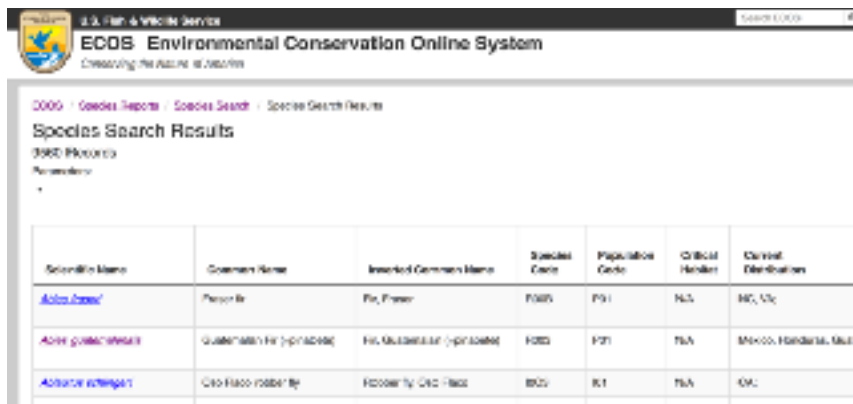


Assignment #2 Scraping and Scrubbing Data

Simon Myunggun Seo (ms9144)

The data I was initially curious to see was the population of cute polar bears over the years. However, I could not find such integrated data that spans several decades.

Inspired from this idea, I started searching for data of other endangered species and came across a database called ECOS (Environmental Conservation Online System) hosted on the U.S. Fish & Wildlife Service website. This seemed reliable since it was maintained by a federal agency. I took the entire database (<https://ecos.fws.gov/ecp0/reports/ad-hoc-species-report-input>) by selecting all fields possible and I saved the webpage and mirrored it on my website (http://i6.cims.nyu.edu/~ms9144/external/Species_All.htm).



The screenshot shows the ECOS web interface with a search results table. The table has columns for Scientific Name, Common Name, Inverted Common Name, Species Code, Population Code, Critical Habitat, and Current Distribution. Three rows are visible, all for the species 'Adelphi'.

Scientific Name	Common Name	Inverted Common Name	Species Code	Population Code	Critical Habitat	Current Distribution
Adelphi	Pearl Bear	Pearl, Pear	R00B	P01	N/A	MC, US
Adelphi guatemalensis	Guatemalan Pear (pradad)	Pear, Guatemalan (pradad)	R00B	P01	N/A	Mexico, Honduras, Guat
Adelphi guatemalensis	Guatemalan Pear (pradad)	Pear, Guatemalan (pradad)	R00B	P01	N/A	MX, US

According to this database, there were at least 9,500 species that were ever considered to be fragile species in all kingdoms of organisms. This database has specific information about the habitat of each species and also records government documents so it could be useful for creating conservation campaigns for each region or state.

There are 17 fields of data for each species in this database:

Scientific Name, Common Name, Inverted Common Name, Species Code, Population Code, Critical Habitat, Current Distribution, Family, First Listed, Group, Lead Region, Federal Listing Status, Regions of Occurrence, Special Rules, U.S. or Foreign Listed, Vertebrate/Invertebrate/Plant, Where Listed

I wrote a Python script to scrap all the data exactly into a CSV file.

However, this data had a lot of useless information for me. The 17 fields include ecological, biological, taxonomical, and administrative data. Some of these were easy to figure out what they mean, but some (like species code: R00B) were of little meaning to

Scientific Name	Common Name	Inverted Common Name	Species Code	Population Size	Current Habitat	Current Distribution	Family	Species Group	Lead Fed. Federal Listing Off.	Regions of Occur.	Species Status	Vertebrate/Invertebrate	
<i>Rorippa nasturtium</i>	Fraser R.	Fr. Rorippa	R004	P01	N/A	BC, WA	Brassicaceae	Cauliflora and Cystost.	5	Not Listed	4, 5	N/A	P
<i>Rorippa nasturtium</i>	Rorippa (Fr. (nasturtium))	Fr. Rorippa (nasturtium)	R004	P01	N/A	Mexico, Honduras, Gu.	Brassicaceae	Cauliflora and Cystost.	Foreign Threatened	NA	N/A	P	
<i>Rorippa nasturtium</i>	Das Flaco (Rorippa Fr.)	Rorippa Fr. Das Flaco	R000	R01	N/A	CA	Brassicaceae	Insects	1	Not Listed	8	N/A	I
<i>Rorippa sp.</i>	Hardy (Fr. Rorippa) (nasturtium)	Hardy (Fr. Rorippa) (nasturtium)	R000	P01	N/A	CA	Brassicaceae	Flowering Plants	8	Not Listed	8	N/A	P
<i>Rorippa nasturtium</i>	Rorippa (Fr. Rorippa)	Rorippa (Fr. Rorippa) (nasturtium)	R000	P01	N/A		Brassicaceae	Flowering Plants	5	Not Listed	NA	N/A	P
<i>Rorippa nasturtium</i>	No common name	No common name	R004	P01	N/A		Brassicaceae	Flowering Plants	5	Not Listed	NA	N/A	P
<i>Rorippa nasturtium</i>	(nasturtium) (nasturtium)	(nasturtium) (nasturtium)	R004	P01	N/A	WY	Brassicaceae	Flowering Plants	8	Not Listed	8	N/A	P
<i>Rorippa nasturtium</i>	(nasturtium) (nasturtium)	(nasturtium) (nasturtium)	R004	P01	N/A	WY	Brassicaceae	Flowering Plants	2	Not Listed	2	N/A	P
<i>Rorippa nasturtium</i>	Large (Fr. Rorippa) (nasturtium)	Large (Fr. Rorippa) (nasturtium)	R000	P01	N/A	Texas, E.R., CO	Brassicaceae	Flowering Plants	2	Endangered	2	N/A	P
<i>Rorippa nasturtium</i>	(nasturtium) (nasturtium)	(nasturtium) (nasturtium)	R000	P01	N/A	WY	Brassicaceae	Flowering Plants	1	Not Listed	8	N/A	P
<i>Rorippa nasturtium</i>	Rorippa (Fr. Rorippa) (nasturtium)	Rorippa (Fr. Rorippa) (nasturtium)	R000	P01	N/A	WY, Possibly extinct	Brassicaceae	Flowering Plants	1	Not Listed	1	N/A	P
<i>Rorippa nasturtium</i>	Pin (Fr. Rorippa) (nasturtium)	Pin (Fr. Rorippa) (nasturtium)	R000	P01	N/A	CA, OR	Brassicaceae	Flowering Plants	1	Not Listed	1, 2	N/A	P
<i>Rorippa nasturtium</i>	No common name	No common name	R004	P01	17,000	IN, U.S.A., 3B	Brassicaceae	Flowering Plants	1	Endangered	1	N/A	P
<i>Rorippa nasturtium</i>	Teddybear	Teddybear	R000	P01	N/A	BR, VA	Brassicaceae	Flowering Plants	4	Not Listed	4	N/A	P
<i>Rorippa nasturtium</i>	Rorippa	Rorippa	R004	P01	N/A	IN, U.S.A., 3B	Brassicaceae	Flowering Plants	1	Endangered	1	N/A	P
<i>Rorippa nasturtium</i>	No common name	No common name	R004	P01	N/A	WY	Brassicaceae	Flowering Plants	2	Not Listed	2	N/A	P
<i>Rorippa nasturtium</i>	No common name	No common name	R004	P01	17,000	IN, U.S.A., 3B	Brassicaceae	Flowering Plants	1	Endangered	1	N/A	P
<i>Rorippa nasturtium</i>	No common name	No common name	R004	P01	N/A	BR, VA, Possibly extinct	Brassicaceae	Flowering Plants	4	Not Listed	4	N/A	P
<i>Rorippa nasturtium</i>	Rorippa	Rorippa	R000	P01	N/A	IN	Brassicaceae	Flowering Plants	1	Not Listed	1	N/A	P
<i>Rorippa nasturtium</i>	U.S.A.	U.S.A.	R004	P01	17,000	IN, U.S.A., 3B	Brassicaceae	Flowering Plants	1	Endangered	1	N/A	P
<i>Rorippa nasturtium</i>	(nasturtium) (nasturtium)	(nasturtium) (nasturtium)	R004	R01	N/A	WY	Brassicaceae	Insects	4	Not Listed	4	N/A	I
<i>Rorippa nasturtium</i>	Deposited	Deposited	R004	R01	N/A		Compositae	Deciduous	NA	Not Listed	NA	N/A	I
<i>Rorippa nasturtium</i>	Phacelia (Fr. Rorippa)	Phacelia (Fr. Rorippa)	R000	R01	N/A	IL, VA	Compositae	Insects	3	Not Listed	3	N/A	I

me. Excluding fields that I did not have an intuitive understanding of, I decided to keep the following 8 fields:

Scientific Name, Common Name, Current Distribution, Family, Group, Federal Listing Status, Regions of Occurrence, Vertebrate/Invertebrate/Plant

Hopefully someone with more expertise can give an explanation for what each fields exactly describe.

As I could not simply visualize this qualitative data easily using graphs, I sorted each column and scrolled through, focusing on the top and bottom of the table for anomalies. There weren't significant anomalies although I found that some cells needed a small cleanup. Each field noted unavailable data in different ways.

The "Current Distribution" field, which describes the names of countries or states that each species is observed, left unavailable data empty.

The "Regions of Occurrence" field, which describes broad geographical regions in the US where each species is observed, marked it as "NA" which made me confused whether it signified North America or Not Applicable.

The "Common Name" field, which describes common, as opposed to academic, names of species, marked it as "No common name".

Some fields (that I decided to remove) marked it as "N/A".

I decided to reiterate through the data and unify everything as "N/A". However, I left this one alone: "Federal listing status" which describes whether the species is listed as endangered/threatened according to US federal data marked it as "Not Listed".

