

Combining Mailing Lists

I was recently called in by a client who has built up a couple of database tables holding names and addresses of customers and possible customers. It was a surprisingly small task to combine the tables and clean up the data. This is probably quite a common situation and I thought I'd give you an idea of how to approach this.

I won't be specific to, say Access, as most database programs have the same features. If your lists are stored in spreadsheets most of the techniques will still apply. Most of the techniques are quite easy, although the fun really starts when you come to remove duplicate records.

On a matter of terminology, spreadsheets hold individual fields in columns and most database programs also use the term column. Some however will use the term field. I have used the term column here. I have also used the database term table. If your data is in spreadsheets a table will correspond to a worksheet.

First Steps

It goes without saying that you should back up all your data before you start.

Make copies of the tables. You will need to make some changes to these tables and clean up the data to some degree. It's worth making a second copy of one of the tables; this will turn into your new, combined table.

Design the Output Table

One problem will be that you will almost certainly have different columns in the different tables. So decide what columns you want in the combined table and what they will be called.

A bigger problem is that there may not be a one-to-one relationship between the columns in the tables. You may find data held in one column in one table is stored in more than one column in the other table. Decide what columns you actually need.

It may be necessary to modify some of the data. Ideally you would write a macro to do this. Or, in a spreadsheet, to concatenate the values of two columns you could enter a formula 'CONCATENATE (A5, " ", B5)' into a cell and copy it down through the rest of the column.

You may still need to do some manual cleaning and correct or input data by hand. This is quite easy in a spreadsheet, you may be able to use a table view in a database. Consider using a temporary column to hold working values.

Rename the column names in the copy tables to the names you will use in the combined table. Add any missing columns to the first table. Then copy the whole of this table to the output table.

Construct a new form to access the new record format.

Combining Mailing Lists

Append the data from the second table to the output table. A database will put all the data into the correct column. If you are merging spreadsheets, the columns will obviously need to be in the same order. It may be easier to copy data a column at a time.

Remove Duplicates

Locating duplicates is the biggest problem. To do this, you will have to sort the combined data. Start by inspecting the records and look for suitable ways to sort.

The best way to sort is by post code followed by first line of address – post code and house number uniquely identifies a property. This provides a good start point, and will show up most duplicate names. Some caution is needed, it is surprisingly difficult to identify duplicates. My household for instance contains me, J R McMillan and my son J M McMillan. Irritatingly he rarely uses his middle initial so he receives letters addressed to J McMillan. You will probably find entries for J Smith, J A Smith, John Smith, J. Smith, Mr J Smith and so on.

You need to check all the addresses in a post code. Some people may use a house name one time and a number another time. Punctuation will mess up a sort badly in names, addresses and postcodes. For example CO10 9XX with two spaces will come before CO10 1AA with one space.

Errors will also occur because of incorrect post codes. After eliminating duplicates by post code, do another sort by name. That may throw up more duplicates.

Finally

When you've finished, don't forget to remove all the working tables, otherwise in 2 years time you'll be wondering what they are.