http://www.natsca.org

NatSCA News

Title: Excel for Data cleaning

Author(s): Mark Pajak

Source: Pajak, M. (2009). Excel for Data cleaning. *NatSCA News*, *Issue 18*, 27 - 30.

URL: http://www.natsca.org/article/131

NatSCA supports open access publication as part of its mission is to promote and support natural science collections. NatSCA uses the Creative Commons Attribution License (CCAL) http://creativecommons.org/licenses/by/2.5/ for all works we publish. Under CCAL authors retain ownership of the copyright for their article, but authors allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in NatSCA publications, so long as the original authors and source are cited.

Excel for Data cleaning

Mark Pajak

Documentation Assistant
Bristol City Museum and Art Gallery, Queen's Road, Bristol, BS8 1RL
Email: mark.pajak@bristol.gov.uk

Introduction

Every day museum collections catalogues swell with an increasing number of computerised records. However for some reason or other the data we have been plugging away so arduously will have to come out eventually. The problem is that this might not be in the appropriate format for whatever future purpose the data has, and it is not always possible to predict the format of data retrieval needed in the future.

With an increasing push for online public access to our databases, the implications for data quality and format are forcing many of us to rethink the way in which our data is organised. These events reveal how inappropriate or convoluted the data can be: bizarre relics from a previous system, misplaced capital letters, lack of common names, reversed dates, cryptic locality details and jumbled up initials are but some of the extravagant conventions that are hiding in collections databases. Data clean-up, therefore, can look like a horrifically monotonous if not impossible task.

Whether or not museum staff pride themselves in logical or repetitive tasks, the fact is that a computer will do them better. If you find yourself amending records in such a fashion then it is possible a bit of computer code will do the job for you. I was not aware of the full potential of Microsoft Excel when a year ago I was given the task of cleaning an entire set of natural history records prior to a data migration. I hope that the following article may reveal a few tricks that could save countless hours spent amending database entries.

Extracting Data

The first step is to get the data out of a database and into Excel by exporting the set of records to be cleaned. Most museum databases seem to display a single record at a time and do not allow space for manipulating or formatting data. Excel is designed to handle long lists of data, making it easy to spot erroneous entries in a column, and make bulk amendments with the trial and error approach accommodated using the undo facility. It should be possible to export only the fields that you want cleaned into excel. Exporting records creates a copy of the selected data which can then be opened and edited in Excel. This will display each exported record as a new horizontal row in a spreadsheet, with the field names as headings at the top of each column (Fig. 1). To facilitate import of the data back into a database it is vital that each record has a unique code that is not altered, to match back up with the original records – in most cases the accession number should perform this function. The entire dataset will be visible by scrolling down the worksheet. The resulting list of data can then be scrutinised for inaccuracies and formatted in specific ways using excel tools.

Basic Excel tools

In addition to being a number crunching program, Excel is very clever with what it can to with text, for example it 'knows' when you are repetitively entering in a term and will autocomplete terms for you, it can convert terms to Upper, Lower and Proper case and allows the usual Copy, Paste and Undo operations. Although it sounds trivial, being able to conjure up a drop down list of all the terms used in a field for your whole data set is extremely useful when cleaning data, and is more than many simple databases are capable of. I have found using the basic Excel tools in combination to be far more quick and reliable than cleaning raw data within a database one record at a time. It is a relatively simple matter to spot and change mistakes with these tools and then to import the cleaned data back into the collections database.

Advanced Excel Tools

Excel can perform more complex operations by inputting specially formatted code into the formula bar for each cell. Although somewhat off-putting (and beyond the scope of this article to fully explain) the code works like a simple language, allowing you to come up with novel functions and amend records in complex logical ways. If the task you need to be done can be written out in a logical statement, i.e. '**if *x*, do *y*'**, Excel will search for a term that matches your criteria x (Table 1), and then perform your specified function y (Table 1).

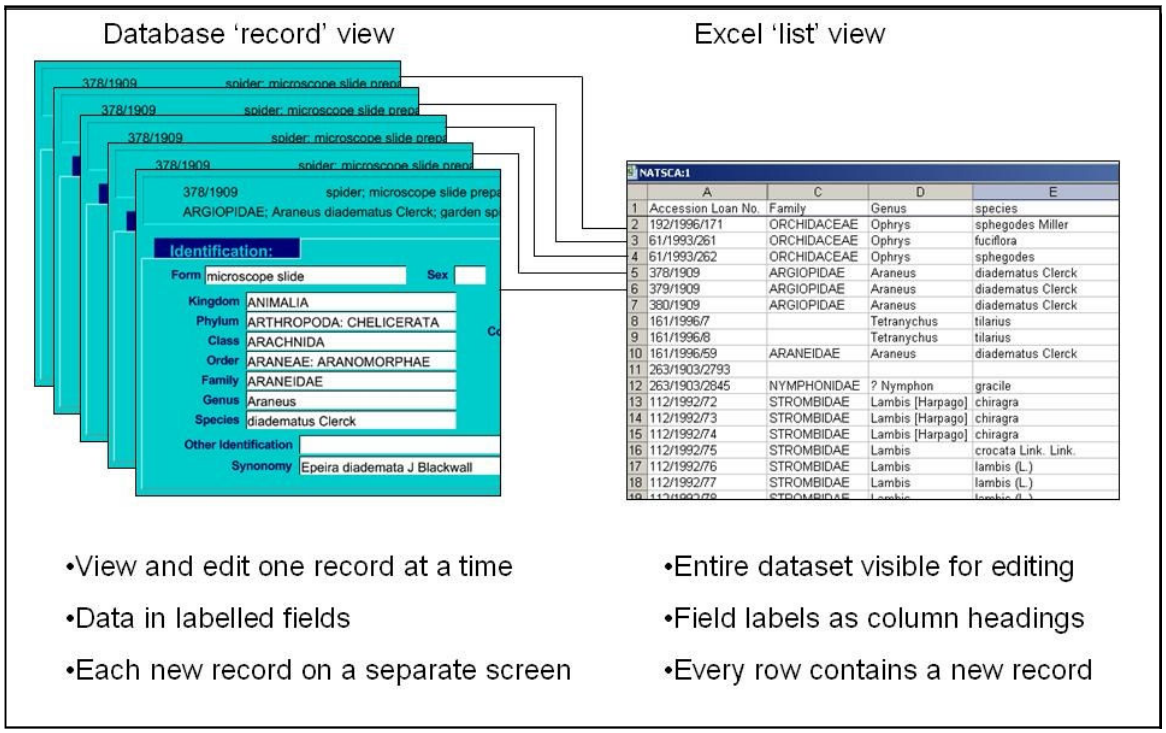

Fig. 1. Database view of records vs. Excel view of records

Table 1. Some examples of basic operations and the associated code.

Example criteria "*if x….*"	**Code For use in place of phrase "*if x….*"**
Contains phrase x (case sensitive)	=ISNUMBER(FIND("x",A1))
does not contain the phrase x	=NOT(ISNUMBER(SEARCH("x",A1)))
contains either phrase x or y	=OR(ISNUMBER(SEARCH("x",A1)),ISNUMBER(SEARCH("y",A1)))
contains both x and y	=AND(ISNUMBER(SEARCH("x",A1)),ISNUMBER(SEARCH ("y",A1)))

Table 2: Logic statements using example criteria

TaskTask "…do y"	Logic "if x do y, if not x leave blank"
Change the case of specific phrases	If cell A1 contains the letters idae (as in taxonomic rank=family), convert to upper case.
Code for above task:	=IF(ISNUMBER(SEARCH("*idae",A1)),UPPER(A1),"")
Join specific phrases	If cell A contains only the phrase 'var.'(as in botanical variety), join data in cells A1 and B1
Code for above task:	=IF(A1="var.",CONCATENATE(A1," ",B1),"")
Expand abbreviations, such as Authorities	If cell A is either 'Lin', 'L.' or 'Linne', change to 'Linnaeus'
Code for above task:	=IF(OR(ISNUMBER(SEARCH("Lin",A1)),ISNUMBER(SEARCH("L.",A1)), ISNUMBER(SEARCH("Linne",A1))),"Linnaeus","")

Table 2 illustrates some logic statements using example criteria. **Fig 2.** below shows the basic format of the logic code: commas separate both criteria and the operations to perform if the criteria are true, or if they are false.

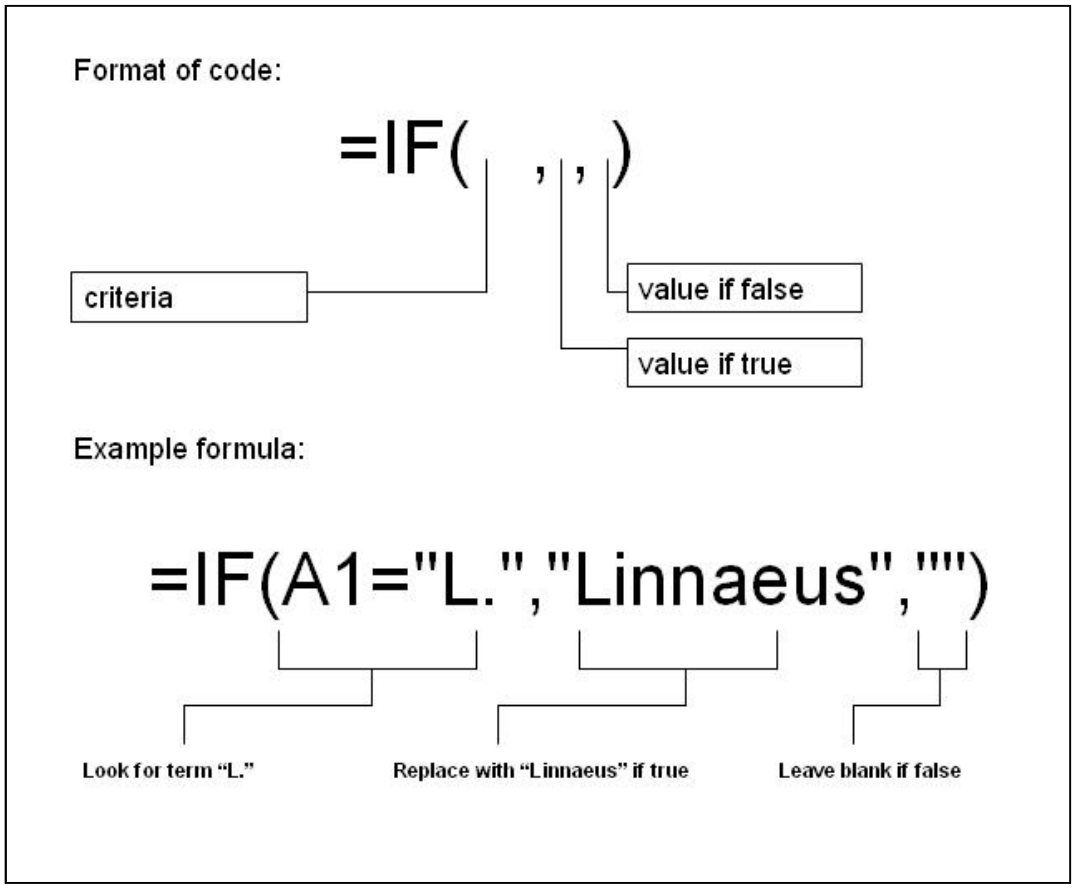

Fig. 2. Format of the logic code

Creating a data cleaning tool

On first glance it may seem that manually finding and replacing a single phrase with "Linnaeus" is a more simple task than getting to grips with Excel code, but when you have multiple terms to sort and 10,000 records to clean the time saved is significant: Multiple criteria and functions can be nested within brackets for complex conditional formatting, and you can be quite creative in the sort of functions you want to do. Once you have written the code in a cell in a new column, all you do is drag down to the bottom of the sheet and excel will automatically process each record – this is the same procedure as using the autofill function and excel will 'know' when you want it to move onto the next record and update the code accordingly.

Needless to say, it can get complicated, but the key is to recognise which data problems could be better tackled in Excel than flicking through database records. The current version of Excel will do about 65,000 records at a time. There are also many ways to protect cells, formulas and columns in a worksheet from accidental editing; this opens up the possibility of creating a data cleaning tool for other staff to use, without them ever coming into contact with an Excel formula (once you have put in the hard work in creating it!).

Summary

I put this into practise recently when records were due to be made available online for our public to browse. We did not until then have a designated "Common Name" field for our specimens and common names were tacked onto the end of a long scientific name in the "Full Name" field. This would not have been a very user friendly way of displaying our data and the only way around this seemed to be to create a new field and manually copy and paste common names from the end of the Full Name text string into the new empty field in the collections database. Using the "Text to columns" function in Excel, as well as a few other tricks outlined in this article I was able to export all of the scientific names and create an algorhythm to split them up automatically. Given a string of data such as "Anas platyrhynchos (Linnaeus): ANATIDAE: mallard", the algorhythm splits up the genus, species, authority, family and common names into separate fields which can then be re-imported into a taxonomic hierarchy within a database. I had to delve a bit deeper into Excel to automate the process but now our museum has an automatic tool for splitting scientific names where they occur into their separate parts. If this "scientific name parser" sounds like it would come in handy to others then please feel free to contact me – it is still in the development stage but could be tailored to suit other specific data cleaning problems.

Appendix – Useful Excel Functions for data cleaning

Autofill: Used to fill a range of cells with a series of identical values. Use this to quickly populate many records with identical data by clicking on the small black square icon and dragging down. You can also use this to automatically fill in a changing series of values, such as dates.

Autocomplete: Excel learns which entries you are filling in and so will complete a value after the first few letters are entered, speeding up data entry.

Data sorting: Whilst checking a particular field, you can sort the column alphabetically. This is a good way to spot typos as records with similar values are grouped together. Use "Expand Selection" to keep all other data associated with each record together.

Autofilter: This tool gives a glance at the distribution of data; any typos will show up here and can be selected and eliminated. Probably the most useful list processing tool, you can filter an entire sheet to show just a few particular terms for editing at any one time. You can also filter multiple terms at once to be more selective about which records you are editing, e.g. if you only want to edit records with containing data on a specific collector. Make sure to reset the auto filter when done to reveal the whole dataset again.

Date functions: Convert numerical dates to text and vice versa, select how many digits for the year and convert American to English formats. All that is required is to input the Excel code to format the data:

Freeze panes: Keeps the column headings visible when you scroll down a very long list.

Validate: Set data entry controls—convert numerical dates to text, only allows specific terms.

Text to columns: A problem that I had was fields in which multiple terms were entered in a string, separated by semi colons. Excel can split these terms into separate fields automatically, based on whichever 'separator' is used. So ,for example, a field in which multiple taxonomic terms have been entered can be split into separate columns in Excel and imported back into their proper fields in the database.