

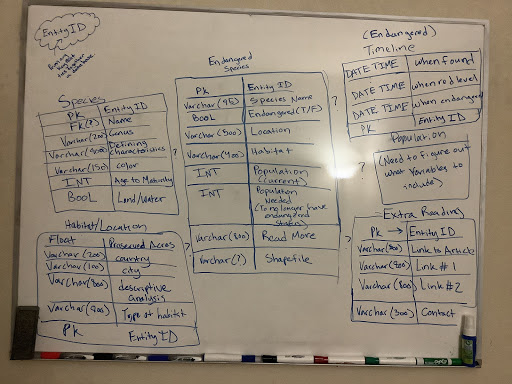
* + Species
  + There are a TON of variables to consider here. Name, Genus, Defining Characteristics, Color, Age to Maturity, Does it walk on land or live underwater? (Though this could technically go to habitat)
  + All connected by an Entity ID
  + It is not necessarily true that all of these descriptive variables for a species will be filled out. Maybe some datasets provide details like color, or genus, or maturity age, whereas some datasets don't. As long as I have an entity ID for each unique species, it'll be easy to just populate missing data with null values.
  + Habitat and Location
  + Again, this is a very important table like Species. Many factors need to go into this, such as Preserved Areas, Country, City, Type of Habitat, etc.
  + Connected via Entity ID
  + It seems like my main dataset doesn't do much with habitat/location. It's not easy to get this data from their 3D shapefiles, so perhaps for the ECOS data, this table may be empty entirely. (Efficient? Potential design Complications?)

Database Design is NOT easy! This is such a taxing process, figuring out how you're going to format different datasets, how to make a schema that will accommodate different variables coming from different websites (How about same species listed on multiple websites? What about more descriptive datasets on said species?)

There's a lot of tradeoffs to consider, and things I am learning throughout this entire process about what things I need to keep and what things I may just be having too many high hopes for.

A lot of this design is realizing that you will need to work- and rework- and rework your database model probably a hundred times before you find something even remotely close to a finished product. Planning out my various sketches took around 20-25 hours a week of brainstorming, trying and retrying tables, variables, and thinking of how to best fit multiple data sources into something relatively easy to understand.

My "SUPER- SUPER" Rough Outline



After erasing the board close to a hundred times, I finally found my "SUPER SUPER" rough sketch of a potential data model. I say SUPER because I bet this will change way more once I start compiling my data and populating the tables, and realizing things might be more efficient to be done said way as opposed to some other way.

I need to put all this into mySQL, make some last minute touch ups (I'm planning to possibly add a table or two, or even remove some from this current sketch, depending on the type of data I have), and then start populating the tables!

Relations also need to be formed still. I left those out because I plan to solidify them in my sql data model.