# LibreOffice Calc 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…

Reference: Naidoo, H (Hush Naidoo Jade Photography), October 2020, Unsplash License, https://unsplash.com/photos/SutAduhzdRQ

# Agenda

| | |
|---|---|
| What are Spreadsheet Tools? | 01 |
| Getting Data into Calc | 02 |
| Summarizing/Slaughtering Data in Calc | 03 |
| Manipulating/Mangling Data in Calc | 04 |
| Getting Data Out of Calc | 05 |
| Recap & Alternatives to Calc | 06 |

# What are Spreadsheet Tools?

# Examples of Spreadsheet Tools



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

**LibreOffice Calc** is the Document Foundation's free open-source spreadsheet program. Available here: 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 LibreOffice Calc Do?



| | Excel | LibreOffice Calc | Google Slides |
|---|---|---|---|
| Formulas | ☑ | ☑ | ☑ |
| Charts & Graphs | ☑ | ☑ | ☑ |
| Auto-formats fields | ⚠ | | ⚠ |
| Allow creation of lookup lists | ☑ | ☑ | ☑ |
| Allow character sets besides Latin1 (e.g., UTF-8) | Limited | ☑ | ☑ |
| Track changes after closing | Limited | Limited | ☑ |
| 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.

github.com/datacarpentry/spreadsheet-ecology-lesson/blob/gh-pages/_episodes/02-common-mistakes.md

FIELD MUSEUM

# Tables vs. Spreadsheets

**Tables:**

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

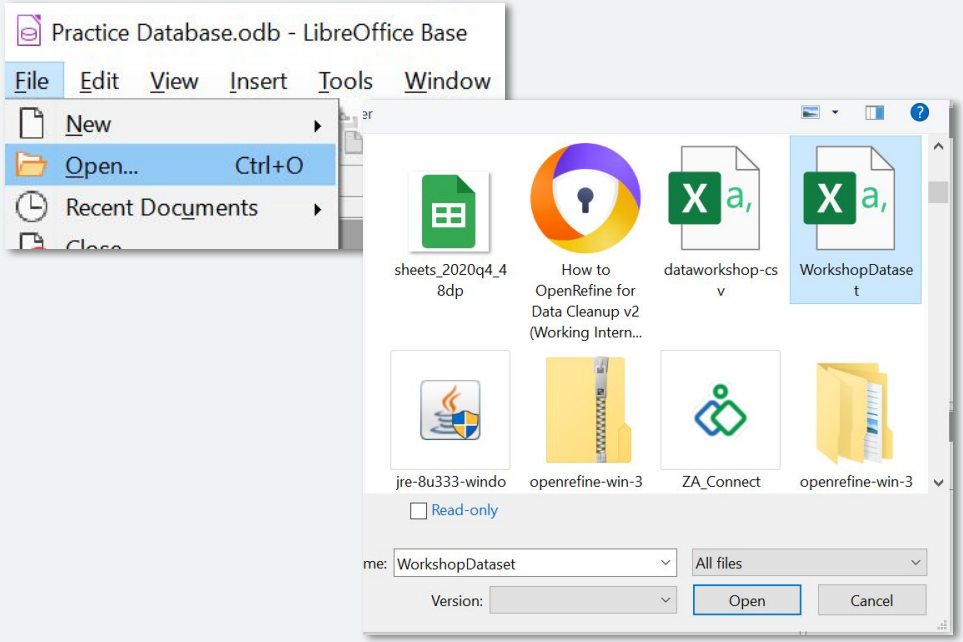
**Spreadsheets:**

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

# Getting Data into Calc

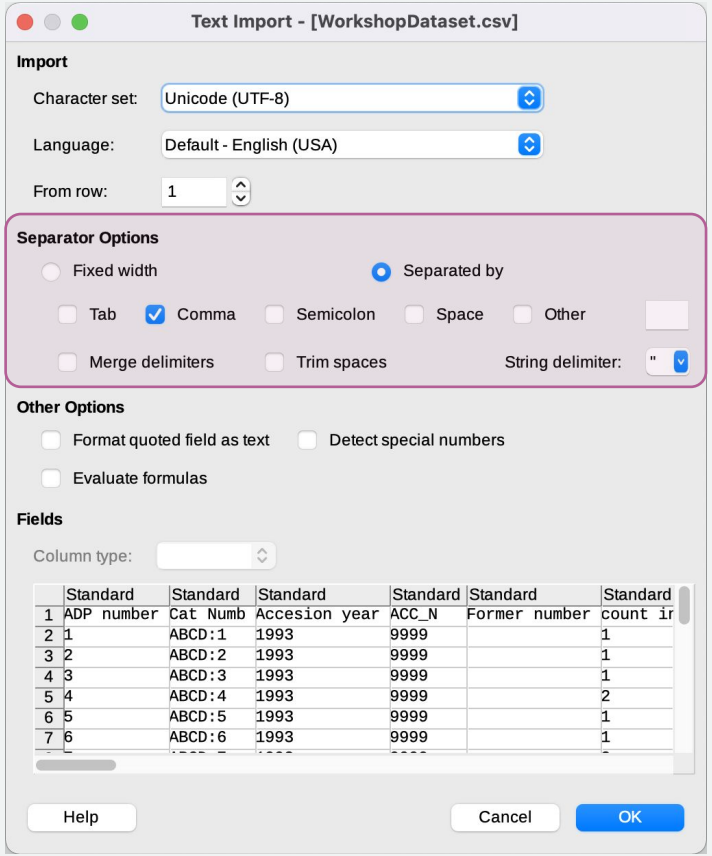
# Getting Data into Calc

- Open LibreOffice.
- From the File menu, click Open… to open the data file. If the data is saved in a standard format such as a .csv, LibreOffice will open it in Calc by default.



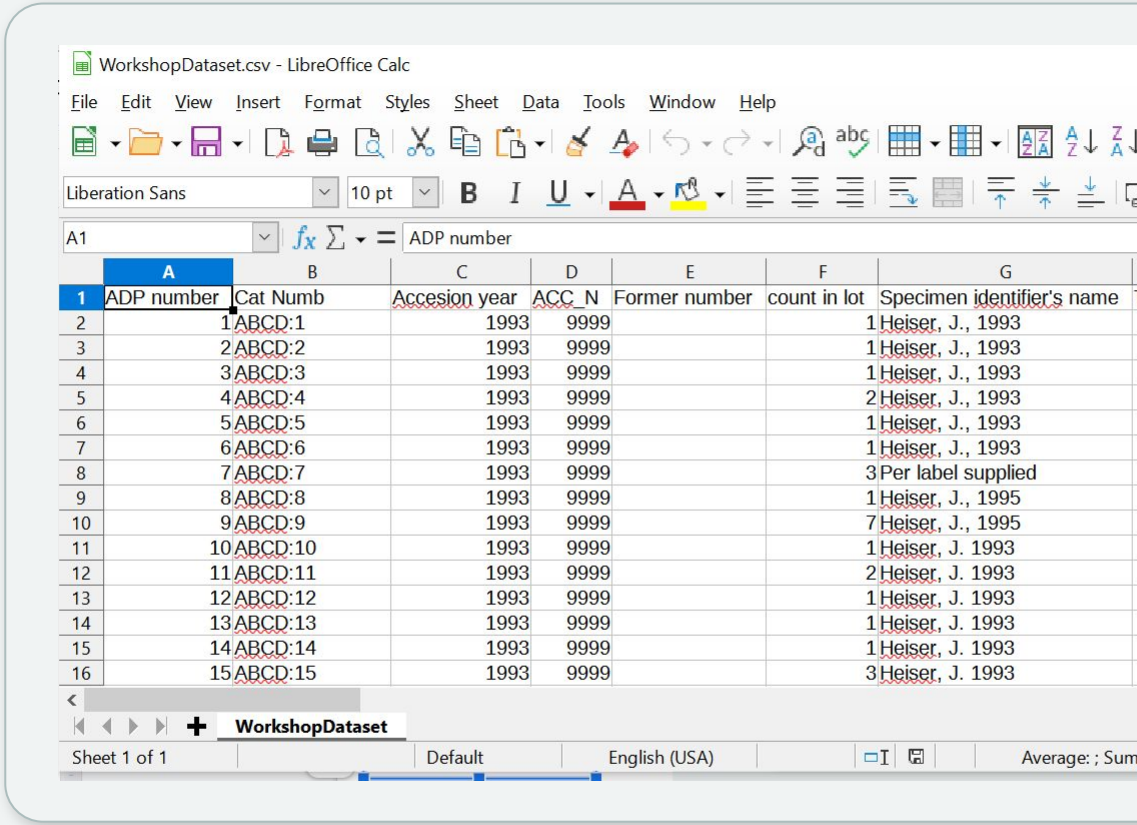
# Getting Data into Calc

- A Text Import wizard will open.
- Certain options will be selected by default. Review them to make sure the settings are fit for the data file.
  - Latin-1 and UTF-8 are the most common character sets.
- If something seems off in the preview window at the bottom of the wizard, try adjusting the Separator Options.



# Getting Data into Calc

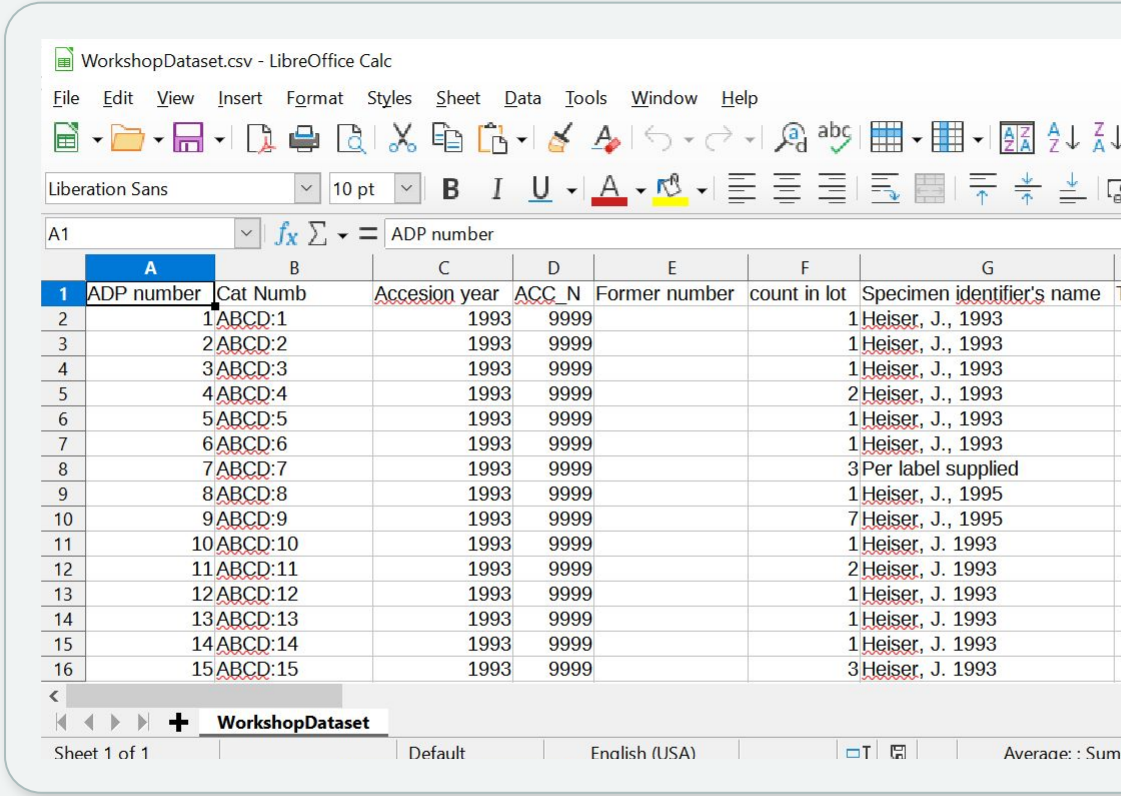
- When everything looks good, click OK. The data will open in a Calc file.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ADP number | Cat Numb | Accesion year | ACC_N | Former number | count in lot | Specimen identifier's name |
| 2 | 1 | ABCD:1 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 3 | 2 | ABCD:2 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 4 | 3 | ABCD:3 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 5 | 4 | ABCD:4 | 1993 | 9999 | | 2 | Heiser, J., 1993 |
| 6 | 5 | ABCD:5 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 7 | 6 | ABCD:6 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 8 | 7 | ABCD:7 | 1993 | 9999 | | 3 | Per label supplied |
| 9 | 8 | ABCD:8 | 1993 | 9999 | | 1 | Heiser, J., 1995 |
| 10 | 9 | ABCD:9 | 1993 | 9999 | | 7 | Heiser, J., 1995 |
| 11 | 10 | ABCD:10 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 12 | 11 | ABCD:11 | 1993 | 9999 | | 2 | Heiser, J. 1993 |
| 13 | 12 | ABCD:12 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 14 | 13 | ABCD:13 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 15 | 14 | ABCD:14 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 16 | 15 | ABCD:15 | 1993 | 9999 | | 3 | Heiser, J. 1993 |

# Getting Data into Calc

- When everything looks correct, click OK.
- Calc will open as a live date file.
  - Using Save (from File menu) after making any edits and selecting the Use Text CSV Format option will overwrite the original data file.
  - Instead, select Use ODF Format or use Save As….




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ADP number | Cat Numb | Accesion year | ACC_N | Former number | count in lot | Specimen identifier's name |
| 2 | 1 | ABCD:1 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 3 | 2 | ABCD:2 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 4 | 3 | ABCD:3 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 5 | 4 | ABCD:4 | 1993 | 9999 | | 2 | Heiser, J., 1993 |
| 6 | 5 | ABCD:5 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 7 | 6 | ABCD:6 | 1993 | 9999 | | 1 | Heiser, J., 1993 |
| 8 | 7 | ABCD:7 | 1993 | 9999 | | 3 | Per label supplied |
| 9 | 8 | ABCD:8 | 1993 | 9999 | | 1 | Heiser, J., 1995 |
| 10 | 9 | ABCD:9 | 1993 | 9999 | | 7 | Heiser, J., 1995 |
| 11 | 10 | ABCD:10 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 12 | 11 | ABCD:11 | 1993 | 9999 | | 2 | Heiser, J. 1993 |
| 13 | 12 | ABCD:12 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 14 | 13 | ABCD:13 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 15 | 14 | ABCD:14 | 1993 | 9999 | | 1 | Heiser, J. 1993 |
| 16 | 15 | ABCD:15 | 1993 | 9999 | | 3 | Heiser, J. 1993 |

# Getting Data into Calc



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

- Most commonly they are either **Latin-1** or **UTF-8.**

# Summarizing/ Slaughtering Data in Calc

# Summarizing/ Slaughtering Data in Calc

## Sorting and Filtering

There are four sort/filter buttons in the menu bar (left to right):

1. Sort
2. Sort Ascending
3. Sort Descending
4. AutoFilter

# Sorting



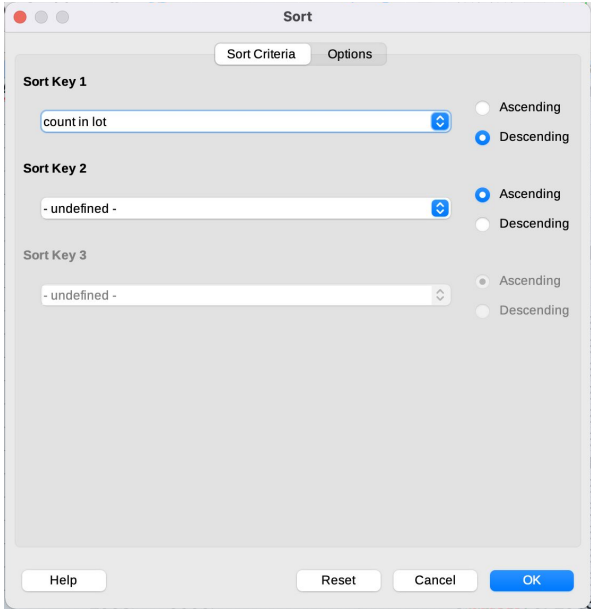
Sort, Sort Ascending, Sort Descending

## Sort Ascending and Descending

- The two buttons are contextual. They will sort based on the highlighted cell.
- **Note:** Calc automatically selects all data.

## Sort

- The Sort button will open the Sort window.
- Set Sort Key 1 to “count in lot” or the column you’d like to sort.
- Select Descending or the desired sorting direction.

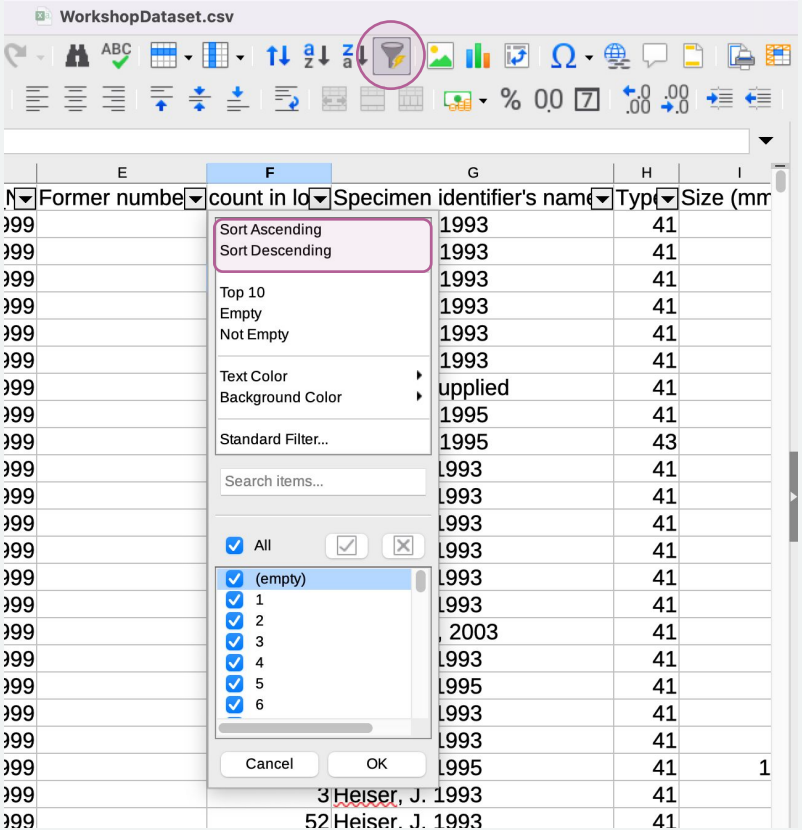


# Sorting

…or use AutoFilter.

To sort data using AutoFilter:

1. Click the AutoFilter button
2. Click the dropdown menu button for the desire column
3. Select Sort Ascending or Sort Descending



# What is the difference between sorting and filtering?

### Sorting

Sorting rearranges the presentation of the data within a table — including **all** values.

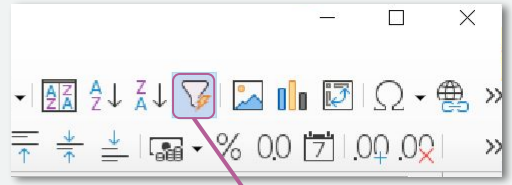
### Filtering

Filtering rearrange the presentation of the data within a table — including only the values that meet the specified criteria.

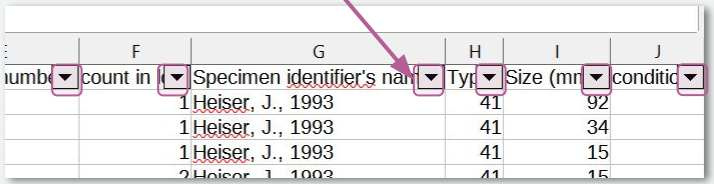
In Calc, the AutoFilter button places drop-down menus at the top of each column with options for both sorting and filtering.

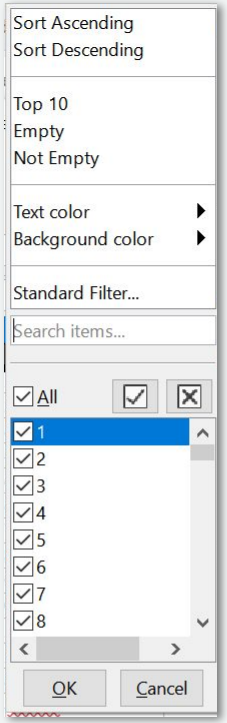
# Summarizing/ Slaughtering Data in Calc

Filtering



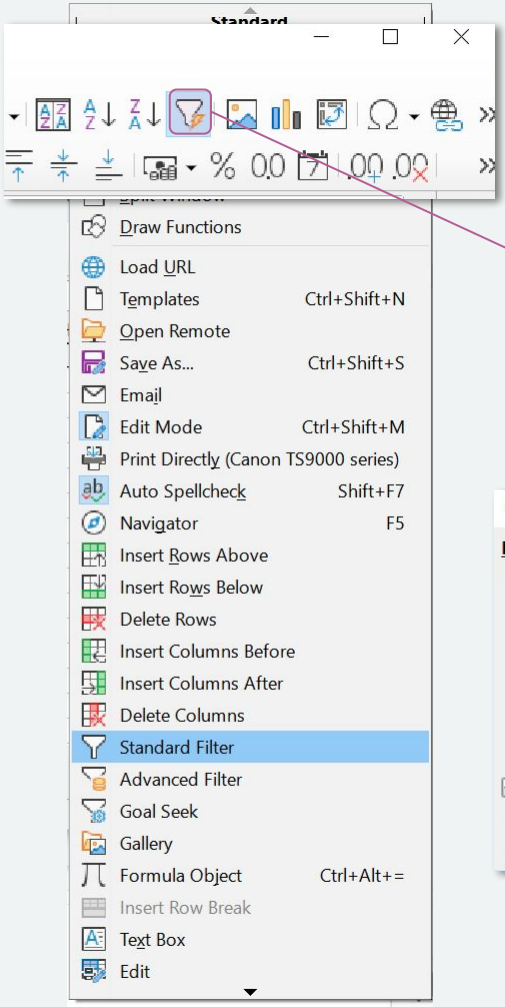
If you left click the icon, it will “Autofilter,” placing grey dropdown menus on all the columns.



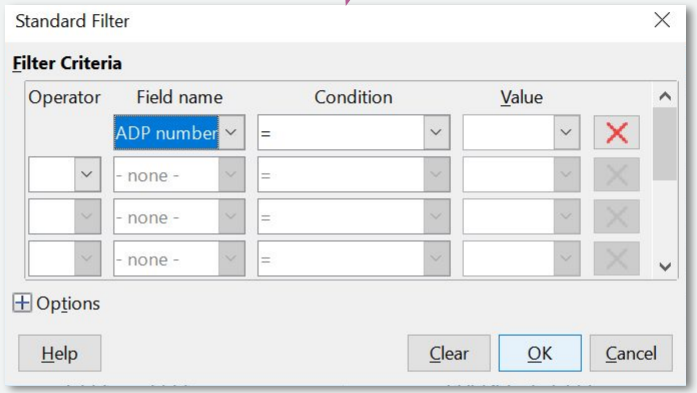



# Summarizing/ Slaughtering Data in Calc

Filtering




If you right-click on the icon, it will open a menu with more options. You can select **Standard Filter** to create your own filter.



# Manipulating/ Mangling Data in Calc

# Manipulating/Mangling Data in Calc

## Functions

Build your own:

- `CONCATENATE`
- `VLOOKUP`

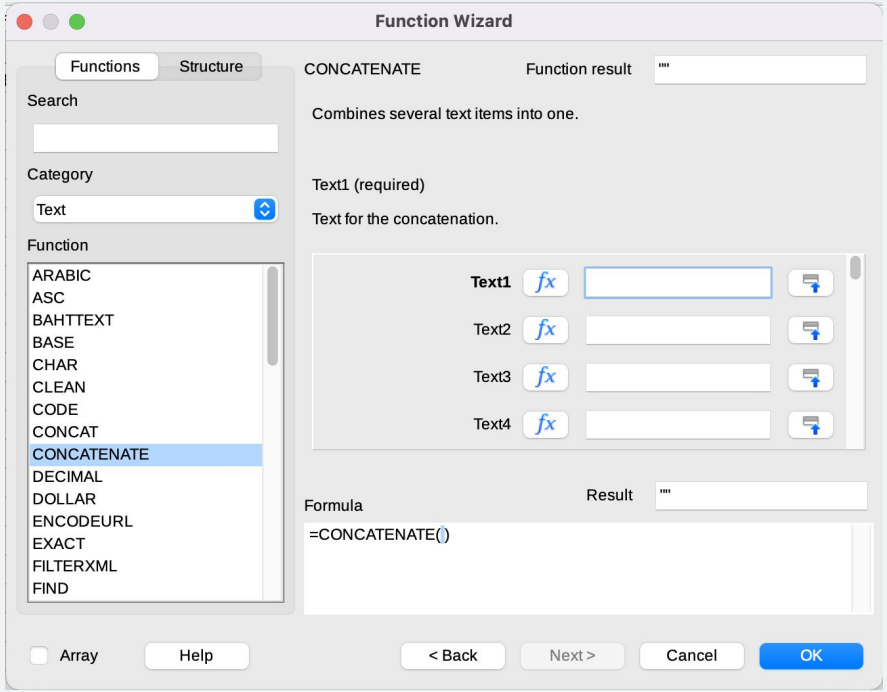
Built-in:

- Split Text-to-Columns
- Deduplication

# Build Your Own Functions

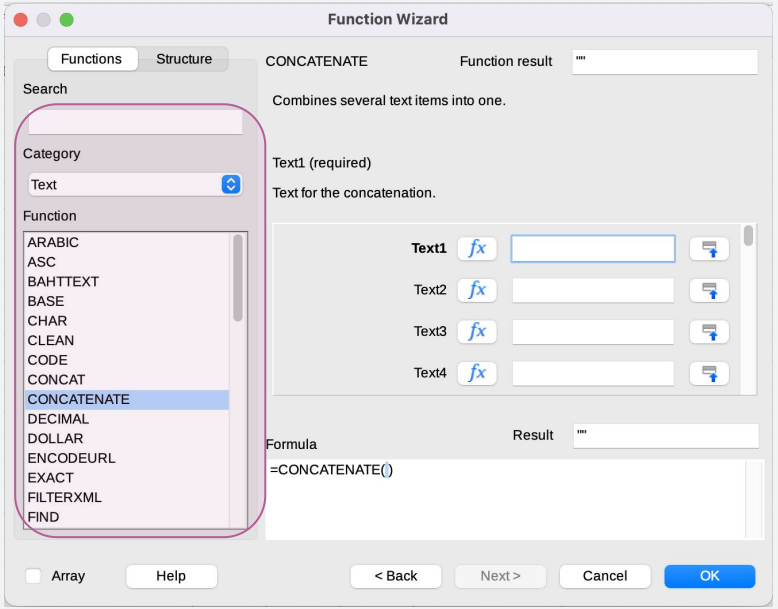
## Function Wizard: CONCATENATE

- Syntax:
  `=FUNCTION(arguments)`
- Built-in help offers a syntax guide:



# Build Your Own Functions

## Function Wizard: CONCATENATE




To build the function using the Function Wizard:

1. Click the Function Wizard button with the target cell highlighted.

2. Select the function from the Function box. Use the Search bar to search for the `CONCATENATE` function.

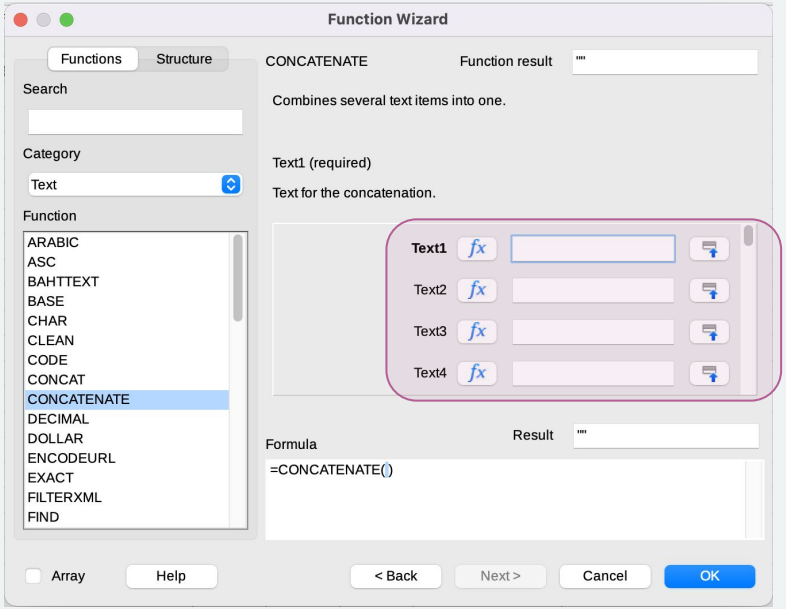
3. Click the Next button.

# Build Your Own Functions

## Function Wizard: CONCATENATE

4. Input cell names and text into the Text for the concatenation boxes. For this example: `AD3, “-”, AE3, “-”, AF3` for fields Text1 through Text5.

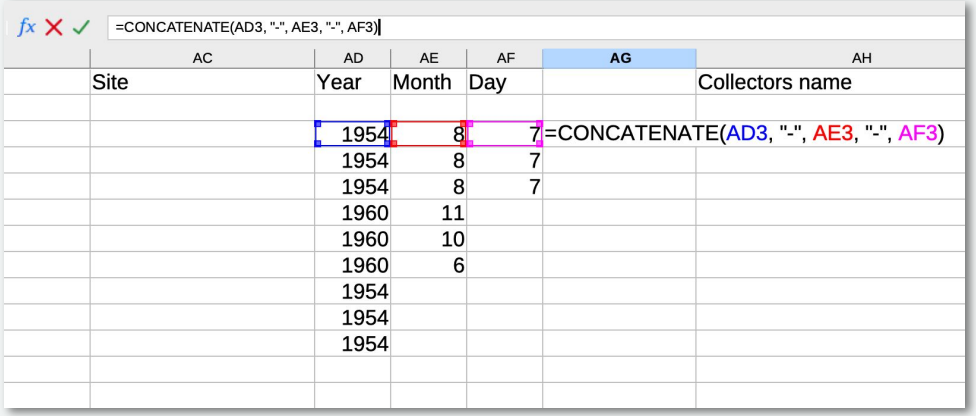
5. The Result box will allow you to preview the displayed value of the cell. Click OK.

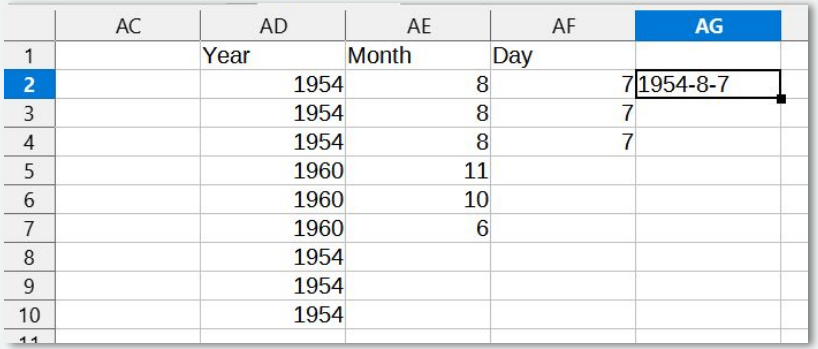


# Build Your Own Functions

Function Wizard:
CONCATENATE

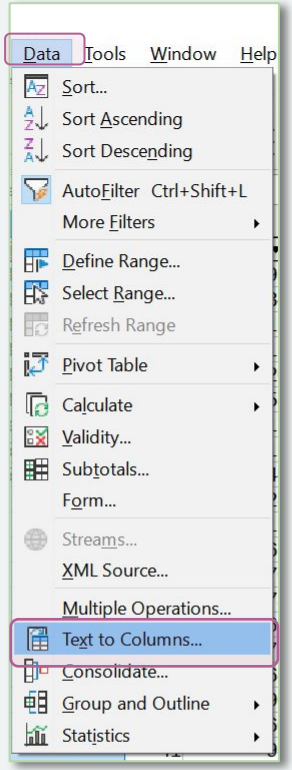
- `CONCATENATE(arguments)`



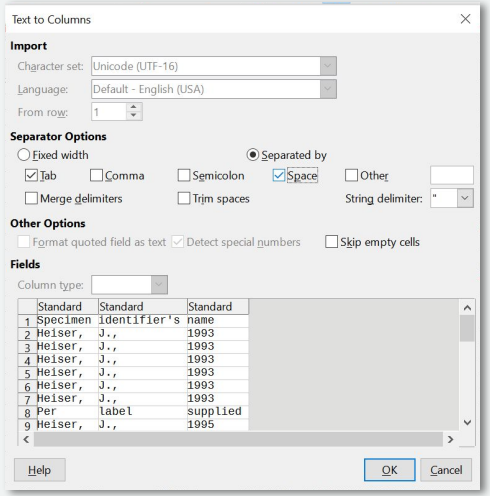



# Built-in Functions

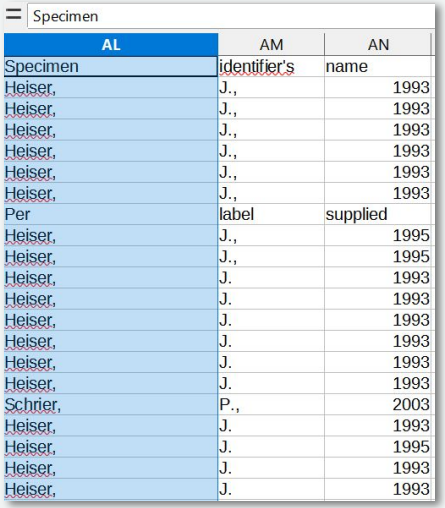
## Split Text → Columns




Select Text to Columns… from the Data menu.




Select your settings in the Text to Column Wizard. Make sure the right delimiter is selected by previewing the data in the bottom window.




| AL | AM | AN |
|---|---|---|
| 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 |
| Heiser, | J. | 1993 |
| Schrier, | P., | 2003 |
| Heiser, | J. | 1993 |
| Heiser, | J. | 1995 |
| Heiser, | J. | 1993 |
| Heiser, | J. | 1993 |

# Functions

Calc does not have a built-in deduplication tool. There are two options:

Install an Extension:

- Remove Duplicates is a free extension for Calc.

Use Formatting:

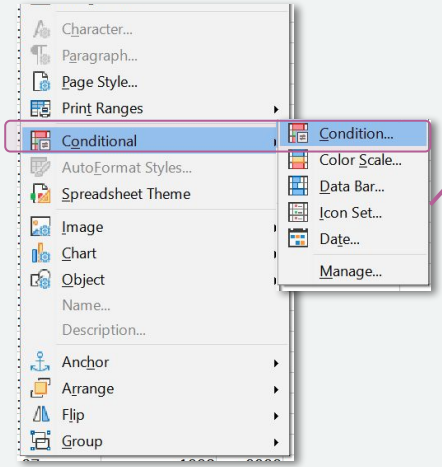
- Calc includes functions for conditional formatting and filtering, which can be used to find and remove duplicates.

# Functions

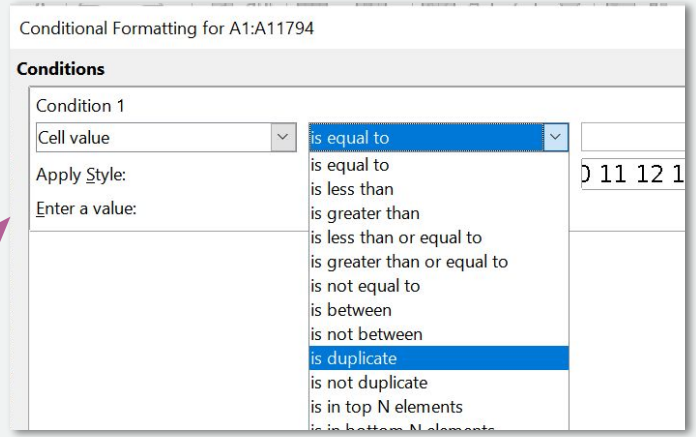
## Deduplicating Rows

| | A | B | C | D | |
|---|---|---|---|---|---|
| 1 | ADP number | Cat Numb | Accesion year | ACC_N | Fo |
| 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 | |
| 21 | 20 | ABCD:20 | 1993 | 9999 | |
| 22 | 21 | ABCD:21 | 1993 | 9999 | |
| 23 | 22 | ABCD:22 | 1993 | 9999 | |
| 24 | 23 | ABCD:23 | 1993 | 9999 | |
| 25 | 24 | ABCD:24 | 1993 | 9999 | |
| 26 | 25 | ABCD:25 | 1993 | 9999 | |
| 27 | 26 | ABCD:26 | 1993 | 9999 | |
| 28 | 27 | ABCD:27 | 1993 | 9999 | |

To deduplicate rows, highlight the column to sort for duplicates. Under Format, select Conditional → Condition…



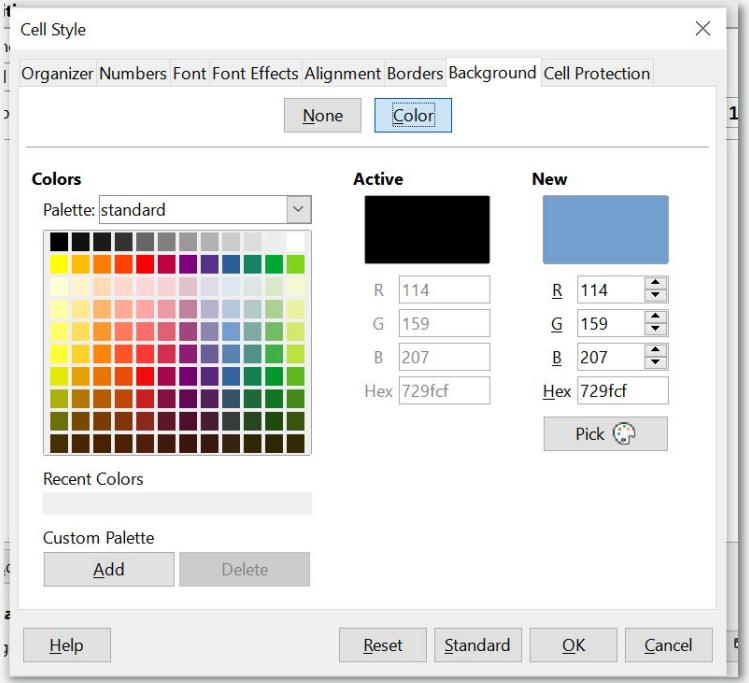

Set the **condition** to “is duplicate” and set the style for the conditionally formatted boxes.



# Functions

## Deduplicating Rows

If you haven't used conditional formatting this way in the program, you may have to create a **new style** that changes the background color rather than font style.

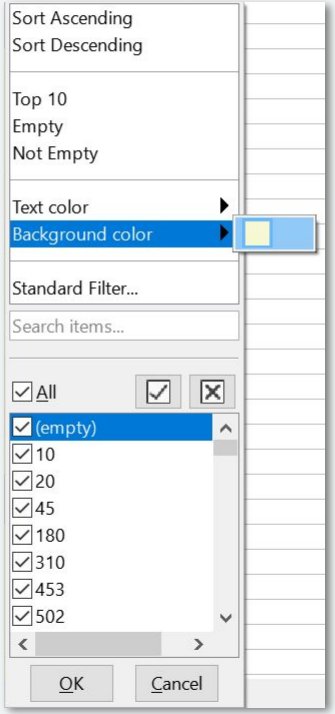


# Functions

## Deduplicating Rows

Once you've formatted any duplicates, use the autofilter setting. Select **Background color** from the dropdown menu to bring duplicates to the top.

From there, check and remove any rows that are duplicates.



# Functions

## Deduplicating Rows

Typically, it's best to avoid background colors because they can make a spreadsheet look messy.

When removing duplicates, however, all rows with a different background color will be deleted.

| | Z | AA | AB |
|---|---|---|---|
| 1 | fieldnumber | accession numbe | date collected |
| 4 | HQ926 | 9-Sep-85 | |
| 5 | HQ833 B-3 | 9-Sep-85 | |
| 6 | HQ832 | 9-Sep-85 | |
| 7 | HQ515 | 9-Sep-85 | |
| 8 | HQ177 | 9-Sep-85 | |
| 9 | | 6267 2-VI-84 | |
| 10 | | 6267 2-VI-84 | |
| 11 | | 6267 2-VI-84 | |
| 12 | | 6267 2-VI-84 | |
| 13 | | 6267 2-VI-84 | |
| 14 | F027 | 6267 2-VI-84 | |
| 15 | | 6267 2-VI-84 | |

# Build Your Own Functions

## Lookup Lists with VLOOKUP

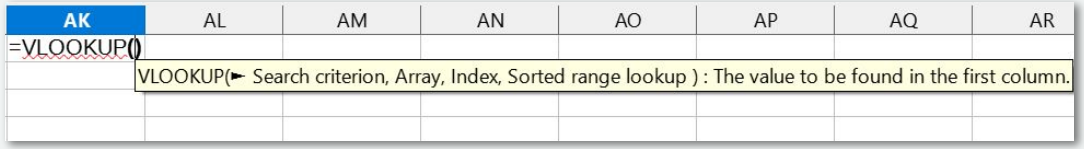
=VLOOKUP(

[cell with a value to replace]

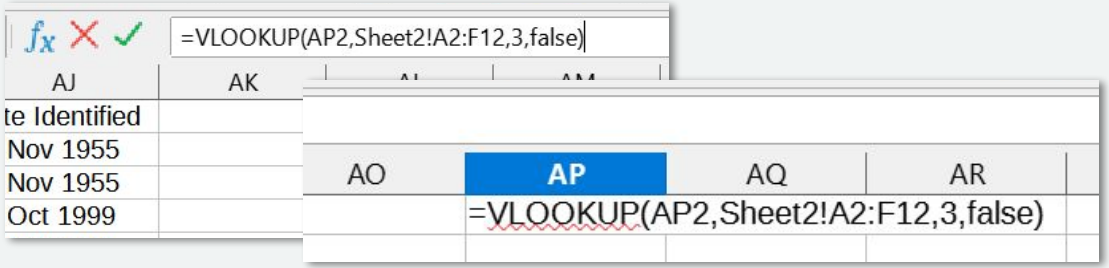
[range where column 1 = matches & column X = replacements]

[column number X]

FALSE to prevent approximate matches )




You can type =VLOOKUP directly into the spreadsheet to use the function. Calc will give you minimal help.



# Build Your Own Functions

## Lookup Lists with VLOOKUP

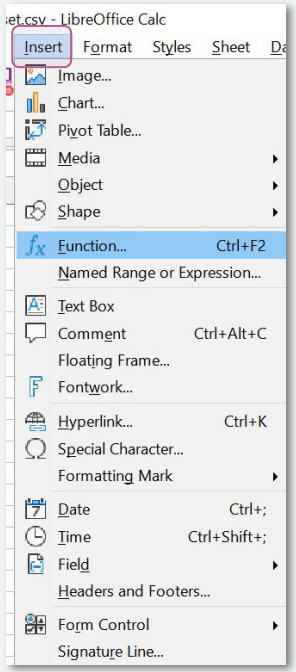
`=VLOOKUP(` [cell with a value to replace]

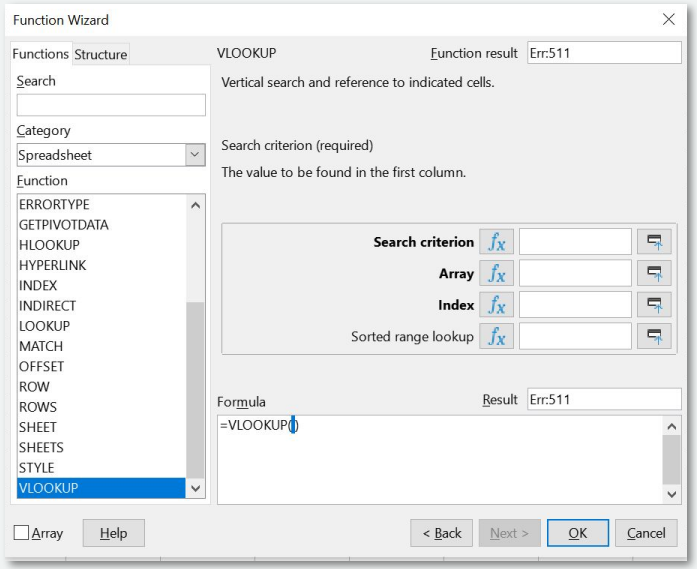
[range where column 1 = matches & column X = replacements]

[column number X]

FALSE to prevent approximate matches `)`

If you want more help writing your function, you can **Insert → Function** to open the Function Wizard.



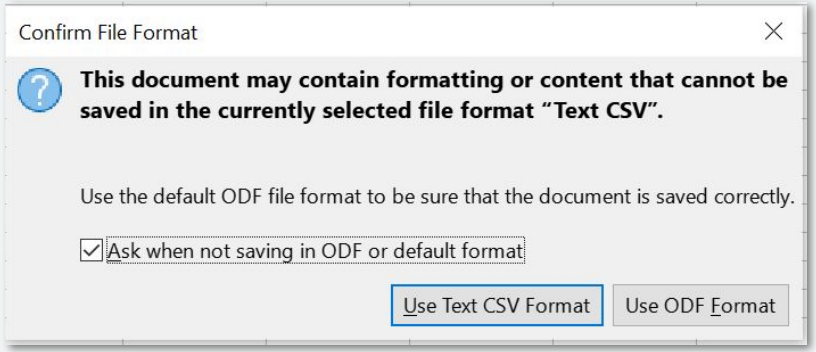



# Getting Data Out of Calc

# Getting Data Out of Calc

You can use the “Save As” function to save the file in a variety of formats.

- Text CSV
  - Where text delimiter is specified and applied to all fields
  - Readable by most programs
- An ODF Spreadsheet
  - The default spreadsheet format for LibreOffice, comparable to Microsoft Excel XLS and XLSX files.




If you attempt to **Save** directly to the original file, you will get a warning. To ensure you don’t mangle any data, **Save As** to make a copy, rather than overwriting the original.

# Recap

# Recap



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

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

# Questions?

**Ask now; we might have answers!**

# Presenter name

Title or credentials

- Bio
- Contact Information

Insert Presenter Image

# Thank you!


**Sharon Grant**
sgrant@fieldmuseum.org
https://www.fieldmuseum.org/

**Janeen Jones**
jjones@fieldmuseum.org
https://www.fieldmuseum.org/

**Kate Webbink**
kwebbink@fieldmusuem.org
https://www.fieldmuseum.org/

**Abigail McArthur-Self**
amcarthur-self@fieldmusuem.org
https://www.fieldmuseum.org/

**Alexis Ramirez**
aramirez@fieldmusuem.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**