Cleaning Data is (Not!) for the Birds!

Amanda Devine and Julia Steier

Global Genome Initiative

SI Carpentries Brown Bag

January 27, 2021

"Keel-billed Toucan" by Andy Morffew is licensed under CC BY 2.0

"bird" by Mathias Appel is marked with CC0 1.0

Research Question

How many bird families, genera, and species are there on each continent?

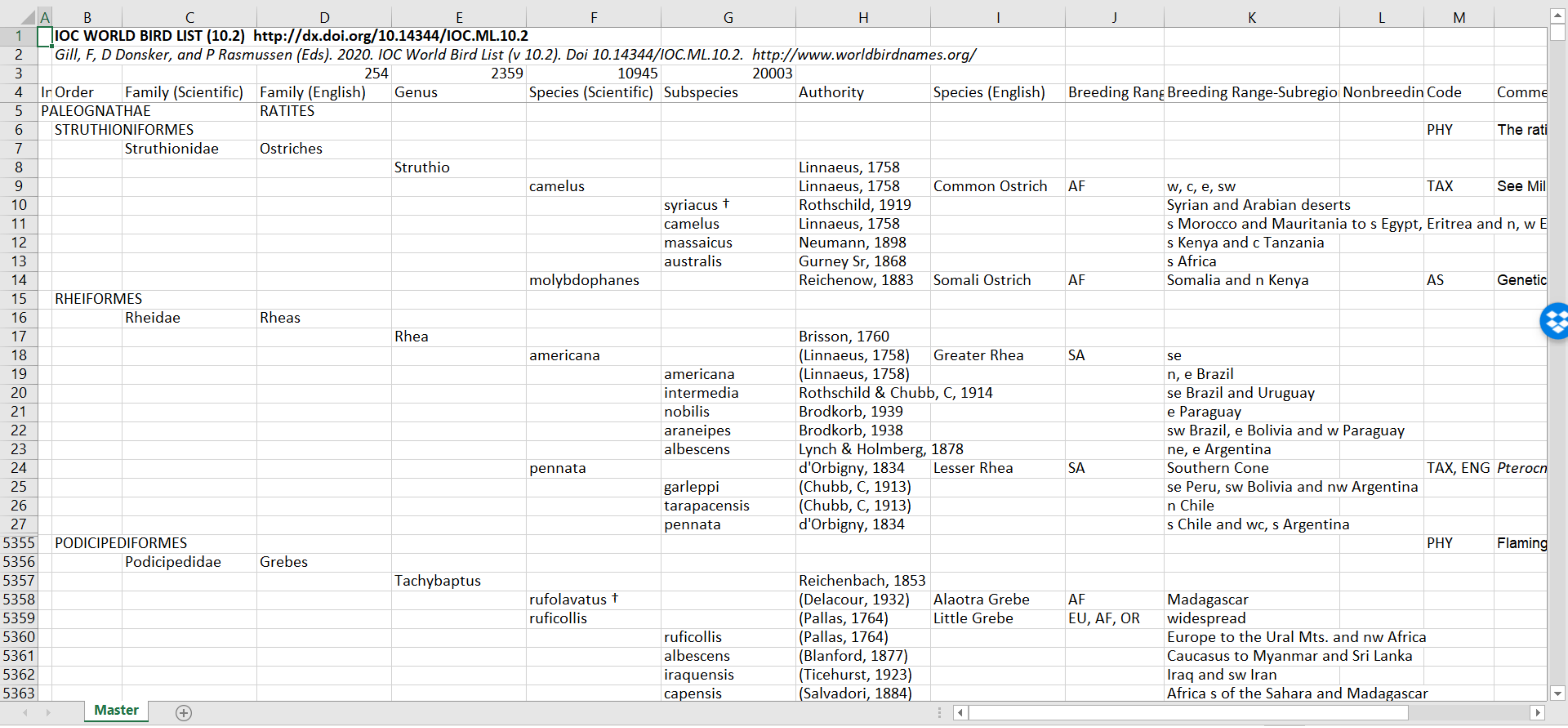
IOC World Bird List

- Up-to-date names and evolutionary classification of world birds
- ~34,000 species and subspecies
- **Range of each species**
  - **geographical region (‘NA’, ‘SA’, ‘AF’)**
  - qualifiers (‘widespread’, ‘e’, ‘se’)
  - specific countries
- Updated semi-annually

Gill F, D Donsker & P Rasmussen (Eds). 2021. IOC World Bird List (v11.1). doi : 10.14344/IOC.ML.11.1.

https://www.worldbirdnames.org/new/ioc-lists/master-list-2/

https://www.worldbirdnames.org/new/ioc-lists/range-terminology/

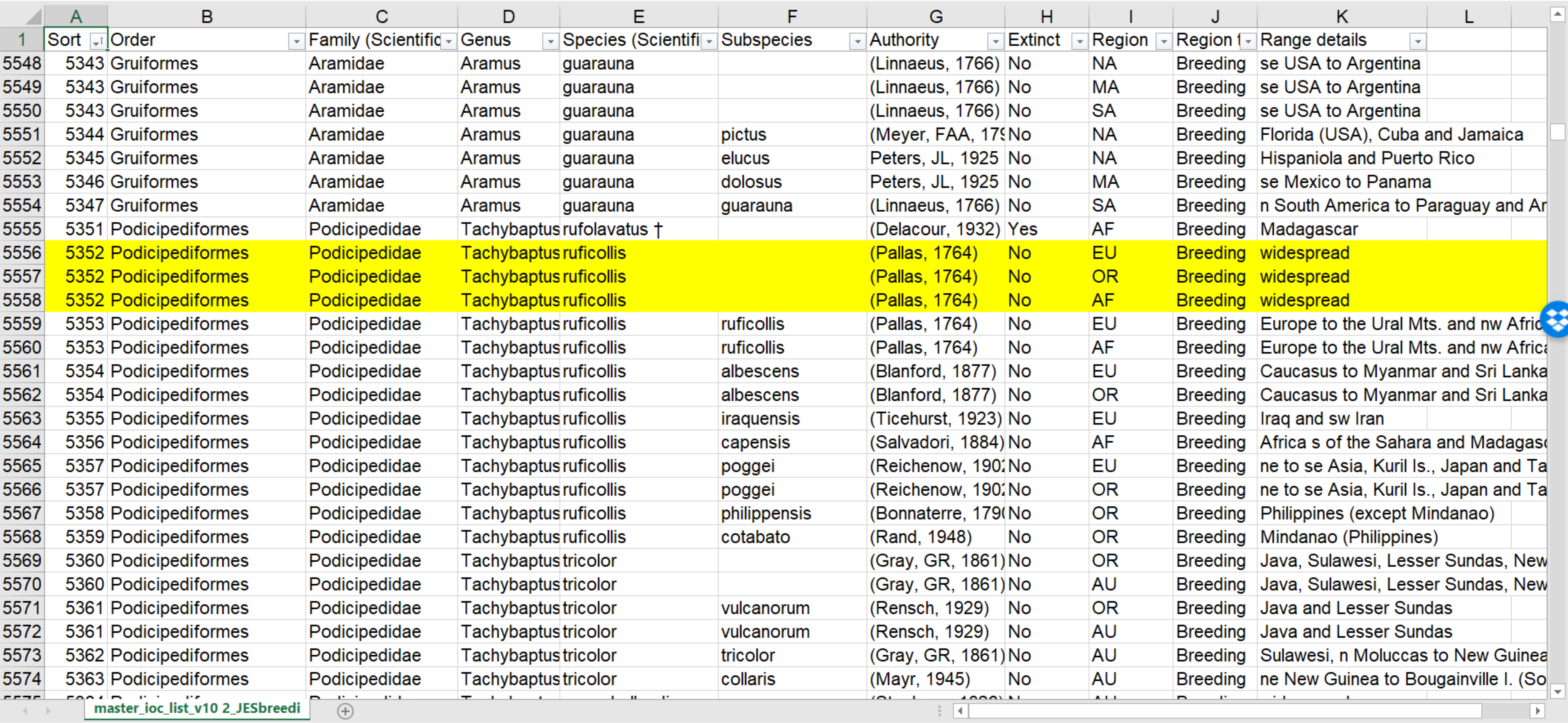


	A	B	C	D	E	F	G	H	I	J	K	L	M	
1		**IOC WORLD BIRD LIST (10.2) http://dx.doi.org/10.14344/IOC.ML.10.2**												
2		*Gill, F, D Donsker, and P Rasmussen (Eds). 2020. IOC World Bird List (v 10.2). Doi 10.14344/IOC.ML.10.2. http://www.worldbirdnames.org/*												
3				254	2359	10945	20003							
4	Ir	Order	Family (Scientific)	Family (English)	Genus	Species (Scientific)	Subspecies	Authority	Species (English)	Breeding Rang	Breeding Range-Subregio	Nonbreedin	Code	Comme
5	PALEOGNATHAE			RATITES										
6		STRUTHIONIFORMES											PHY	**The rati**
7			Struthionidae	Ostriches										
8					Struthio			Linnaeus, 1758						
9						camelus		Linnaeus, 1758	Common Ostrich	AF	w, c, e, sw		TAX	**See Mil**
10							syriacus †	Rothschild, 1919			Syrian and Arabian deserts			
11							camelus	Linnaeus, 1758			s Morocco and Mauritania to s Egypt, Eritrea and n, w E			
12							massaicus	Neumann, 1898			s Kenya and c Tanzania			
13							australis	Gurney Sr, 1868			s Africa			
14						molybdophanes		Reichenow, 1883	Somali Ostrich	AF	Somalia and n Kenya		AS	**Genetic**
15		RHEIFORMES												
16			Rheidae	Rheas										
17					Rhea			Brisson, 1760						
18						americana		(Linnaeus, 1758)	Greater Rhea	SA	se			
19							americana	(Linnaeus, 1758)			n, e Brazil			
20							intermedia	Rothschild & Chubb, C, 1914			se Brazil and Uruguay			
21							nobilis	Brodkorb, 1939			e Paraguay			
22							araneipes	Brodkorb, 1938			sw Brazil, e Bolivia and w Paraguay			
23							albescens	Lynch & Holmberg, 1878			ne, e Argentina			
24						pennata		d'Orbigny, 1834	Lesser Rhea	SA	Southern Cone		TAX, ENG	*Pterocn*
25							garleppi	(Chubb, C, 1913)			se Peru, sw Bolivia and nw Argentina			
26							tarapacensis	(Chubb, C, 1913)			n Chile			
27							pennata	d'Orbigny, 1834			s Chile and wc, s Argentina			
5355		PODICIPEDIFORMES											PHY	**Flaming**
5356			Podicipedidae	Grebes										
5357					Tachybaptus			Reichenbach, 1853						
5358						rufolavatus †		(Delacour, 1932)	Alaotra Grebe	AF	Madagascar			
5359						ruficollis		(Pallas, 1764)	Little Grebe	EU, AF, OR	widespread			
5360							ruficollis	(Pallas, 1764)			Europe to the Ural Mts. and nw Africa			
5361							albescens	(Blanford, 1877)			Caucasus to Myanmar and Sri Lanka			
5362							iraquensis	(Ticehurst, 1923)			Iraq and sw Iran			
5363							capensis	(Salvadori, 1884)			Africa s of the Sahara and Madagascar			

Master

Dataset Improvements

Analytical Task	**Data Cleaning Task**
Group and summarize by different taxonomic ranks	Populate full taxonomic hierarchies for each species and subspecies record
Group and summarize by different geographic regions	Split multi-value range cells into separate records
Filter out extinct species/subspecies	Create a column indicating whether a species/subspecies is extinct
Filter by breeding/nonbreeding region, or look at all regions in aggregate	Convert columns “Breeding region” and “Nonbreeding region” into columns “Region” and “Region type”



	A	B	C	D	E	F	G	H	I	J	K	L
1	Sort	Order	Family (Scientific	Genus	Species (Scientifi	Subspecies	Authority	Extinct	Region	Region t	Range details	
5548	5343	Gruiformes	Aramidae	Aramus	guarauna		(Linnaeus, 1766)	No	NA	Breeding	se USA to Argentina	
5549	5343	Gruiformes	Aramidae	Aramus	guarauna		(Linnaeus, 1766)	No	MA	Breeding	se USA to Argentina	
5550	5343	Gruiformes	Aramidae	Aramus	guarauna		(Linnaeus, 1766)	No	SA	Breeding	se USA to Argentina	
5551	5344	Gruiformes	Aramidae	Aramus	guarauna	pictus	(Meyer, FAA, 17	No	NA	Breeding	Florida (USA), Cuba and Jamaica	
5552	5345	Gruiformes	Aramidae	Aramus	guarauna	elucus	Peters, JL, 1925	No	NA	Breeding	Hispaniola and Puerto Rico	
5553	5346	Gruiformes	Aramidae	Aramus	guarauna	dolosus	Peters, JL, 1925	No	MA	Breeding	se Mexico to Panama	
5554	5347	Gruiformes	Aramidae	Aramus	guarauna	guarauna	(Linnaeus, 1766)	No	SA	Breeding	n South America to Paraguay and Ar	
5555	5351	Podicipediformes	Podicipedidae	Tachybaptus	rufolavatus †		(Delacour, 1932)	Yes	AF	Breeding	Madagascar	
5556	5352	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis		(Pallas, 1764)	No	EU	Breeding	widespread	
5557	5352	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis		(Pallas, 1764)	No	OR	Breeding	widespread	
5558	5352	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis		(Pallas, 1764)	No	AF	Breeding	widespread	
5559	5353	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	ruficollis	(Pallas, 1764)	No	EU	Breeding	Europe to the Ural Mts. and nw Afric	
5560	5353	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	ruficollis	(Pallas, 1764)	No	AF	Breeding	Europe to the Ural Mts. and nw Africa	
5561	5354	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	albescens	(Blanford, 1877)	No	EU	Breeding	Caucasus to Myanmar and Sri Lanka	
5562	5354	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	albescens	(Blanford, 1877)	No	OR	Breeding	Caucasus to Myanmar and Sri Lanka	
5563	5355	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	iraquensis	(Ticehurst, 1923)	No	EU	Breeding	Iraq and sw Iran	
5564	5356	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	capensis	(Salvadori, 1884)	No	AF	Breeding	Africa s of the Sahara and Madagas	
5565	5357	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	poggei	(Reichenow, 190	No	EU	Breeding	ne to se Asia, Kuril Is., Japan and Ta	
5566	5357	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	poggei	(Reichenow, 190	No	OR	Breeding	ne to se Asia, Kuril Is., Japan and Ta	
5567	5358	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	philippensis	(Bonnaterre, 179	No	OR	Breeding	Philippines (except Mindanao)	
5568	5359	Podicipediformes	Podicipedidae	Tachybaptus	ruficollis	cotabato	(Rand, 1948)	No	OR	Breeding	Mindanao (Philippines)	
5569	5360	Podicipediformes	Podicipedidae	Tachybaptus	tricolor		(Gray, GR, 1861)	No	OR	Breeding	Java, Sulawesi, Lesser Sundas, New	
5570	5360	Podicipediformes	Podicipedidae	Tachybaptus	tricolor		(Gray, GR, 1861)	No	AU	Breeding	Java, Sulawesi, Lesser Sundas, New	
5571	5361	Podicipediformes	Podicipedidae	Tachybaptus	tricolor	vulcanorum	(Rensch, 1929)	No	OR	Breeding	Java and Lesser Sundas	
5572	5361	Podicipediformes	Podicipedidae	Tachybaptus	tricolor	vulcanorum	(Rensch, 1929)	No	AU	Breeding	Java and Lesser Sundas	
5573	5362	Podicipediformes	Podicipedidae	Tachybaptus	tricolor	tricolor	(Gray, GR, 1861)	No	AU	Breeding	Sulawesi, n Moluccas to New Guinea	
5574	5363	Podicipediformes	Podicipedidae	Tachybaptus	tricolor	collaris	(Mayr, 1945)	No	AU	Breeding	ne New Guinea to Bougainville I. (So	

master_ioc_list_v10 2_JESbreedi

Demo: Fill down blank values in Excel

1. Fill all intentionally blank cells with a dummy value, e.g. zzzBLANK
2. Select range of cells that need to be “filled down”
3. Select all blank cells in range: Home → Find & Select → Go To Special… → Blanks
4. In active cell, type **=<up arrow key>** to set the cell equal to the one above it
5. Use Ctrl + Enter (or Cmd + Enter) to fill every blank cell with a reference to the one above it
6. Paste Special to replace formulas with values
7. Use Home → Replace to replace dummy value with an empty string

For more details: https://www.ablebits.com/office-addins-blog/2014/05/02/fill-blanks-excel/

Demo: Split multi-value cells into separate rows

1. **Important! Show as: records**
2. On the column containing the multi-value cells: Edit cells → Split multi-valued cells…
3. Optional: trim extra whitespace and convert to standard upper or lowercase
4. For each field that you would like to fill down: Edit cells → Fill down. **Important: save the first data column for last!**

For another method of handling this:
https://kb.refinepro.com/2012/03/fill-down-right-and-secure-way.html