Tools for Managing Digital Collections Series: Spreadsheets for Data Management
Written for Excel 2007 on the Windows OS

2013-03-19

Prepared by
Chris Lacinak
AVPreserve
chris@avpreserve.com
www.avpreserve.com

Table of Contents

Introduction

This workshop and group of exercises is designed to help people manage metadata more accurately, efficiently and effectively. While databases and database applications are widely recognized as being the most effective for management and use of metadata, for a variety of reasons use of spreadsheets is still pervasive for capturing, editing, storing and reporting information. However, there are many caveats to using spreadsheets that you must be aware of. There are also many opportunities to leverage sophisticated features within spreadsheets that can have you working faster, smarter and with greater accuracy.

Activity 1: Controlled Input

Excel is frequently used for capturing metadata.

1. Open *Activity_1.xls.* Note that there are 3 tabs, labeled *Data Entry*, *Vocabularies* and *Inherited*

2. Take a look at *Data Entry*

3. What types of controls might we want to put in place for such a form?

4. Based on outcome of conversation create and implement controls all together, using Vocabularies sheet

5. Spend 5 minutes experimenting on your own with data validation in the *Data entry* sheet.

6. Go to the tab labeled *Inherited*

7. Apply validation rules to columns in that tab

8. Select the **`Data Validation dropdown menu`** and select **`Circle Invalid Data.`** Review results.

9. As a group walk through identifying duplicates and unique values
    a. Conditional formatting, sorting and filtering
    b. Data → Remove Duplicates

10. As a group walk through counting
    a. Counting the number of occurrences of a given value (Count)
    b. Counting the number of times a field is populated with a value (Countif)

Activity 2: Importing data

CSV files are the currency of data exchange. They are used everywhere for exporting, editing, importing and reporting. They are also commonly used in such a way that degrades the integrity of the data that they carry with potentially very big implications. In this activity you will learn about CSV files and how to work with them in the proper way.

You will learn:

- The difference between opening CSVs and Importing CSVs
- What a CSV file is and how to create and edit one

Walk through together:

1. Open *Activity_2.csv* by double-clicking on it
2. **Save as** *Activity_2_open.xlsx* to the desktop and keep it open in Excel
3. Within Excel select the tab called **Data** and then select **From Text**
4. Navigate to *Activity_2.csv* and select it
5. Select **Import**
6. Select the **Delimited** radio button and then select **Next**
7. Select **Comma** and deselect any other items under the **Delimiters** section and select **Next**
8. Select all of the columns by scrolling all the way to the right within the dialog window, holding down shift key and selecting the furthest most right column.
9. Select **Text** and then select **Finish**
10. Select **New Worksheet** and then **OK**
11. **Save as** *Activity_2_import.xlsx*
12. Compare the differences between *Activity_2_open.xlsx* and *Activity_2_import.xlsx* and discuss as a group

Exploring and creating a CSV

13. Go to the desktop and navigate to *Activity_2.csv*

14. Right-click *Activity_2.csv*, select **Open With** and select **Notepad**

15. Maximize window and review in reference to the *Activity_2_open.xlsx* and *Activity_2_import.xlsx* to compare the relationship between the information represented in Notepad and the information represented in Excel.

16. Within Notepad go to **File** and select **New**

17. Type the following into Notepad including the line returns as shown:

    Place, Time, Title, Note, Rating 1, Rating 2, Rating 3
    METRO, 04:00 PM, Spreadsheet Workshop, Awesome!, 0001, 1, 1001
    Work, 09:00 AM, Monday, Bleh!, 0, , 0000

18. **Save as** *my.csv* to the desktop

19. Open and import into Excel

Activity 3: Working with existing data part 1

In this exercise you will learn how to take existing data and transform it into more structured information that enables more effective data analysis, use and management. Specifically you will learn how to:

- Arrange data in a spreadsheet according to your layout of choice
- Map data contained in one field to multiple fields
- Use a text editor as a tool for preparing and cleaning data in spreadsheets
- Keep Excel from behaving badly with numbers and dates
- Get Excel to do what you actually want it to do

Walk through:

- *Selecting multiple columns, rows, cells*
- *Using functions*
- *Typing equations*
- *Paste Special and Values*
- *Using text editors to cleanup text*
- *Using text to columns*
- *How to select from current cell to the last cell in a series*
- *Using Sort*
- *Populating a range in a dialog window by selecting*  *and selecting the cells*
- *Formatting cells (General, Text, Number)*
- *Cut and paste columns and rows*
- *Automatically sizing several columns*
- *The $ sign and it's role*

1. Go to http://en.wikipedia.org/wiki/List_of_best-selling_books#Between_10_million_and_20_million_copies

2. Copy the table (including the header row) under the heading "*Between 10 million and 20 million copies*"**.**

3. Create a new Excel workbook and paste the data into cell A1. If this does not work (It depends on a number of factors) follow the steps below. If it does work, notice that there are hyperlinks in the spreadsheet which can be problematic. Follow the steps below.

    a. Open the application called Notepad and create a new document.
    b. Paste the data from Wikipedia into the new Notepad document

c. Save the document as *authors.csv* (Note that you should select the **Unicode** encoding option when saving and replacing the .txt with .csv in the filename) to the desktop.
d. Double-click the file and it should open in Excel. If not, go to Excel and open the file from Excel, using the **Windows Icon** in the top left of the Excel Window and select **Open.**

4. Row 1 should have the column headers, *Book, Author(s),* etc. and the data should be in the rows beneath row 1.

5. **Save as** *authors.xlsx* to the desktop

Turning *Author(s)* into Last, Middle and First Names

6. Column B should be *Author(s)*. Select entire column by putting your mouse over the column letter "B" until you see a black arrow pointing downward, and then click. You should see the entire column highlighted.

7. Select **Data** and then select **Text to Columns**

8. Select **Delimited** and **Next**

9. Under **Delimiters** deselect everything and select **Space** so there is a checkbox next to it. Make sure **Treat consecutive delimiters as one** is checked

10. You should see the data split into 3 columns within the dialog window. Select **Next.**

11. You will note that at the top of each column in the dialog window the word "General" appears. This is speaking to the formatting of cells. Select all columns in the dialog window by holding down shift and selecting the right most column. When all columns are selected they should be highlighted in black

12. Select **Text** under **Column Data Format**. You will note that the "General" changes to "Text" at the top of each column.

13. **Destination** will say "*B1*". To the right of "*B1*" you will see a little icon that looks like this  . Select it and then select cell F1, which should be blank, and press Enter/Return. **Destination** should now be populated with "*F1*".

14. Select **Finish**

15. You should see the name split into Columns F, G and H.

16. In F1 replace "*Author(s)*" with "*First Name*". In G1 write "*Middle Name*". In H1 write "*Last Name*".

17. Select F1 through H1 and make them Bold text.

18. Select columns A through H

19. Select **Data** and then **Sort**

20. Make sure that **My data has headers** in the top right of the dialog window is checked.

21. In the dialog window, select the area under **Column** where it says *Sort by* and select **Last Name.** Select **OK.**

22. Starting in row 12, there should only be text in columns G and H. G12 through G54 represent instances of names which had no middle name. Select G12 through G54. Cut and paste these values into column H.

23. Select columns F, G and H. Select **Home** and then select **Cut.**

24. Mouse over the letter B for column B until a black downward arrow appears and right-click. Select **Insert Cut Cells.** First, Middle and Last name should now be columns B, C and D.

Activity 3A: Date Normalization using Text to Columns and Concatenation

This exercise demonstrates another use of **Text to Columns** for mapping data from one field to many. It also demonstrates how to map data from multiple fields to one field through use of the function called **Concatenation**.

Walk through:

- ***Concatenate***

1. Open Activity_3a.xls. Ideal formatting is ISO 8601 date format of yyyy-mm-dd.
2. Cut and paste date values from column B into column A, row 7
3. Cut and paste date values from column C into column A, row 12
4. Format all cells in column A as Date. Note that there are issues with the results of this.
5. Delete headers for column B and C
6. Select A2 through A11
7. Select **Data** and then select **Text to Columns…**
8. Select **Delimited** and then select **Next**
9. The only box that should be selected is **Other** and the value should be *"/"*. Select **Next**
10. Select all columns in the dialog window and format them as **Text**
11. Select *B2* as the **Destination** and select **Finish**
12. Swap B7 through B11 with C7 through C11 to make the B column represent the months and the C column to represent the days
13. Select A12 through A16
14. Select **Data** and then select **Text to Columns…**
15. Select **Delimited** and then select **Next**

16. The only box that should be selected is **Other** and the value should be *"-"*. Select **Next**

17. Select all columns in the dialog window and format them as **Text**

18. Select *B12* as the **Destination** and select **Finish**

19. In E1 type *yyyy-mm-dd* and make it bold

20. In E2 type *=CONCATENATE("0",B2,"-0",C2,"-",D2)*

21. Copy E2 and paste into E3 through E11

22. In E12 type *=CONCATENATE(B12,"-",C12,"-",D12)*

23. Copy E12 and paste into E13 through E16

24. Copy E2 through E16

25. Select F2, then select **Edit** followed by **Paste Special**

26. Select **Values** and then select **Ok**

27. Select Column F and format as text

28. Cut Column F and paste over Column E

Activity 4: Working with existing data part 2

In this exercise you will continue to learn how to take existing data and transform it into more structured information that enables more effective data analysis, use and management. Specifically you will learn how to:

- Transform a text-based expression of a number into its numeric expression
- Use filters to help identify variances
- Parse specific characters from a cell

Walk through:

- *Find and Replace*
- *Wildcards*
- *If/Then statements*
- *Search statements*
- *Left statement*
- *Concatenate*
- *Filter to show the value that is 11-12 million*

1. Go back to *authors.xlsx*

2. Column H should have the header *Approximate Sales*. We want to turn this into an actual number.

3. Select Column H and press Ctrl-H to bring up the Find and Replace dialog.

4. In **`Find what:`** Type *[**]*

5. In **`Replace with:`** leave it blank and press **`Replace All`**

6. In cell I2 type *=if(search("million",H2)>0,1000000,"")* and press Enter/Return. You should see the value *1000000.* This generates our base value.

7. In cell J2 type *=left(H2,2)*. This pulls the first 2 characters from the string in H2, generating the multiplier.

8. In cell K2 type *=J2*I2* and press Enter/Return. The result should be the numeric representation of the value in column H.

9. Copy I2, J2 and K2 and copy them all the way down to the end of the list.

10. Copy the data in column K.

11. Select column L and right-click. Select **Paste Special…**

12. Select **Values** and then **OK**

13. Select column L and press ctrl-1. You should see a dialog window labeled *Format Cells* appear.

14. Select **Number**. Select **Use 1000 Separator (,).** Change value in **Decimal Places** from *2* to *0*. Select **OK.**

15. Delete columns I, J and K. Cut column I and Paste it over column H.

Activity 5: Cross-referencing with other data sets

In this example we'll download a second data set and use it to answer the question: *What is the number of sales relative to the number of native speakers of the books original language?* Specifically you will learn how to:

- Perform cross-references across spreadsheets
- How to retrieve and record information based on a user-identified key/ID

Walk through:

- *Find and Replace*
- *Wildcards*
- *Linking across spreadsheets*

1. Go to http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers#Nationalencyklopedin_.282007.29

2. Copy the entire table under the header *Nationalencyklopedin (2007)*

3. Paste into Excel. Note that it pastes in correctly but with Hyperlinks. **Save as** *languages.csv* on the desktop.

4. Copy the data (not including the header) in column A and paste it into the application called Notepad.

5. Copy the data from notepad and paste it into column F in the spreadsheet containing the languages. Note that the values have pasted back in without the hyperlinks.

6. Cut the values in column F and paste them over the values in column A to overwrite the original hyperlinked data.

7. Select columns A through E. Select **`Data`** and then **`Sort.`**

8. Make sure that **`My data has headers`** in the top right of the dialog window is checked.

9. In the dialog window, select the area under **`Column`** where it says *Sort by* and select **`Language.`** Select **`OK`**

10. We would like to turn the values in column B into the correct numeric representation. Note that we couldn't do this because of the parentheticals. Select column B and press ctrl-f for find.

11. Select the tab labeled **Replace.** Type " *(***)"*, noting the space before the open parentheses, next to **Find What.** Don't type anything in **Replace with** because we want to simply delete. Select **Replace All**. Note that because we selected column B before entering into this dialog that "Replacing All" will only work within column B. Note that the parenthetical numbers were deleted.

12. In column F type *=B3*1000000*. This will create the full numeric representation of the number in column B.

13. Copy and paste down to all rows where there are data.

14. Copy the data in column F and **Paste Special, Values** into column G Format as a number with separators and no characters to the right of the decimal point.

15. In the *Authors*.xls spreadsheet column I should be the first available column. In I1 type "*Native Speakers in the World*" and make it bold.

16. Position the *languages.csv* spreadsheet where you can easily get to it. You will be switching back and forth between the two spreadsheets. You may need to resize and/or reposition your windows to make it easy.

17. In Authors.xls I2 type *=vlookup(F2,*

18. Without selecting any other cells switch over to the languages.csv spreadsheet and select the full table (A2 through E101) and type *,2,FALSE)* and press Enter/Return.

19. In *Authors.xls* Copy and paste this down all rows in the table

20. Looking back at *Authors.xls*, why did we get *#N/A* in some of the cells in column I? Think about why this is and what you might do to get around this. Be prepared to discuss.

21. In cell J2 type *=H2/I2* and hit Enter/Return

22. Go to **Home** and where it says *General* with a dropdown menu select **Percentage.**

23. Copy and paste this formula down the entire table.

24. Create a new spreadsheet and copy and paste columns F, H, I and J into a new spreadsheet.

25. Why do you get *#REF!* in columns C and D? How might you remedy that? See what you can do to figure it out.

Other

- Ways of identifying unique and duplicate values in a list
    - Conditional Formatting followed by sorting and filtering by color/font
    - Data → Remove Duplicates
- Transposing
- If/Then for Quality Control (using sorting and filtering by color/font)
- Counting the number of occurrences of a value (countif) or the number of times a field has a numeric value (count)