Excel Tips

for handling large spreadsheets of collections data

Melissa Belvadi
University of Prince Edward Island

October 2017

Excel Spreadsheet Vocabulary and Limits

- Spreadsheet = Workbook = Book = Entire File
- Sheet = Tab = Worksheet = one table within a file

Spreadsheet Limitations Excel 2007-2016:

Feature	**Maximum limit**
Total number of rows and columns on a worksheet	1,048,576 rows by 16,384 columns
Column width	255 characters
Total number of characters that a cell can contain	32,767 characters
Hyperlinks in a worksheet	66,530 hyperlinks

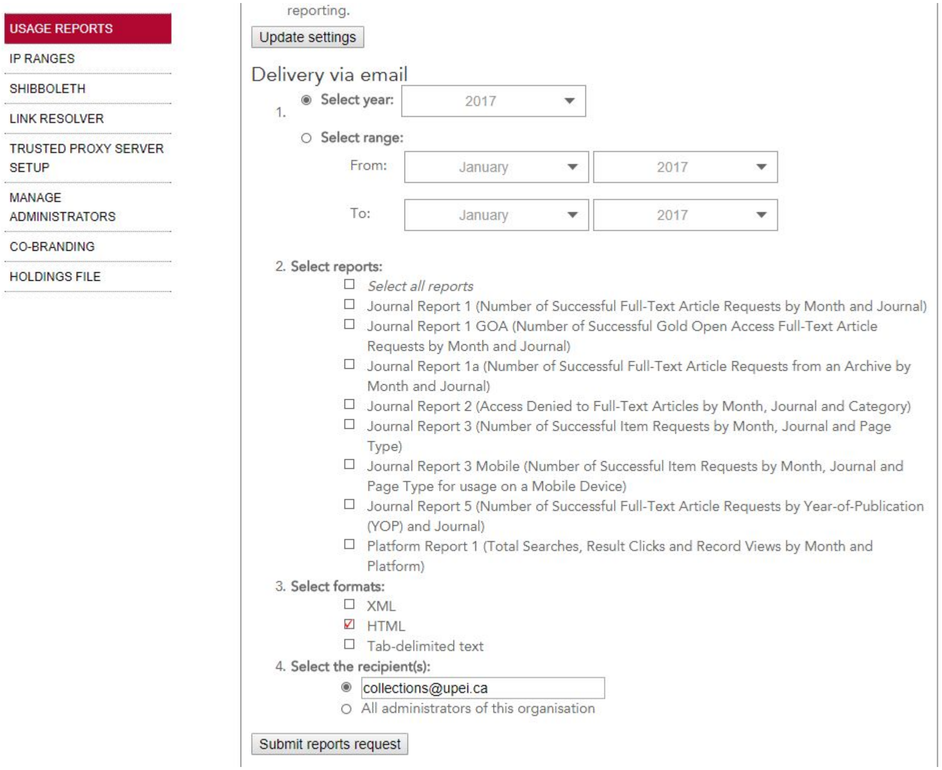
Best Practices - File Management

- Always (!) make a copy of original data file, save in separate folders
- Naming convention - name originals and copies in pattern, taking sort order into account, for example:
  - T&F COUNTER JR1 2016
  - an T&F COUNTER JR1 2016

    - start with "an" ("Analysis") so see it first if names truncated - don't accidentally edit original

Best Practices - Export Data from Vendor

- Prefer tsv (tab separated values) first, then csv (comma separated values) over "Excel"
  - why not Excel - mangled headers, merged cells
  - why tsv - tab separated values - never possible in data
  - csv - can get corrupted if comma in actual data
  - sometimes what vendor claims is "Excel" is actually tsv or csv anyway
- tsv exports often named .txt - if Google Sheets, fix extension to .tsv first
- COUNTER reports - one at a time - shortcut to select multiple is a trap!



reporting.

Delivery via email

1. ● Select year: 2017
   ○ Select range:
   From: January 2017
   To: January 2017

2. Select reports:
   - □ *Select all reports*
   - □ Journal Report 1 (Number of Successful Full-Text Article Requests by Month and Journal)
   - □ Journal Report 1 GOA (Number of Successful Gold Open Access Full-Text Article Requests by Month and Journal)
   - □ Journal Report 1a (Number of Successful Full-Text Article Requests from an Archive by Month and Journal)
   - □ Journal Report 2 (Access Denied to Full-Text Articles by Month, Journal and Category)
   - □ Journal Report 3 (Number of Successful Item Requests by Month, Journal and Page Type)
   - □ Journal Report 3 Mobile (Number of Successful Item Requests by Month, Journal and Page Type for usage on a Mobile Device)
   - □ Journal Report 5 (Number of Successful Full-Text Article Requests by Year-of-Publication (YOP) and Journal)
   - □ Platform Report 1 (Total Searches, Result Clicks and Record Views by Month and Platform)

3. Select formats:
   - □ XML
   - ☑ HTML
   - □ Tab-delimited text

4. Select the recipient(s):
   - ● collections@upei.ca
   - ○ All administrators of this organisation

Demo: an TF combined report

Best Practices - Import tsv or csv

- Two choices: open directly, selecting Excel as "open with" OR
- Open empty workbook first, then Data - Get External Data - From text

Import Wizard allows you to change data type of columns like ISSN and ISBN to "text" during import, saving from having to fix later and restore lost leading zeros.

Excel 2016 for Windows may use Query Editor instead of Import Wizard but same idea

IMMEDIATELY after import complete, "Save As" Excel format file

- watch location and filename

Get External Data creates internal link to original data file - hard to break

2016 may immediately import as a Table...

Sheets versus Tables

"Tables" in Excel:

- convert to table: Home - Styles - Format as Table (or ctl-L)
- convert to normal: Table - Design - Convert to Range
- Advantages:
    - Assumes header row for you and turns on Filter automatically
    - Bands the rows into alternate shading for you
    - Excel will autofill formula columns for you
    - Some functions/add-ons may require your data to be in a Table (Fuzzy Lookup)
- Disadvantages : those first two - clutters display

Demo: Sciencedirect\an-jr1 file

Best Practices - First Edits to Sheet

- Convert to/from Table as you prefer
- Rename sheet to something like "main" or "raw" or other specific ("JR1")
- Want first row = column headers, nothing else
- All other header info - create new sheet, copy info including summary totals, rename sheet to something like "JR1 header", delete those rows from "JR1"
- Check bottom for totals, also move that data to other sheet
- Freeze top row: View - Freeze Panes - Freeze top row
- Wrap text on header row if needed
- Add new first column - "original order" - 1,2,3 - use autofill to complete down

Demo: Sciencedirect\an-jr1 file

About Autofill

- "Copy cells" versus "fill series" - need two numbers started to indicate series
- Formulas and Autofill
  - Will change referenced cells in same direction as autofill, e.g. if column C is formula:
  - =a1+b1
  - autofill down column C will make row 2 =a2+b2
  - autofill into column D will make d1 =b1+c1
  - To set a row or column reference so autofill won't change it, put a $ in front of part to set
  - =$b1+$c1
  - =b1+c1 - this will be exactly the same everywhere autofilled - use to hard-reference exact cell but will change if rows/columns added to follow the data
- Example: apply percentage increase to entire column with autofill, in blank column:
  =a2*1.05 - will add 5% more to all of the column A values

Demo: Sciencedirect\an journal print price list

Basic numeric functions

SUM, SUMIF

=sum([range of cells]) =sum(a2:a254)
but also =sum(a:a) (don't put this anywhere in column A)

=sumif(range,criteria,sum_range)
=sumif(a2:a4,"=0",b2:b4)
but also
=sumif(a:a,">1999",b:b)   [the criteria can be just a text string too]
You can leave out the sum_range if the range itself is what you want to add

COUNT, COUNTA, COUNTIF

=count(a1:a200) - counts number of cells with numeric values - 0 will count but blank will not
=counta(a1:a200) - counts anything non-blank, use for text
=countif(a1:a200,">1999")

Basic text string functions

=left(a2,3)   first 3 characters of cell a2

=right(a2,3)   last 3 characters of cell a2

=mid(a2,4,3) the characters in cell a2 starting with the 4th for a length of 3 (4,5,6)

Use double-quotation marks for literal characters in formulas involving strings

Joining strings with different formulas using &

    =left(a2,4)&"-"&right(a2,4)

    will take the text: 1234567X and return back 1234-567X

=len(a2) Length of the string, "1234x" will return 5

Basic text string functions - IF and related

```
=IF(criterion,value-if-true,value-if-false)
=if(left(a2,3)="978","this is a 13-digit ISBN","this is a 10-digit ISBN")
```

some functions useful in IFs

<> means not equal to: =if(a2<>"","a has a value","a is blank")

isblank(a2) = returns TRUE if the cell is empty

NOT() - reverse the logic of any criterion

AND(), combine several criteria:
=if(AND(a2="",b2<>""),"no start date",a2&"-"&b2)

OR(), like AND but only one has to be TRUE

Note that =, <> on strings is case-insensitive: a will equal A

Basic text string functions - SEARCH and related

Search: find a string and return its starting position

=search(pattern, cell to search within [, starting position])
=search("login?url=",a2)

Example: find a specific pattern in ezproxy log files (use case: small publisher not COUNTER compliant, but has consistent URL syntax for PDF calls)

=iferror(search("http://www.hh.um.es:80/pdf",a2),"")

Then =count(a:a) to find the total

Demo: ezproxy\ez2017-sample-raw
=iferror(SEARCH("/reserves/Psyc",G2),"")

Data Types

Numbers, Text, Dates, others

- What looks like a number may not be stored as one
- Data that you intend as text (ISBN) may be imported as number, loses leading zeros
- Tell numbers versus text by default horizontal alignment - numbers right, text left (dates right-align)
- Verify with test column using formula, autofill, and sort or filter:
  =type(a2)   [1=number, 2 = text, 4 = true/false, 16 = error]
- Dates are stored as numbers (type 1), displayed as dates as you choose
- When entering data by hand that you want as text, start with apostrophe '

Demo: sciencedirect\an sd-sample numeric-text books tab

Convert data types - ISBN

From number to text, restoring leading zeros, ISBN:

```
=text(a2,"0000000000000")
```

If mix of 10-digit and 13-digit ISBNs:

```
=if(left(text(a2,"0"),3)="978",text(a2,"0000000000000"),text(a2,"0000000000"))
```

If we convert a2 to text without any padding, then check the first 3 characters, and if they are "978" we are going to convert a2 to text with padding to 13 digits, otherwise we convert a2 to text with padding to 10 digits

Demo: sciencedirect\an sd-sample numeric-text books tab

Normalize ISBNs - remove all hyphens

Don't use find/replace - will turn values into numbers instead of text!

Instead use formula SUBSTITUTE:

=substitute(a2,"-","")   [change all instances of hyphen with nothing]

Convert data types - ISSN

Always normalize ISSNs to include the hyphen which guarantees it can't be a number.

From number to text, restoring leading zeros, ISSN:

```
=text(a2,"0000-0000")
```

If value is already text, this will leave it alone, which is a problem for 0123034X

so final formula for column autofill is:

```
=if(type(a2)=2,left(a2,4)&"-"&right(a2,4),text(a2,"0000-0000"))
```

If a2 is text, I want the first 4 characters, then a hyphen then the last 4 characters, otherwise convert a2 to text formatted as 0000-0000

Demo: sciencedirect\an sd-sample numeric-text journals tab

Best Practice - Convert Formulas to plain text

After normalizing data, convert entire column from formula to direct values

Column select - right-click Copy, Right-Click Paste Special - Values

Why?

- Once fixed, only risk of accidental change, no benefit
- Performance drag on large spreadsheets

Demo: sciencedirect\an sd-sample numeric-text
journals tab

Converting text dates to true dates

How to tell: dates are right-aligned, selecting multiple ones will show average/sum at bottom

If date string is fully padded, e.g. 01/31/1999:

Month column: =left(a2,2)
Day =mid(a2,4,2)
Year =right(a2,4)

Date formula wants (year, month, day)

True date value is   =DATE(right(a2,4), left(a2,2), mid(a2,4,2))

Useful date formulas

When have true date, not "date as text"

determine day of week

=weekday(a2)
returns number: 1 = Sunday through 7 = Saturday
For readable text:
=text(weekday(a2),"ddd") will convert 1 to "Sun"
=text(weekday(a2),"dddd") will convert 1 to "Sunday"

month of year - same but use "mmm" for "Jan" or "mmmm" for "January"

Demo: sciencedirect\an sd-sample numeric-text books tab

Convert text to date if not fully padded

Depends on syntax

Problem would be something like 1/31/2007 and 10/2/2007 in m/d/yyyy format

Sometimes "datevalue" can figure it out =datevalue(a2)

Otherwise may need to use "text to columns", then convert each number and combine. If split on "/" end up with month in D, day in E, year in F:

```
=date(text(f2,"0000"),text(e2,"00"),text(d2,"00"))
```

Clean up hidden characters

Trim - removes leading and ending spaces and extra spaces between words

e.g. the value " Journal of   Abnormal Psychology "
will become "Journal of Abnormal Psychology"

Clean - removes line breaks and non-printable characters

If you suspect these problems with a text column, do a replacement column combining both just to be safe:

=clean(trim(a2))

Note that Clean will NOT remove accented characters, no worries

LC Call Number - extract the LC class

Painful formula:

```
=if(iserror(value(mid(d2,3,1))),left(d2,3),if(iserror(value(mid(d2,2,1))),left(d2,2),left(d2,1)))
```

Logic: if the 3rd character is not a number, then use the first 3, otherwise if the second character is not a number, then take the first 2, otherwise use just the first character.

function iserror(formula) returns true/false

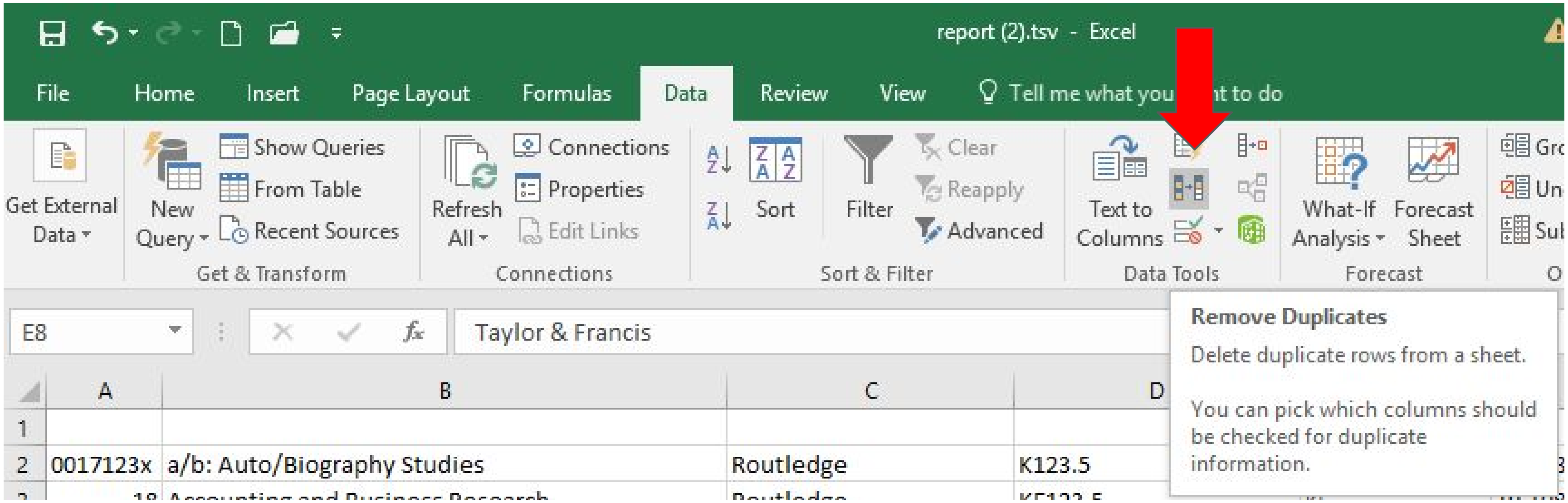
function value returns an error if applied to a non-digit

Demo: circulation\an Circulation activity A-F...

Removing duplicate rows

Excel has function built in - Data - Data tools
Allows you to select which columns to consider if not all

Will remove immediately, no option to inspect



Remove Duplicates

Delete duplicate rows from a sheet.

You can pick which columns should be checked for duplicate information.

	A	B	C	D
1				
2	0017123x	a/b: Auto/Biography Studies	Routledge	K123.5

Demo: circulation\an duplicates Circ activity A-F...

Sorting

Make sure ALL columns have headers

Make sure you have no row or column selected, so it will sort everything (otherwise always "expand selection" when prompted to)

Home - Sort & Filter - Custom Sort - to sort on multiple columns, e.g. first by the LC class, then within that grouping, by the pubyear

Demo: circulation\an duplicates Circ activity A-F...

Filters

Highlight your entire header row, then:
Home - Sort & Filter - Filter
Pulldown arrow next to first column to Filter by
Unclick "Select all"
Either use Text filters or use search to select matching basic pattern

Applying a filter on a second column is like an "AND" on the first one

Demo: circulation\an duplicates Circ activity A-F...

Cleaning Up Bad Data - Basics

Use extra columns as needed to catch invalid values:

=type(a2) - if expect all numbers, make this column then sort/filter to find anything that's not a "1"

=len(a2) if expect all strings to be a certain length (eg 13-digit ISBNs that have been normalized to text), make this column then sort to find less than or longer than 13

Usually you would delete these columns after removing/fixing the bad data

Vlookup - Combining data from different lists

Must get tables into same file/workbook.

Have both open.
Right-click sheet tab - Move or copy…

Order of sheets doesn't matter.

Be sure to select checkbox "create a copy"!

Rename the sheet copied to a nice simple name, no spaces, punctuation

Demo: springer\ combine two files

Vlookup

Need an "index" column that is basis for match - usually ISBN, ISSN, vendor ID.

Titles make very poor matches even from same vendor.

Be sure your index column is normalized the same way in both sheets.

Decide which sheet is getting the data from the other. The one getting the data is the target. The one getting looked-up from is the source.

Source: make sure the index column comes BEFORE the desired data to lookup.

Demo: springer\ combine two files

Vlookup

Source ("prices")- column A has the ISSN, column F has the price

Target ("main") - column D has the ISSN, you've opened blank column E to put the price from the source

In cell E2, **=vlookup(D2,prices!A:F,6,0)**

The "6" means report back the 6th of the range of columns (which is what F is relative to A)
The "0" means only report an exact match - will return an #N/A error if no match

Better:

=iferror(vlookup(D2,prices!A:F,6,0),"no match") [or whatever you want if no match]

Demo: springer\ combine two files

Vlookup with wildcards

Example use case: if have multiple ISSNs in single column like in CRKN JUP report

Works if the target value to match is a substring of the source

=vlookup("*"&D2&"*",prices!A:F,6,0)

Note: this solves problem of having two ISSN columns and not sure which to match on. First combine both in one column =f2&"; "&g2 then do wildcard match

Demo: sciencedirect\..combined - show JR1 then UPEI(2)

Vlookup with wildcards and dealing with two ISSNs

Note: this solves problem of having two ISSN columns in BOTH sheets and not sure which to match on. First combine both in one column C =a2&"; "&b2 then do wildcard match on that.

If have two ISSNs in BOTH documents and not sure which might match to which, combine using nested IF and the method above:

1. combine as above in the source document where C is combined column and F is still the prices you want to pull into the main sheet
2. in the target document if ISSN columns are D and E, do:

```
=iferror(vlookup("*"&D2&"*",prices!C:F,4,0),iferror(vlookup("*"&E2&"*",prices!C:F,4,0),"no match")
```

Pivot Tables

Gets Excel to summarize lots of raw data for you - most often sum and count
A critical tool for Data Viz of large data sets

Select all of the columns you want involved in the pivot first

Insert - Pivot Table - accept defaults - range already selected, create new worksheet

Drag column header for Rows first, then Columns if any, then Values (may be same as Rows)

Decide if you want Values to be Summed or Counted (always counted if text!) - default is count, pulldown arrow - Value Field Settings to change to Sum or other

Pivot table as cleanup tool

Simple pivot table on column of suspicious data, e.g. publisher names

Quickly exposes variants, e.g.
John Wiley & Sons
John Wiley & Sons, Inc.
John Wiley & Sons, Incorporated
John Wiley & Sons, Ltd
John Wiley & Sons, Ltd.

Filter on this column in main sheet, use autofill copy to fix, check pivot for more until pivot list is 'clean'

Demo: an ebook cleanup publisher

Pivot Tables - more

If cross-tab two variables, like LC Class and shelving location, make the one with the most possible values be rows - very hard to deal with lots of columns

Can also subdivide rows instead of using columns - best only two levels - add second variable/column to Rows below the first

Filter rows if want to only include certain values - especially to get rid of "(blank)" row.

Group rows - useful for creating subtotals by decade eg of publication year of books or grouping LC class values - but easier if you do as separate column in the actual data first then pivot on that

Use first column header right-click to "refresh" as main data is changed. Not affected by main data "filter".

Pivot Tables - more

Can add same values multiple times to show count, sum, average, etc. as different calculated columns

Can edit the labels, headings, etc. in the formula bar at top

Can copy-paste entire pivot table:

- Grouping/Ungrouping changes affect both even on different sheets
- Changes to columns included, filters, values, independent

If copy-paste-special, only displayed lines/groups will copy

Demo: circulation\an circulation...