



# Excel for Data Cleaning

What is a spreadsheet?

A tool for holding data in rows and columns.

Why use spreadsheets?

For disposable data-cleanup or visualization...



# Agenda

---

What are Spreadsheet Tools?

**01**

---

Getting Data into a Spreadsheet Tool

**02**

---

Summarizing Data in Excel

**03**

---

Manipulating/Mangling Data in Excel

**04**

---

Getting Stuff Out of Excel

**05**

---

Recap & Alternatives to Excel

**06**



# Examples of Spreadsheet Tools



**Excel** is Microsoft's standalone **spreadsheet** program. Available with the Office suite.



**LibreOffice Calc** is the "Document Foundation" free open-source spreadsheet program. Available here: [www.LibreOffice.org](http://www.LibreOffice.org)



**Google Sheets** is part of the Google Apps Suite. If you have a Gmail and Google Drive, you have it.

# What Can/Can't Excel Do?

	Excel	LibreOffice Calc	Google Slides
Formulas	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Charts & Graphs	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Auto-formats fields	<input type="checkbox"/>		<input type="checkbox"/>
Allow creation of lookup lists	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Allow character sets besides Latin1 (e.g., UTF-8)	Limited	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Track changes after closing	Limited	Limited	<input checked="" type="checkbox"/>
Maintain integrity between rows and columns	Limited	Limited	Limited
Enforce referential integrity between different sheets	Limited	Limited	Limited



# Spreadsheet Don'ts

What should you **AVOID** doing in a spreadsheet?

- Mixing multiple datasets in a single spreadsheet
- Using multiple tabs in a workbook
- Not filling in zeros
- Using problematic null values (e.g., -999, +/-, Null, NA)

- Using visual formatting [color-highlights, fonts, borders] to convey information
- Using visual formatting to make the data sheet look pretty

- Placing comments or units in cells
- Entering more than one piece of information in a cell
- Using problematic field names

- Using special characters in data (e.g., line breaks, em-dashes, quotation marks)
- Inclusion of metadata in data table

**P.S. DO NOT MERGE CELLS. DO NOT. BAD. SHAME.**



# Tables vs. Spreadsheets

## Tables:

- = records & attributes
  - = actual structure
  - = *safer data...*

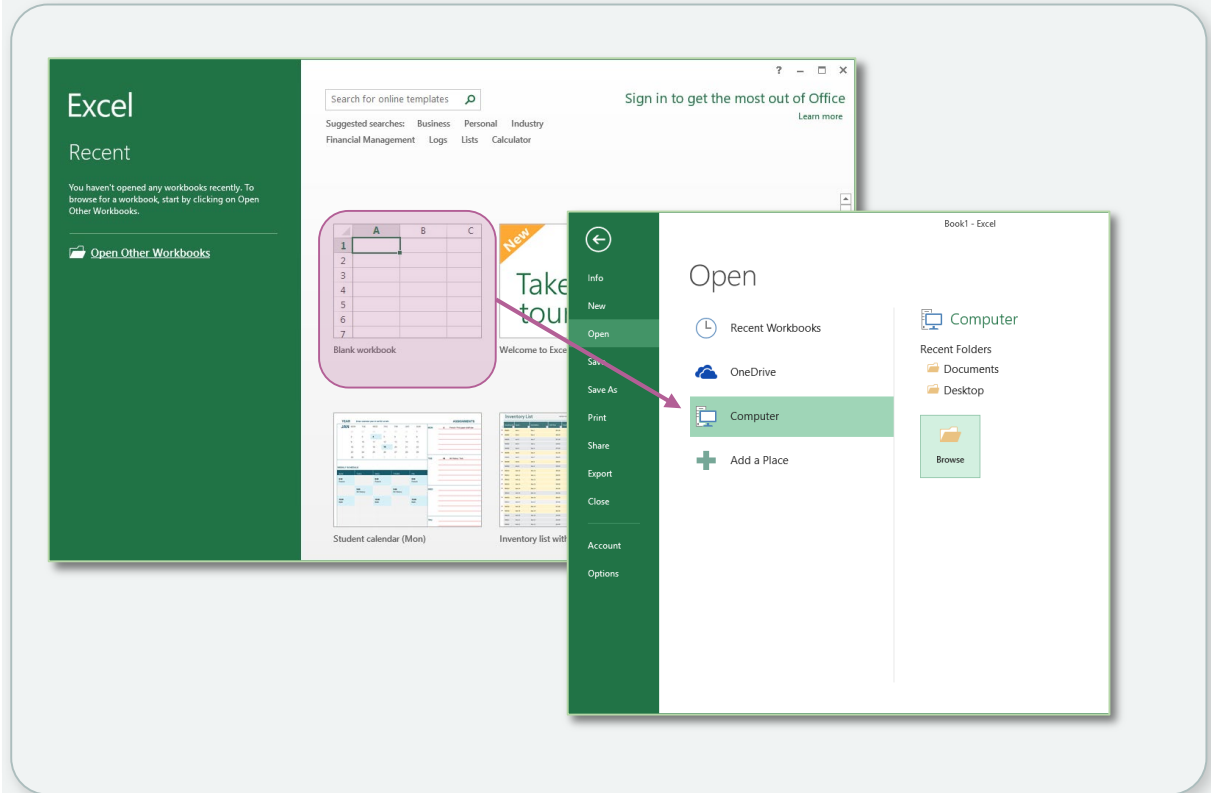
## Spreadsheets:

- = rows & columns
  - = illusion of structure
  - = *accident-prone data...*
  
- 1 Excel file can hold multiple spreadsheets = more *illusion...*

# Getting Data into Excel

[the not-safe way]

- File → Open...



# Getting Data into Excel

[the not-safe way]

- File → Open...
- can **subtly mangle** CSV data.

The image shows a spreadsheet with columns D, E, F, and G. Column D contains dates, column E contains event codes, column F contains measurement fractions, and column G contains measurement types. Two callouts highlight parsing issues: an orange callout points to the date column, and a purple callout points to the fraction column.

D	E	F	G
AdmDateModified	StaEventC	MeaMeasurementFraction	MeaMea:
1/5/2017	SPE-279	2/3/2004	height
1/5/2017	SPE-279	6/15/2016	depth
1/5/2017	SPE-279	9/3/2016	width
1/5/2017	SPE-279		weight
12/1/2016	RHB-67-B	3/3/2008	height
12/1/2016	RHB-67-B	2/3/2008	depth
12/1/2016	RHB-67-B	1/15/2016	width
12/1/2016	RHB-67-B		weight
12/1/2016	RHB-67-C	3/3/2008	height
12/1/2016	RHB-67-C	2/7/2008	depth
12/1/2016	RHB-67-C	1/3/2004	width
12/1/2016	RHB-67-C		weight
12/1/2016	RHB-64-C	7	height
12/1/2016	RHB-64-C	8-Jan	depth

Dates!  
But are they D/M/Y or M/D/Y...

Not dates!  
These should be fractions.

# Getting Data into Excel

## The Safer Way

Import it as External Data...

1. Click the **From Text** option in the **Get External Data** group of the **Data** tab.
2. Follow three steps in the **Text Import Wizard**
3. The data should now display properly.

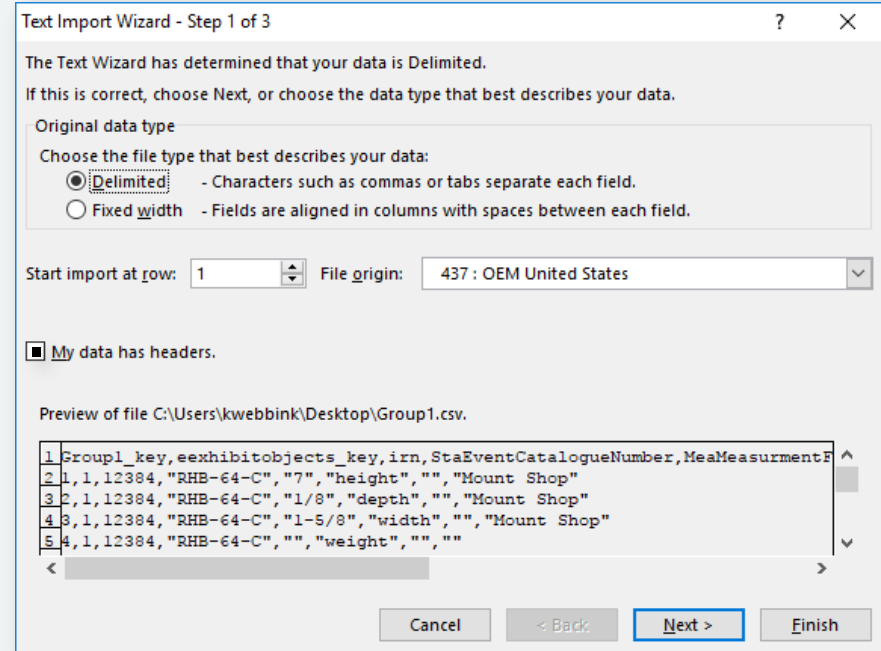
The screenshot illustrates the process of importing data from a text file into Excel. The 'Data' tab is active, and the 'From Text' option is selected. The Text Import Wizard is open, showing the 'Delimited' option chosen for the file type. The preview shows a CSV file with columns: IDP number, Cat Numb, Accession year, ACC\_N, Former number, count in lot, Specim. The resulting table in Excel is as follows:

AdmDateModified	StaEventCat	MeaMeasurementFractio
2017-01-05	SPE-279	2-3/4
2017-01-05	SPE-279	6-15/16
2017-01-05	SPE-279	9-3/16
2017-01-05	SPE-279	
2016-12-01	RHB-67-B	3-3/8
2016-12-01	RHB-67-B	2-3/8
2016-12-01	RHB-67-B	1-15/16
2016-12-01	RHB-67-B	
2016-12-01	RHB-67-C	3-3/8
2016-12-01	RHB-67-C	2-7/8
2016-12-01	RHB-67-C	1-3/4
2016-12-01	RHB-67-C	

# Text Import Wizard

## Step 1 of 3

- Original data type:
  - “Delimited” for CSV files
  - Check “My data has headers”
- Check the preview:
  - ...Does data look mangled?  
(e.g.:  $\diamond$ ,  $\square$ , ?)
  - ...If so, check the **File origin** and try “65001 Unicode (UTF-8)” [requires Excel 2007 or later]
  - Double-check the Preview...



# Getting Data into Excel

The safer way – but beware

Data files (CSV, XLS, TXT, etc.) are encoded in a specific character-set.

- Most commonly they are either **Latin-1** or **UTF-8**.
- **Pre-2007 versions of Excel only read/write Latin-1.**

# Text Import Wizard

## Step 2 of 3

1. Select **Comma** from the **Delimiters** options for CSV files
2. Select “” from the Text qualifier options

Again, check the **Data preview**:

- Does data look mangled (columns run together)?
- If so, try different delimiters.

Double-check the **Data preview**.

Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

**Delimiters**

Tab

Semicolon

Comma

Space

Other:

Treat consecutive delimiters as one

Text qualifier: "

**Data preview**

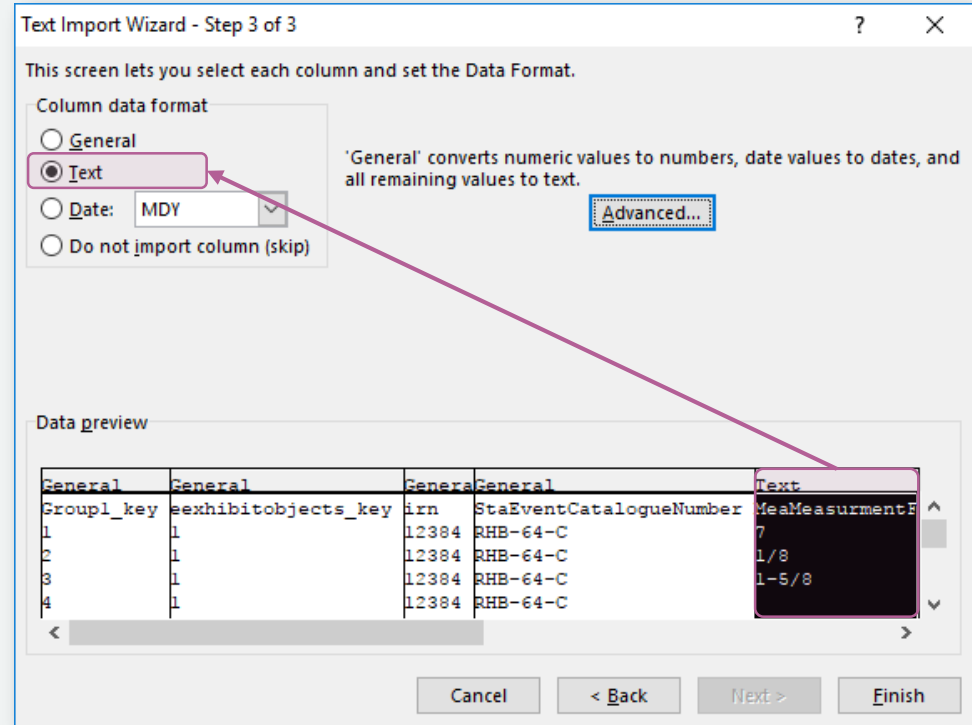
Group1_key	exhibitobjects_key	irn	StaEventCatalogueNumber	MeaMeasurementF
1	1	12384	RHB-64-C	7
2	1	12384	RHB-64-C	1/8
3	1	12384	RHB-64-C	1-5/8
4	1	12384	RHB-64-C	

Buttons: Cancel, < Back, Next >, Finish

# Text Import Wizard

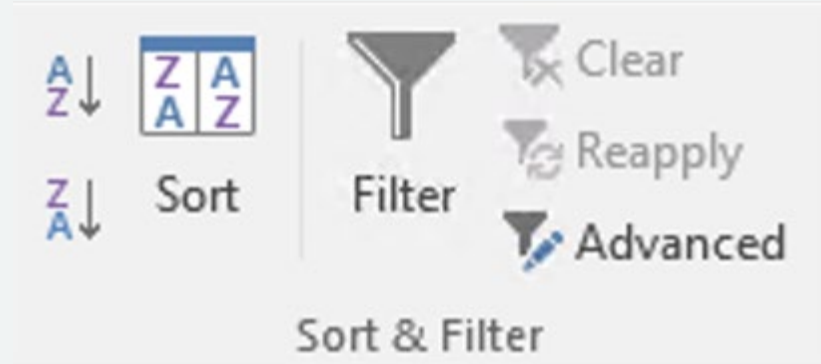
## Step 3 of 3

1. Make sure to check each column's data-type (Column data format).
2. If a column appears to be specially-formatted numeric data (dates, fractions, etc.):
  - a. Click on the column in the Data preview to select it
  - b. Change its data format to Text.



# Summarizing Data in Excel

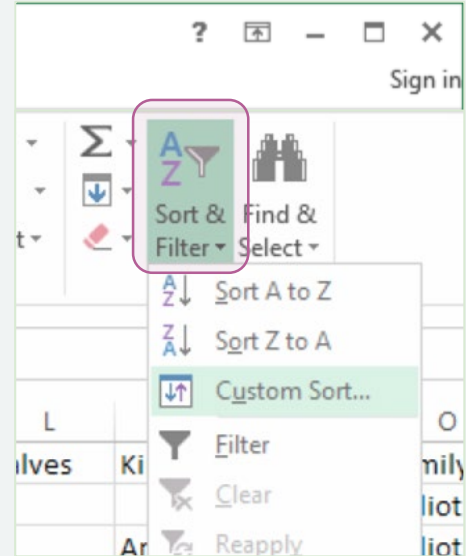
Sorting and Filtering



# Summarizing Data in Excel

## Sorting

- Beware of grabbing only part of your dataset
- Beware of one-way trips...to bad places...



# Sorting Errors

Before Sort:

L	M	N	O	P	Q	R	S	T
Valves	Kingdom	Superfam	Family	Subfamily	Genus	Subgenus	Species	Subspecie
		Pleurotom	Haliotidae		Haliotis		cracherodii	
	Animalia	Pleurotom	Haliotidae		Haliotis		ancile	
	Animalia	Trochoide	Trochidae	Trochinae	Clanculus		puniceus	
	Animalia	Trochoide	Trochidae	Calliostom	Calliostoma		ligatum	
	Animalia	Trochoidea			Unidentified			
	Animalia	Trochoide	Trochidae	Gibbulinae	Cittarium		pica	
	Animalia	Trochoide	Trochidae	Monodon	Monodonta		labis	
	Animalia	Trochoide	Trochidae	Monodon	Tegula		mariana	
	Animalia	Lymnaeoi	Lymnaeidae		Lymnaea		stagnalis	
	Animalia	Cerithioid	Vermetidae		Dendropoma		irregularis	

After Sort:

L	M	N	O	P	Q	R	S	T
Valves	Kingdom	2	Family	Subfamily	Genus	Subgenus	Species	Subspecie
		10	Haliotidae		Haliotis		cracherodii	
	Animalia	10	Haliotidae		Haliotis		ancile	
	Animalia	10	Trochidae	Trochinae	Clanculus		puniceus	
	Animalia	10	Trochidae	Calliostom	Calliostoma		ligatum	
	Animalia	10			Unidentified			
	Animalia	Acavoidea	Trochidae	Gibbulinae	Cittarium		pica	
	Animalia	Acavoidea	Trochidae	Monodon	Monodonta		labis	
	Animalia	Achatellin	Trochidae	Monodon	Tegula		mariana	
	Animalia	Achatellin	Lymnaeidae		Lymnaea		stagnalis	
	Animalia	Achatinell	Vermetidae		Dendropoma		irregularis	

If you sort without expanding the selection, only the highlighted column will reorder, leading to mismatched data that is no longer connected to its record.

# Summarizing Data in Excel

## Filtering


The screenshot shows an Excel spreadsheet titled 'WorkshopDataset - Excel'. The ribbon is set to the 'HOME' tab. The 'Sort & Filter' button is highlighted in the ribbon, and a dropdown menu is open, showing the 'Filter' option selected. The spreadsheet data is as follows:

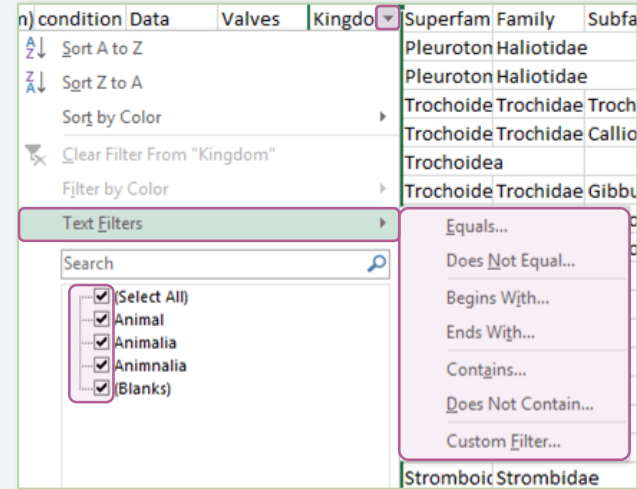
	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Type	Size (mm)	condition	Data	Valves	Kingdom	Superfam	Family	Subfamily	Genus	Subgenus	Species	Subspecie	Full na
2	41	92		None		Animalia	Pleuroton	Haliotidae		Haliotis		crache		
3	41	34		None		Animalia	Pleuroton	Haliotidae		Haliotis		ancile		
4	41	15		None		Animalia	Trochoide	Trochidae	Trochinae	Clanculus		punic		
5	41	15		None		Animalia	Trochoide	Trochidae	Callioston	Calliostoma		ligatu		
6	41	20		None		Animalia	Trochoidea			Unidentified				
7	41	55		None		Animalia	Trochoide	Trochidae	Gibbulina	Cittarium		pica		

- Select the desired column by clicking on its header, e.g., **M**.
- Click the **Sort & Filter** button in the **Home** or **Data** tab and select **Filter** from the dropdown menu.

# Summarizing Data in Excel

## Filtering

- Click the dropdown button  and use the checkboxes to select and exclude values.
- Hover over the **Text Filters** to change the text filter condition from the default **Equals...**





# Manipulating/Mangling Data in Excel

## Functions

### Build your own:

- concatenate
- vlookup

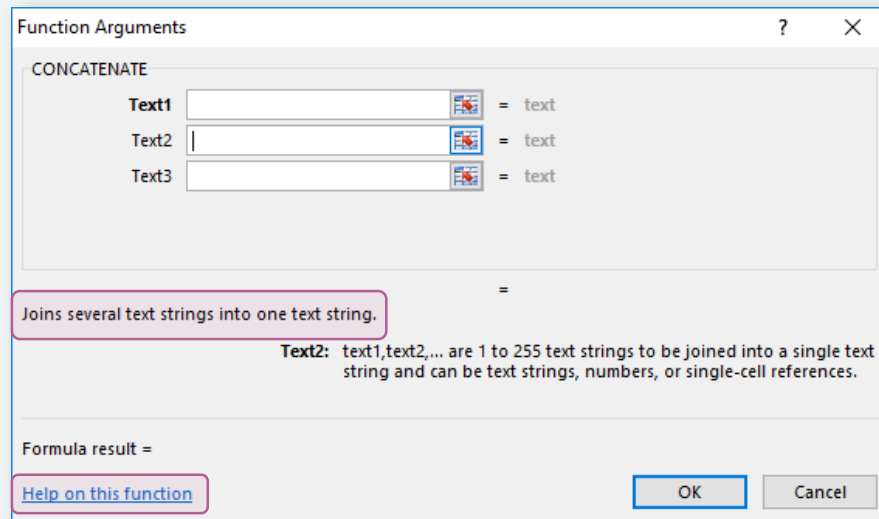
### Built-in:

- Split Text-to-Columns
- Deduplication

# Functions

## Build your own: CONCATENATE

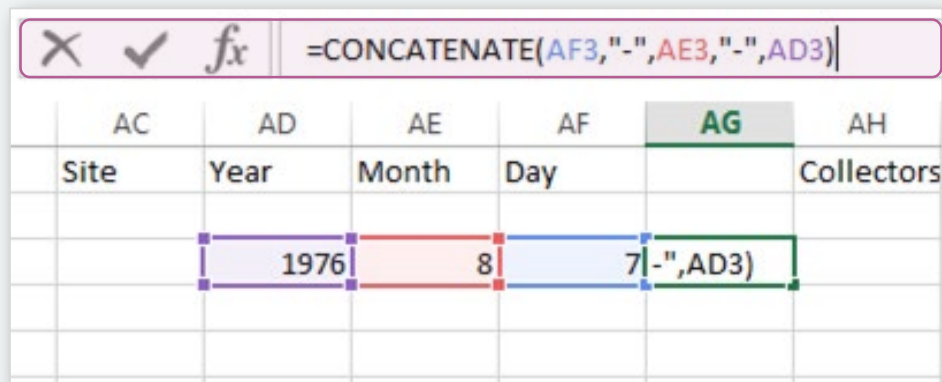
- The general syntax for function is:  
`=FUNCTIONNAME(argument1, ...)`
  - The Function wizard takes care of the syntax, with users needing only to input the arguments.
  - The wizard offers built-in help and a link to a help page.



# Functions

Build your own:  
CONCATENATE

=CONCATENATE(arguments)



The screenshot shows an Excel spreadsheet with the following data:

	AC	AD	AE	AF	AG	AH
	Site	Year	Month	Day		Collectors
		1976	8	7	=CONCATENATE(AF3,"-",AE3,"-",AD3)	

The formula bar at the top shows the formula: `=CONCATENATE(AF3,"-",AE3,"-",AD3)`. The cells containing the values 1976, 8, and 7 are highlighted with colored boxes (purple, red, and blue respectively) to show they are the arguments being concatenated. The result of the formula is shown in cell AG3.

# Functions

## Built-in: Split Text → Columns

The screenshot shows the Excel Data ribbon with the 'Text to Columns' button highlighted in a pink box. Below the ribbon, a data table is visible with columns for 'Specimen identifier's name', 'Type', 'Size (mm)', and 'condition'. A 'Text to Columns' dialog box is open, showing instructions on how to split text into multiple columns.

The screenshots show the 'Convert Text to Columns Wizard' steps. Step 1 of 3 shows the wizard determining if the data is delimited. Step 2 of 3 shows the user selecting a delimiter (comma) and previewing the data. Step 3 of 3 shows the user selecting the column data format (General) and previewing the data.

	AJ	AK	AL
Specimen identifier's name			
Heiser	J.		1993
Heiser	J.		1993
Heiser	J.		1993
Heiser	J.		1993
Heiser	J.		1993
Heiser	J.		1993
Per label supplied			
Heiser	J.		1995
Heiser	J.		1995
Heiser	J.		1993
Heiser	J.		1993
Heiser	J.		1993
Heiser	J.		1993
Heiser	J.		1993

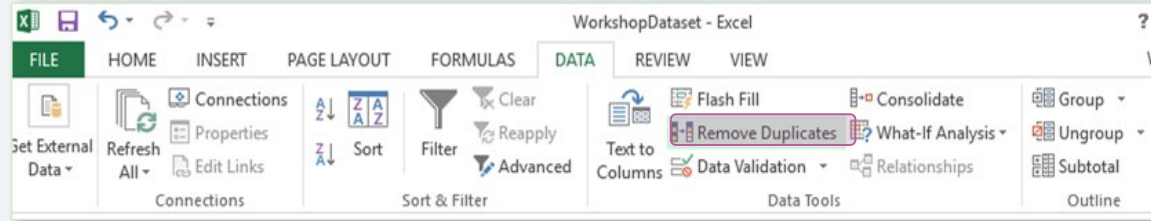
Select **Text to Column** from the **Data** menu.

Follow the steps in the **Text to Column Wizard**. Make sure the right delimiter is selected (typically a , . “ ” or ;) by previewing the data in the bottom window.

# Functions

## Built-in: Deduplicating Rows

	A	B
1	ADP num1	Cat Num
2		1 ABCD:1
3		2 ABCD:2
4		3 ABCD:3
5		4 ABCD:4
6		5 ABCD:5
7		6 ABCD:6
8		7 ABCD:7
9		8 ABCD:8
10		9 ABCD:9
11		10 ABCD:10
12		11 ABCD:11
13		12 ABCD:12
14		13 ABCD:13
15		14 ABCD:14
16		15 ABCD:15

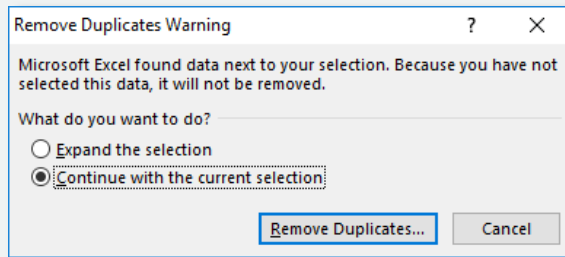


To deduplicate rows, first highlight the column you wish to remove duplicates from. Under the **Data** heading, click **Remove Duplicates**.

# Functions

## Built-in: Deduplicating Rows

the scary way:



Excel will ask if you want to expand the selection. If you say no, only the duplicates from one column will be cleaned. The rest of the data will remained unchanged, and the cleaned column will no longer line up with the same entries.

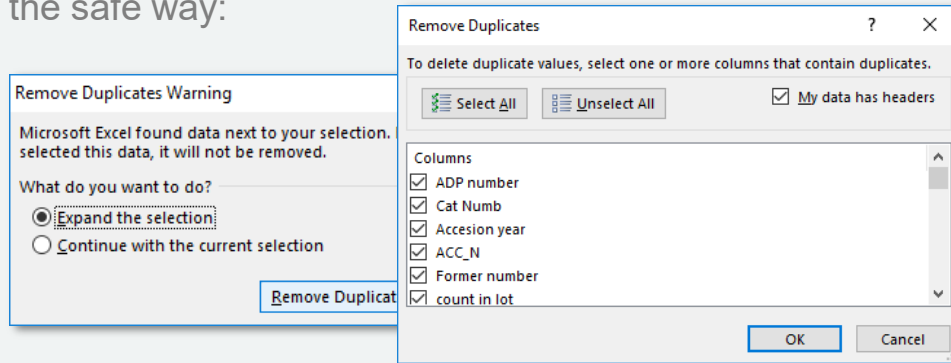
1	ADP num	Cat Nur	Accession	ACC_N	F
2	1 ABCD:1		1993	9999	
3	2 ABCD:2		1993	9999	
4	3 ABCD:3		1993	9999	
5	4 ABCD:4		1993	9999	
6	5 ABCD:5		1993	9999	
7	6 ABCD:6		1993	9999	
8	7 ABCD:7		1993	9999	
9	8 ABCD:8		1993	9999	
10	9 ABCD:9		1993	9999	
11	10 ABCD:10		1993	9999	
12	11 ABCD:11		1993	9999	
13	12 ABCD:12		1993	9999	
14	13 ABCD:13		1993	9999	
15	14 ABCD:14		1993	9999	
16	15 ABCD:15		1993	9999	
17	16 ABCD:16		1993	9999	
18	17 ABCD:17		1993	9999	
19	18 ABCD:18		1993	9999	
20	19 ABCD:19		1993	9999	

1	ADP num	Cat Nur	Accession	ACC_N	F
2	1 ABCD:1		1993	9999	
3	2 ABCD:2		1922	9999	
4	3 ABCD:3		1984	9999	
5	4 ABCD:4		1981	9999	
6	5 ABCD:5		1942	9999	
7	6 ABCD:6		1982	9999	
8	7 ABCD:7		1979	9999	
9	8 ABCD:8		1755	9999	
10	9 ABCD:9		1753	9999	
11	10 ABCD:10		1995	9999	
12	11 ABCD:11		1978	9999	
13	12 ABCD:12		1921	9999	
14	13 ABCD:13		1941	9999	
15	14 ABCD:14		1985	9999	
16	15 ABCD:15		9999	9999	
17	16 ABCD:16		1944	9999	
18	17 ABCD:17		1943	9999	
19	18 ABCD:18		1956	9999	
20	19 ABCD:19		1983	9999	

# Functions

## Built-in: Deduplicating Rows

the safe way:

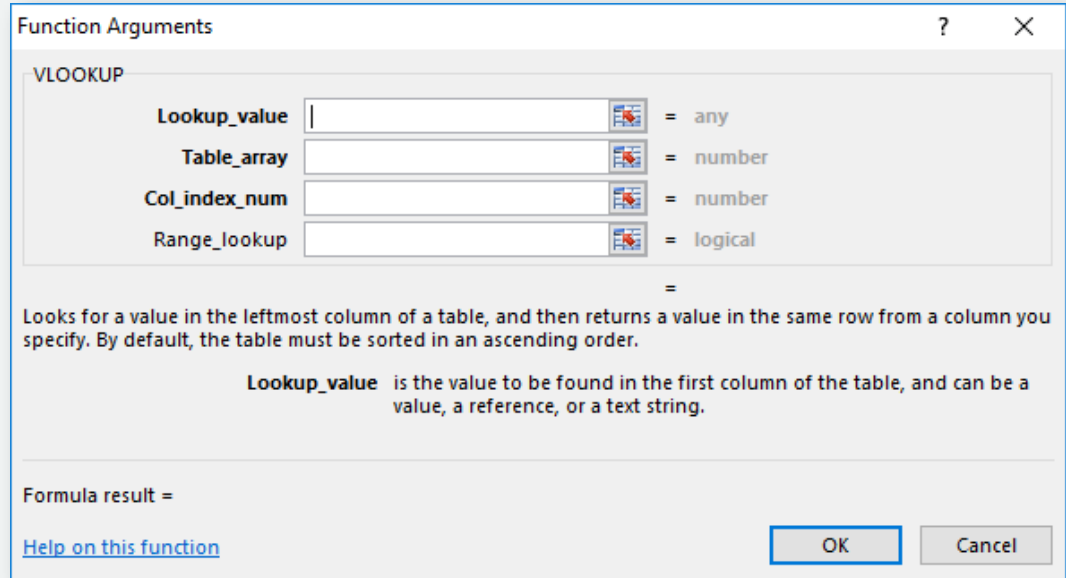


If you expand the selection, all the data will be cleaned at once. The rows will remain intact.

# Functions

## Build your own: Lookup Lists with VLOOKUP

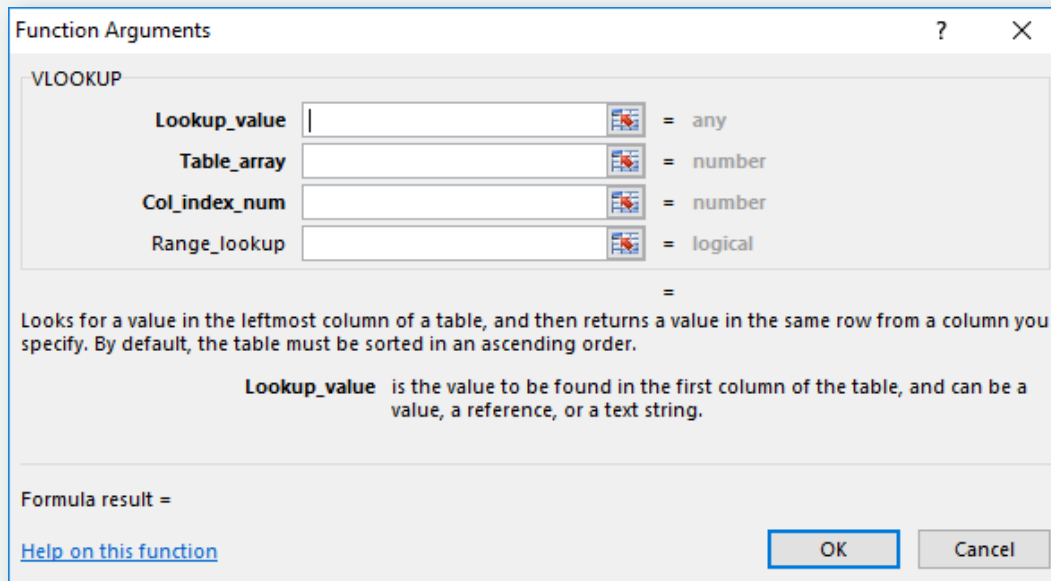
- The VLOOKUP function can be used to find a value in a range or table by row.
  - For example, look up scientific name by IRN.



# Functions

## Build your own: Lookup Lists with VLOOKUP

- The VLOOKUP function requires four arguments:
  - The **Lookup\_value** argument is used to define the value you want to look up. It can be a value, reference or text string.
  - The **Table\_array** argument is used to define the table or range (via reference or table name) where you want to look for the lookup value.



**NOTE:** If the `Lookup_value` is a value or text string, it must be located in the leftmost column of the table or range.

# Functions

## Build your own: Lookup Lists with VLOOKUP

- The VLOOKUP function requires four arguments:
  - The **Col\_index\_num** argument is used to define the column number of the column that contains the value to be returned.
  - The **Range\_lookup** argument is used to return an approximate or exact match. Use **0/FALSE** to prevent approximate matches.

? X

**Function Arguments**

VLOOKUP

<b>Lookup_value</b>	<input type="text"/>	=	any
<b>Table_array</b>	<input type="text"/>	=	number
<b>Col_index_num</b>	<input type="text"/>	=	number
<b>Range_lookup</b>	<input type="text"/>	=	logical

=

Looks for a value in the leftmost column of a table, and then returns a value in the same row from a column you specify. By default, the table must be sorted in an ascending order.

**Lookup\_value** is the value to be found in the first column of the table, and can be a value, a reference, or a text string.

---

Formula result =

[Help on this function](#)

OK
Cancel

**NOTE:** The column number for Col\_index\_num is specific to the selected table or range. Remember Excel uses one-based indexing.

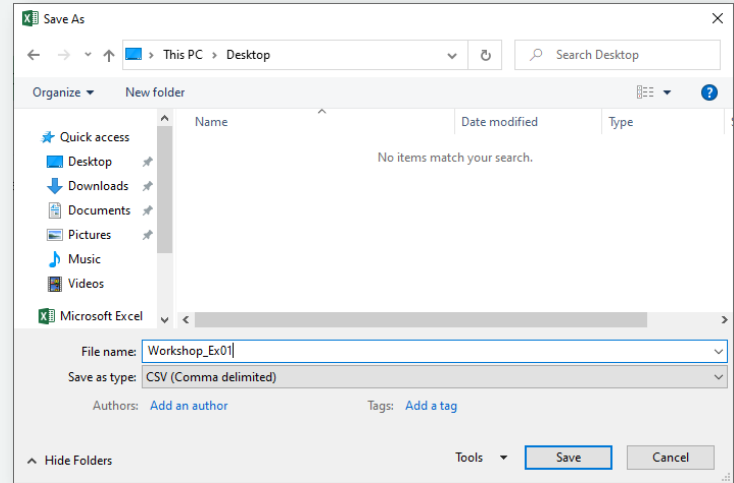
# Getting Data Out of Excel

## The Safe Way

...without mangling date-type data.

Differences between:

- A true CSV
  - where text-delimiter is specified and applied to all fields
- An Excel CSV
  - where text-delimiter is specified but only applied to text fields





# Recap

- Data in spreadsheets:
  - Good for one-time-use
  - Not good for maintaining
  - ...Is it structured? No.

- Alternatives to Excel:
  - Google Sheets
  - LibreOffice
  - Not using spreadsheets...?



# Thank you!

**Sharon Grant**

sgrant@fieldmuseum.org

<https://www.fieldmuseum.org/>

**Janeen Jones**

jjones@fieldmuseum.org

<https://www.fieldmuseum.org/>

**Kate Webbink**

kwebbink@fieldmuseum.org

<https://www.fieldmuseum.org/>

**Abigail McArthur-Self**

amcarthur-self@fieldmuseum.org

<https://www.fieldmuseum.org/>

**Alexis Ramirez**

aramirez@fieldmuseum.org

<https://www.fieldmuseum.org/>



This project was made possible in part by the  
**Institute of Museum and Library Services**  
Grant ME-249136-OMS-21 | [IMLS.gov](https://www.ims.gov)

