Supplementary Material 1:

Data cleaning in MS Excel

This supplementary material provides some basic tips for cleaning biodiversity occurrence data in MS Excel, as this is the software that the staff of conservation organisations are familiar and comfortable with using. However, errors can be introduced to data in MS Excel through human error and automatic cell formatting. Therefore, where possible, we encourage conservation organization to investigate the usage of OpenRefine for its data verification functionality. Capacity building in using tools, such as OpenRefine and Wikibase software, could improve data workflows.

Some common issues requiring data cleaning

Biodiversity data require standardization to make it more useable. For example, data sets need detailed metadata so that data users know what the data represent and how they were collated. Additionally, there are multiple problems that occur across biodiversity data, including:

- Field headings often have inconsistent and unclear naming. Across datasets, it is useful for field headings to be consistent and understandable.
- Latitude and longitude coordinates are often inconsistent and sometimes have errors.
- Measurements may be recorded in multiple different units in one field.
- Dates are often recorded as Microsoft dates, which do not convert easily.
- Species names are often outdated.
- Data are misplaced, having been entered in the wrong field.
- Data are combined in one field and need to be separated into multiple fields.
- Data have been entered incorrectly.

Below, some data checking solutions in MS Excel are outlined. However, other software is also useful for data cleaning, e.g. OpenRefine and R.

Standardising and converting coordinate data in Microsoft Excel

Older biodiversity datasets often have latitude and longitude coordinates stored in different formats, and sometimes, there are other errors such as mixing up latitude and longitude, not including negative values and incorrect data entry. There are various checks and tools that can be useful for getting coordinate data into order.

- Check that the latitude column contains north-south coordinates and the longitude column contains east-west coordinates.
- Check that West and South coordinates are negative.
- Sort by latitude and longitude and check for outliers that don’t make sense.

Calculating decimal degrees

- If data are recorded as degrees, minutes and seconds, these need to be separated to calculate decimal degrees. First check that delimiters (separators) used do not make a problem for decimal seconds e.g. if a coordinate has been entered with a full stop between degrees, minutes, seconds and decimal seconds DD.MM.SS.DS.
- There are different ways of separating degrees, minutes and seconds: one way is using Flash Fill (on the Data ribbon) and another is using text to columns and a consistent delimiter.
- First select your whole spreadsheet (click in the top left of the spreadsheet where there is a triangle) and sort by latitude or longitude to get all rows with coordinate information.
- When using Flash Fill, create four new columns after latitude and longitude: Degrees, Minutes, Seconds, Decimal Degrees. For the first two rows write in the degrees, minutes and seconds and put in the formula for calculating decimal degrees: Deg + (Min/60) + (Sec/3600). Remember to multiply by -1 if south or west. Then, press Flash Fill when the cursor is in the third row of each of the first three columns/fields. You can also use autofill to copy the formula down to all the rows in the decimal degrees column. If the pattern that Flash Fill needs to follow is not entirely clear/consistent between rows you might need to include more examples for MS Excel to pick up the pattern. Remember to check that Flash Fill has converted the data the right way. For more on Flash Fill see https://support.microsoft.com/en-us/office/using-flash-fill-in-excel-3f9bcf1e-db93-4890-94a0-1578341f73f7.
- To use text to columns, you need consistent delimiters between degrees, minutes and seconds and you need to be aware of decimal seconds and how they are reported. If the same delimiter is used between degrees minutes and seconds, you can run text to columns once. Insert 5 or 6 columns to the right of your field. Copy and paste the latitude or longitude field into the first new column. Then, select this column and Click text to columns. Select Delimited. Select other and input the relevant delimiter. If there are different delimiters used, you will need to repeat the process. If seconds and decimal seconds are separated, they will need to be combined.
- Once the degrees, minutes and seconds are in different columns, the formula to calculate decimal degrees can be used in the next new column. = Deg + (Min/60) + (Sec/3600)
- To get the data without the formula, copy the column with the formulas for calculating decimal degrees. In the next column, Paste special > Values

You can also try use the MS Excel GeoCoordinatesParser Add-in, which is available at https:/bit.ly/coordsparser. This add-in works when latitude and longitude coordinates are in the same cell. If your latitude and longitude coordinates are in different cells, then you can put them into the same cell using the CONCATENATE function.

Standardizing measurement data to the same unit

In some biodiversity datasets measurement values and units are recorded in the same field and different units are used in one field. It is good practice to only record values in a field and for all the values to be of the same unit.

There are some tools in Microsoft Excel that can be used to convert inconsistent data with values and units to a standardised value field.

- First make a copy of the column you are working with and keep the original data.
- Sort your whole data sheet for the column you are working with so that the column contains a continuous set of data - this is useful for using autofill.
- To split values from units, you can try using Flash Fill. In two new columns write how you want the data to look (based on data in an adjacent column) for two rows. In the third row in your first new column, press the Flash Fill button on the Data ribbon. In the third row in the second new column, press the Flash Fill button on the Data ribbon.
- An alternative option, is to use formulas. See link for instructions https://exceljet.net/formula/split-numbers-from-units-of-measure. First column =MAX(ISNUMBER(VALUE(MID(A1,{1,2,3,4,5,6,7,8,9},1)))*{1,2,3,4,5,6,7,8,9})+1; Second column, To get unit: =TRIM(RIGHT(B5,LEN(B5)-C5+1)); Third column, To get number: =VALUE(LEFT(B5,C5-1)).
- You can also use FIND and IF functions to find different units in a column with values and units e.g. for a columns with mm, cm, m, ft. In the formula, A1 is the cell with the value and unit in it.
  B1=IFERROR(IF(FIND("mm",A1)>0,"mm"),IFERROR(IF(FIND("cm",A1)>0,"cm"),IFERROR(IF(FIND("ft",A1)>0,"ft"),IF(FIND("m",A1)>0,"m"))))
- When you have a column that tells you the unit of your data (and your data are in mm, cm, m or ft) you can use the following formula. In the formula, C1 is the cell with the unit in it (after you have separated or determined the unit). B1 is the cell with the value in it. The values can be obtained by using find and replace and replacing with nothing.
  =IF(C1="cm",B1,IF(C1="mm",B1/10,IF(C1="m",B1*100,IF(C1="ft",B1*30.48))))

Converting dates in Microsoft Excel

Dates are often in inconvenient formats in MS Excel that cannot be imported into databases. Additionally, MS Excel may sometimes inconsistently recognize dates, introducing errors (e.g. recognizing days 1-12 of the month as months and incorrectly storing these data, but correctly storing data for other days of the month).There are a few useful tools in MS Excel that can help with converting dates to a preferred format. However, it depends what format your data are in to start with and what format you want. For example, some databases need Day, Month and Year in different columns, whereas Darwin Core Archive format is YYYY-MM-DD. Our advice to staff of conservation organizations, who would typically work with spreadsheets, is that in addition to using Darwin Core Date format, Year, Month, Day and Time should be recorded in separate columns as data in date format are prone to becoming corrupt. Also remember to check that your date format does not change in your dataset e.g. from YYYY/MM/DD to MM/DD/YYYY.

- Try using Flash Fill to get the date format you want. If you want to have day month and year in number format in three different columns and they are in one column in date format, Add three columns and make sure that they are formatted as general not date. In the first two rows of the day, month and year columns write the relevant numbers. In the third row of each column use Flash Fill.
- Try using the functions DAY, MONTH and YEAR in three new columns.
- Try using text to columns. Select the column by clicking on the column heading. Click Data > Text to columns. Specify the delimiter/separator.
- Try using LEFT, MID and RIGHT functions.
- If you need to have a particular format, such as YYYY-MM-DD. Make a copy of the date column. Select the new date column and right click to display the context menu and select format cells. In the Format Cells dialog, under the Number tab, select custom from the list, and type yyyy-mm-dd in the text box and click OK.
- If the date is written as 8 numbers YYYYMMDD e.g. 20220331 and you need three different columns. Then, select the column with numbers. On the Data tab of the ribbon, click Text to Columns. Click Next >, then Next > again. Under 'Column data format', select Date, then select YMD from the drop-down next to the Date option button. Click Finish.
- To copy date data without formulas, select the column by clicking on the column heading. Right click. Copy. In the top row of another empty column, right click paste special values.
- To convert event data or time data to a text format, you can use formulas e.g. B2=TEXT(A2,"hh:mm:ss") and B2=TEXT(A2,"yyyy-mm-ddThh:MM:SS")