

OVERVIEW

The intent of this document is to highlight the functionality that is available within Microsoft Excel that can be applied by those that need to manage data to support the administration of HPRP and other funding sources. While most HMIS applications have built in reporting tools, Excel is the least common technical denominator and is the most ubiquitous database application around. Most HMIS applications can export data in a comma separated variable (CSV) format that can be easily imported into Excel and worked with.

These techniques have been used to handle converting HMIS data from various formats into a common format used by the *Commonwealth of Massachusetts* statewide homeless database. This is just a fraction of what Excel has to offer and when combined with a bit of ingenuity they can be used to create rather powerful reports such as the HPRP QPR, the APR, and even AHAR.

Items that are covered in this documentation include the following:

- I. Identifying duplicate records
- II. Manipulate the database columns as needed to match the desired target format.
- III. Populate data quality codes using the length function:
- IV. Stripping out undesirable characters.
- V. Combining data elements that are stored across multiple columns into one column.
- VI. Splitting up one field into several fields using Excel (ie Name into First, Middle and Last):
- VII. Translating values to be consistent with the target format.
- VIII. Removing erroneous service records.

I. Identifying duplicate records

Oftentimes a database does not have edit checks in place to ensure that the information being entered does not already exist in the system. A unique key can be created based on the data you are collecting and then sorted in order to remove these duplicated records. To create this key you have to use the CONCATENATE function in Excel to bring the fields that are supposed to be unique into one key field. For example, only one client should have the same name, date of birth, gender and social. You may wish to perform this analysis using different keys in order to capture as many duplicates as possible. For example you may wish to do one just by first name and social, another by first name, last name and date of birth, etc.

NOTE: In order to check for inter-agency duplicate records it is imperative to perform this process over all programs at once using 1 data set.

To use this technique perform the following:

1. Insert two blank columns to the left of the data you are checking.
2. Use the concatenate function (eg. **=CONCATENATE(A2,B2,C2,D2)**) to populate the first empty column with the newly created concatenated field. A concatenated field that joins together client identifiers might look something like "SmithJohn03211977045-55-1923". Copy this formula down to the last row of data.
3. Click on **Data | Sort** and select the column containing the newly created key.
4. Check for duplicates using the EXACT command. If the new key field is in column A and Column B is blank enter the formula in the first row of Column B as follows **=EXACT(A2,A3)**. Copy the formula down to the remaining cells. The *EXACT()* function returns *TRUE* if the values it is given are identical, and *FALSE* otherwise.
5. Since it may be difficult to locate all of the cells that have a value of TRUE in them you may wish to highlight the TRUE values using conditional formatting. To do this select the test column and click on **Format | Conditional Formatting**. Arrange the drop down boxes to read "Cell Value Is" "Equal To" "TRUE". Select the Format button, followed by the **Patterns** tab, then choose the color to highlight the cell if the function value is true.
6. If there is potentially valid data in multiple records you may wish to merge the data into one record and delete the other(s). A duplicate might not be a data quality issue, but could be an issue of a client receiving services from

multiple programs which may or may not be allowed by program guidelines. HPRP for example does not allow clients to be served for the same reason by multiple agencies.

II. Manipulate the database columns as needed to match the desired target format.

Basic Manipulation

Inserting/Adding Columns – Simply select the column to the right of the column you wish to add, right click, and select “Insert”.

Deleting Columns - This process can be handled simply by selecting the column that is not needed, right clicking and selecting “Delete”.

Moving Columns – Select the column you wish to move, right click, select “Cut”, select the column to the right of where you want the column to go, right click, and select “Paste”.

Advanced Manipulation

1. Create a separate worksheet in the same workbook that has the source data.
2. Copy the target format field names into a header record (the first row) on the worksheet. You may have to use Copy | “Paste Special” and select “Transpose” in order to copy field names listed vertically to be listed across the worksheet.
3. Enter the equals sign (=) in the first column and first row under the newly copied file header row. Click over to the source file, select the corresponding field in the first row under the source file’s header row and hit “Enter”. Repeat this process until you have mapped one field for all of the values in your target database.
4. Copy the row containing these new formulas down your spreadsheet to at least as many rows as you have records in your source file.
5. Create a blank worksheet. Copy the entire worksheet using Select All (the box in top left corner of the workbook) and use “Paste Special”. Select the option to paste “Values” only.
6. Save the newly created target file.

III. Populate data quality codes using the length function:

Since fields such as zip code and social security number require a data quality code you can derive the code if partial or full based on the length of the data entered using the *LEN()* function to count the number of characters. Enter =**LEN(F2)** into Row 2 of a new column and copy the formula down the column. The database can then be sorted by this new length column and the Data Quality Code can be updated using either the copy/paste function or Find|Replace.

IV. Stripping out undesirable characters.

For the sake of data consistency it is common to remove extra characters such as \$,-/. If you are trying to link a social security number that has dashes with one that does not you will never get a match. To handle this you can use the Find|Replace function by highlighting the column you want to work with and hitting CTRL+F. To remove dashes you would search for – and replace it with blanks and click Replace All. To ensure the Find | Replace works only over the column you wish to work with you may consider copying the column to a blank worksheet and performing the Find|Replace there and then copying it back to the original worksheet.

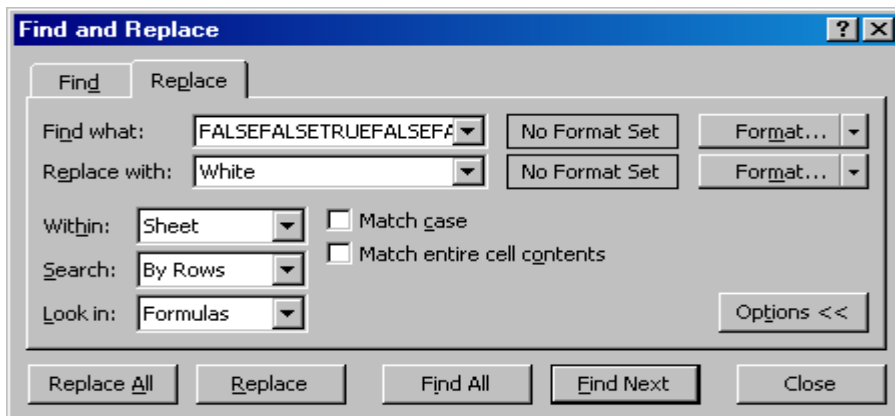
V. Combining data elements that are stored across multiple columns into one column.

In this example the race field is being stored as the individual response categories. In order to have one race field the concatenate function can be used along with the Find|Replace function. In the example below the data being worked on is in columns U to Y and the data being created is in column T.

	American Indian or Alaskan Native	Asian	Black or African American	Native Hawaiian or Other Pacific Islander	White
Race					
=CONCATENATE(U2,V2,W2,X2,Y2)	FALSE	FALSE	FALSE	FALSE	TRUE

In order for Find | Replace to work correctly the values (ie not the formulas) need to be copied into a new column. To handle this simply insert a blank column, select the column you populated data in with the concatenate function, right click and select copy, select and right click the column you just inserted, and click "Paste Special" and click the button next to "Values" and click OK.

Select the column to work, hit CTRL F to initiate Find|Replace and use CTRL C and CTRL V to copy and paste the created text string into the "Find What" box. In this case the created value for "White" is FALSEFALSEFALSEFALSETRUE. By searching for FALSEFALSEFALSEFALSETRUE and replacing it with "White" and clicking "Replace All" the data will now be consolidated correctly into one field. Repeat this process for each possible value (ie for "Alaskan Native" perform the Find|Replace function on TRUEFALSEFALSEFALSEFALSE).



VI. Splitting up one field into several fields using Excel (ie Name into First, Middle and Last):

- A) Right-click column heading for the name field and select **Insert**.
- B) Select the name field column then click **Data** in the menu bar and select **Text to Columns**.
- C) Select the option for **Delimited** and click **Next**.
- D) In the **Delimiters** section click **Space and/or Comma** (depending on how your data is stored) and then click **Finish**. Note if the data is stored as (Simmonds, Matthew D. Sr. you will have to insert three columns and use both comma and space as delimiters).
- E) When prompted, answer **OK** to replace the contents of the destination cells.
- F) Change the text in the column headers to now accurately reflect the change. (i.e. *Last Name, First Name, Middle Name, Suffix*).

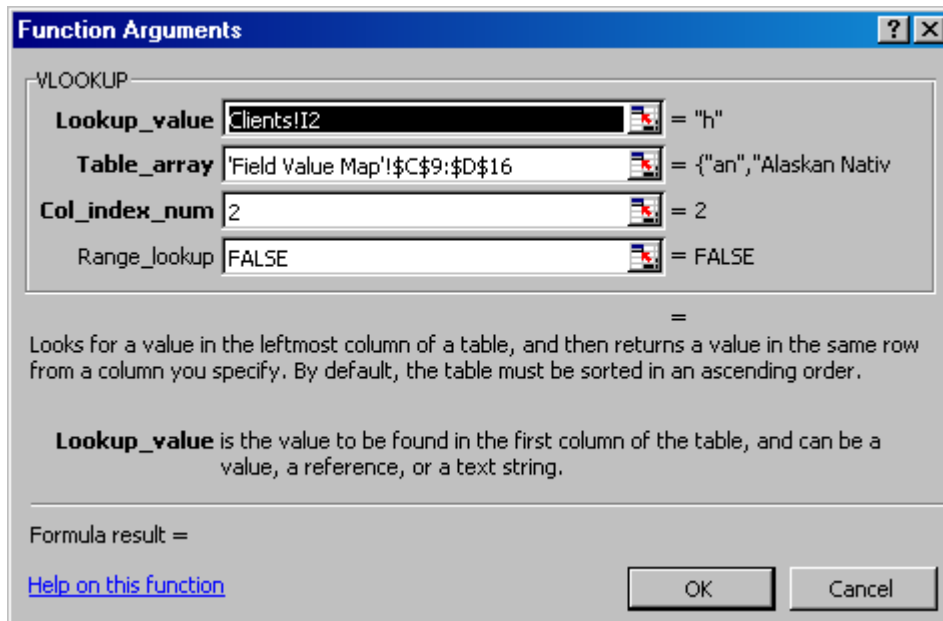
VII. Translating values to be consistent with the target format.

A database without consistency of values lacks integrity and will not serve the desired purpose. For this reason you cannot simply move over data without translating the values in the source system to the anticipated values in the target system. For example, in the source database gender may be represented as M and F while in the target database the values may be Male and Female. The easiest technical method of doing changing values is to simply use Find|Replace to change the values from the source to target value. The downside of this is that it requires more manual work and will not be effective for data integration/exchange that is of a more continuous (ie not a one-time) process. For regular data interchange or for those that wish to automate the process the best approach would be to use a data value translation table that would assist in the translation.

Steps for creating a data value translation table in Excel.

1. Create a master table of values used within the target database on a field by field basis.
2. Distribute the master table to the agencies that will be sending data and ask them to map their values to the target values. An example of mapping would be to enter in "M" (the value they use) next to the gender column value of "Male".
3. Use the VLOOKUP function in Excel to translate the source database values to the target values. (see below)
4. Replace the values with the values generated from the VLOOKUP function.

Using VLOOKUP to replace values:

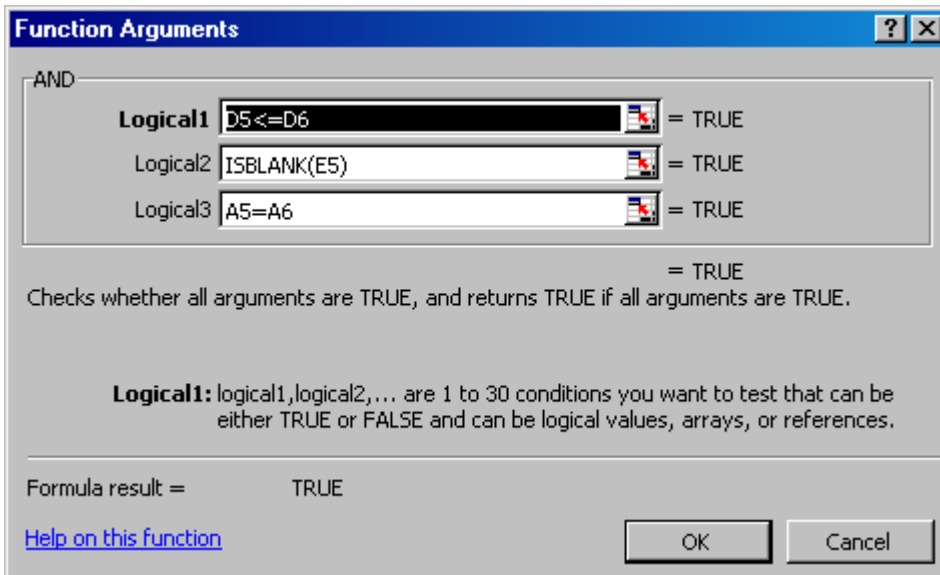


VIII. Removing erroneous service records.

Quite often it is difficult to figure out when a client has left (especially for programs such as street outreach) so the data records for a recorded service end up having blanks for an exit date. Blanks can be valid if you are still servicing the client. One way to check for erroneous blank conditions is to determine if you have provided the same client with the same service since the record where the blank exit occurs. For example, the 3rd record below is invalid because it is followed by a service record that has an exit date for the same service.

	A	B	C	D	E
	PROVIDER_ASSIGNED_NUM	SERVICE_NAM	AIRS_TYPE_CD	START_DTE	END_DTE
3	108493	emergency bed	BH-180.850	10/14/2003	10/15/2003
4	108493	emergency bed	BH-180.850	4/13/2005	5/23/2005
5	108493	emergency bed	BH-180.850	5/25/2005	
6	108493	emergency bed	BH-180.850	5/25/2005	5/26/2005

To check for these conditions you can use the =AND function as follows. The resulting formula would be entered in cell F1 and Copy and Pasted down to all other cells in the column.



The resulting formula is =AND(D5<=D6,ISBLANK(E5),A5=A6) would result in a TRUE condition for cell F4.

Using either conditional formatting (Format | Conditional Format) or a data filter (Data | Filter | Autofilter) will quickly show you all records where there is an invalid service record sandwiched in between two valid service records.

CONCLUSION

Microsoft Excel is filled with powerful functions that can be used to help cleanse your HMIS database. Other functions to look into include filtering out invalid records out using the FILTER command, auditing your data and creating reports using the COUNTIF function and using advanced data validation techniques to ensure clean data entry. Performing these tasks and mastering the functionality available within Excel will help ensure the integrity of your HMIS database. If you have questions or would like assistance in cleansing your HMIS database please feel free to contact Matt Simmonds at Matt@SimtechSolutions.com. You can also check in to the SimtechSolutions.com website as we should be updating this and other helpful documents from time to time and may be adding an HMIS discussion blog to the site in the near future.