Supplementary Material 4:

Matching names to a taxonomic backbone

Although many online taxonomic backbones have tools for species matching, you may need to match names to a list of names that does not have a matching tool. Some of these online species matching tools can be useful for fuzzy matching, which can match names even when there are spelling mistakes in your list of names. However, remember that these tricks for matching names need checking as some names can be quite similar e.g. Corymbia vs Corymbium or Boschia vs Boscia. The tips that are provided here are for the staff of conservation organisations, who are most familiar and comfortable with using MS Excel. Other solutions are also likely to be available. For example, functions in R could be used for name matching and Wikibase software from the Wikimedia foundation could possibly be used to match names against custom taxonomic backbones. Capacity building in using tools, such as OpenRefine and Wikibase software, could improve data workflows.

Use the GBIF Species Lookup Tool to deal with typos and get a canonical name, GBIF taxon key, GBIF scientific name and other taxonomic information

- Create an MS Excel document with three columns headed: ID, scientificName, Kingdom. The scientificName field must only include scientific names.
- Automatically create ID numbers 1, 2, 3 etc. or use any unique ID numbers. Autofill kingdom if all species are from the same kingdom e.g. Plantae or Animalia
- Save your file as CSV (comma delimited)
- Go to https://www.gbif.org/tools/species-lookup
- Make sure that you are logged in to GBIF
- Click on select file and navigate to your file. Or drag and drop the file onto the webpage.
- Click on Match to backbone.
- Wait for matching to finish processing.
- In the bottom right of the screen click Generate CSV to download the results.
- Rename the normalized.csv.
- The csv output file is UTF-8 encoded. The file needs to be imported into MS Excel using this encryption to keep special characters.
- In MS Excel, CLICK Data > From Text.
- Text Import Wizard > File Origin 65001 : Unicode (UTF-8).
- Delimited > separator comma.

The download from the GBIF species lookup tool will include the matched GBIF scientific name and a canonical name as well as rank of the taxon name and the status of the match. The GBIF species lookup tool employs fuzzy matching, so it will match names even if there are typos. As you have a field with the canonical name and a field with taxon rank, you can add subsp. for subspecies and var. for varieties to the canonical names so that you have a name without authors to match with. For example, the lion subspecies *Panthera leo subsp. melanochaita* will be returned as *Panthera leo melanochaita* in GBIF, and you need the version with the subsp. to match names. You can add this by using the Flash Fill or text to columns functionality in MS Excel to separate the components of the name, and then sorting by Taxon rank and adding a column where you can enter “subsp.” for all

subspecies and “var.” for all varieties. Then you can use the TRIM and CONCATENATE functions to put the names back together again as per the below.

Create versions of scientific names without authors

A useful tool for matching names is to create versions of names without authors for both lists of names. However, MS Excel will only match names that are written exactly the same. The example provided here is for a list of species including forma, var. and subsp.

- Make a copy of your list in a new spreadsheet and add a column names order and Auto Fill 1,2,3 consecutive numbers
- Select the column with the list of names and replace “forma” with a delimiter e.g. $. Use find and replace all in column. On the Home ribbon there is Find and Replace or use Ctrl+H.
- Use text to columns with delimiter $
- Label the new forma column ‘forma’
- Use text to columns to remove the author info from the forma column and delete all the author columns.
- Then, add new columns before the forma column to make sure you don’t copy over forma information when using the text to columns function again.
- After that, do the same for “var.” and “subsp.”, replacing var. with a delimiter and using text to columns to separate the var. and again to separate authors and deleting authors. Then, do the same for subsp.
- Check for hybrid names and these will not be caught properly in text to columns. Search for “x”. Other aspects of names that could be problematic are aff. and cf. Search for these.
- When only the genus and species names are left with authors, you can use text to columns and use space as a delimiter. Again, you need to be careful not to copy over columns with information in them. It is often easiest to copy a column with species names in to a new spreadsheet and use text to columns to separate the genus and species and author names and delete all the columns of author names or copy just the genus and species columns back to another spreadsheet.
- Sort by forma (make sure all relevant columns are selected), then add a column before forma and write the word forma and autofill / drag down for all rows that have forma
- Sort by variety and add a column before and write “var.” for all the rows with var.
- Sort by subspecies, add a column before and write “subsp.” for all the rows with subsp.

Genus	Species		subsp.		variety		forma
xxx	yyy	subsp.	zzz				
xxx	yyy	subsp.	zzz	var.	vvvv		
xxx	yyy					forma	ffff

Another way of allocating different ranks to rows is through searching for subsp., var. and forma using a formula. In different columns (e.g. columns B, C, D, E), the following formulas can be used to find and indicate subsp., var., forma and x. B2=IFERROR(IF(FIND("subsp.",A2,1)>0,"subsp."),"-"); C2=IFERROR(IF(FIND("var.",A2,1)>0,"var."),"-");

D2=IFERROR(IF(FIND("forma",A2,1)>0,"forma"),"-"); E2=IFERROR(IF(FIND(" x ",A2,1)>0,"x"),"-"); The same can be done for aff. and cf.

- Use the concatenate function to join the different parts (genus, species, subspecies, variety, forma) of the full name together. You can also use the trim function to remove extra spaces
  =TRIM(CONCATENATE(A2,“ ”,B2,“ ”,C2,“ ”,D2,“ ”,E2,“ ”,F2,“ ”,G2,“ ”,H2)

- There is also a way of using formulas to separate genus and species in Excel. If the full name (text string) is in cell A1. To get the first word in the string (i.e. Genus)
  =IFERROR(TRIM(MID(SUBSTITUTE(TRIM($A1)," ",REPT(" ",LEN(TRIM($A1)))), (1-1)*LEN(TRIM($A1))+1, LEN(TRIM($A1)))), ""). Then, to get the second word in the string (i.e. species)
  =IFERROR(TRIM(MID(SUBSTITUTE(TRIM($A1)," ",REPT(" ",LEN(TRIM($A1)))), (2-1)*LEN(TRIM($A1))+1, LEN(TRIM($A1)))), "")

- You can also edit data in Open Refine. In Open Refine, click the down arrow near the column heading. Select edit column. Select split into several columns. Change the separator to a space.

Dealing with spelling errors in species names when matching names

Some species names might not match because of spelling errors. Online matching tools e.g. The GBIF species lookup tool is useful to use to get fuzzy matches of species names. If the names you are trying to match have spelling mistakes, you can first use the GBIF species lookup tool to get accurately spelled scientific names before matching. However, names will only been matched to existing names in GBIF.

Using INDEX, MATCH, LOOKUP and VLOOKUP functions to match names in Excel

The INDEX and LOOKUP functions can be used to make identical matches between two lists of names in Excel and return the corresponding values in another column.

- Place the checklist/ taxonomic backbone you want to compare your names to in a new spreadsheet.

Sheet 1

A	B	C
Verbatim scientific name	**Name without authors**	**Match to checklist**
		=formula

Sheet 2

A	B
Full scientific name	**Name without author**

- The MATCH function returns the row where it finds the value you are looking for. The INDEX function returns the value in a specified row in a specified column. Therefore, you can use C2=INDEX(‘Sheet2!’A:A, MATCH(B2,‘Sheet2!’B:B,0)) where B2 is the value you want to match, column B in Sheet 2 is where you want to find the row of the matched name without authors, a zero is used to indicate an exact match, and Sheet 2 column A is where you want to find the corresponding scientific full name in the same row as the matched name without authors.

The MATCH and LOOKUP functions can also be used to make identical matches between two lists of names in Excel and return the corresponding values in another column.

- Place the checklist/ taxonomic backbone you want to compare your names to in a new spreadsheet.

Sheet 1

A	B	C
Verbatim scientific name	**Name without authors**	**Match to checklist**
		=formula

Sheet 2

A	B	C
1	**Full scientific name**	**Name without author**
2		
3		

- In this example, the list of names without authors that are being matched are in column B of Sheet 1 and the list they are being matched against is in another spreadsheet: Sheet2, where column A in the second spreadsheet is a consecutive list of numbers starting from 1 in row 1, 2 in row 2 etc.. In column B of the second spreadsheet is the list of names you want to match to. In column C of the second spreadsheet is the list of versions of names without authors. C2 =LOOKUP(MATCH(B2, 'Sheet2'!C:C,0),'Sheet2'!A:A, 'Sheet2'!B:B). Here, 'B2' is the value that you want to match, column C in Sheet 2 is the list you want to match to, '0' indicates that you want an exact match, column A in Sheet 2 is the index of numbers indicating the row number and column B in Sheet 2 is the column with the corresponding value you want to returned. The match function gives the output of the row number the B2 value is found in and the LOOKUP function looks for a value in the first
- You can also use this approach to add other information to your spreadsheet for a particular matched name e.g. you might have other taxonomic information or conservation status as affiliated data for the checklist you are matching to.

The VLOOKUP function can also be used to make identical matches between two lists of names in Excel and return the corresponding values in another column.

- When using VLOOKUP, the values you are looking up need to be in the first column of the lookup table or array.
- In this example, the list of names without authors that are being matched are in column B of Sheet 1 and the list they are being matched against is in another spreadsheet: Sheet2, where column A in the second spreadsheet is the list of scientific names (from the reference checklist) without authors, and column B of the second spreadsheet is the list of full scientific names. The VLOOKUP function requires specifying the lookup array. The extents of your lookup table can be determined/ highlighted using shift+ctrl+across and shift+ctrl+down. For the example in the formula, the array has 10 rows. C2=VLOOKUP(B2, 'Sheet2!'A1:B10, 2, false). Here, 'B2' is the value being looked up, 'A1:B10' is the data lookup array, with the values to lookup in the first column, '2' is the column number of the values that you want returned and 'false' indicates that an exact match is required.
- You can also use this approach to add other information to your spreadsheet for a particular matched name e.g. you might have other taxonomic information or conservation status as affiliated data for the checklist you are matching to.

Sheet 1

A	B	C
Verbatim scientific name	**Name without authors**	**Match to checklist**
		=formula

Sheet 2

A	B
Name without author	**Full scientific name**