

# Tidy data for librarians

<https://librarycarpentry.org/lc-spreadsheets/>

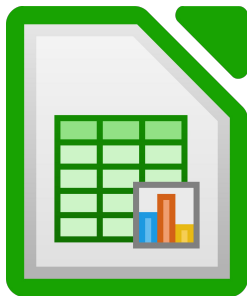
# Today's tools and references

- Zoom interaction
  - Chat, React, and Speak!
- These slides
- Spreadsheet software
  - Microsoft Excel for demos
- The "training\_attendance.xlsx" data file
- The "Tidy data for librarians" lesson

# **Using spreadsheet programs for data organization**

What data do you put into spreadsheets?

Which software tools do you use to work with your spreadsheets?



What kind of operations do you do in  
spreadsheets?

# Limitations

- Long data
  - Lots and lots of records/observations
- Automation
- Transparency and repeatability



# Our topics

- Formatting data tables in spreadsheets
- Avoiding common formatting mistakes
- Being wary of dates
- Quality assurance and quality control features
- Exporting data from spreadsheet software

<https://librarycarpentry.org/lc-spreadsheets/>

# **Formatting data tables in Spreadsheets**

# Tidy tables

Header		Variable	
Observation		Value	

# Exercise

You're asked to evaluate the training program from 2016 and 2017.

You want to analyze the dataset and need to aggregate the data into a single table.

# Rules for tidying tabular data

1. Leave the raw data raw.
2. Start with a header row and put all your variables in columns.
3. Put each observation or record in its own row.
4. Avoid including multiple pieces of information in one cell.
5. (Eventually) Export the cleaned data to a plain text format.

# **Formatting problems**

# Common problematic formatting

- Multiple tables in a single sheet
- Multiple tabs for a single dataset
- Zeroes and null values
- Using formatting to convey information or for visual effect
- More than one piece of information in a cell
- Field name problems
- Special characters in data
- Inclusion of metadata in a data table

**Dates as data**



## PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS ***THE*** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27


THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13

20130227 2013.02.27 27.02.13 27-02-13

27.2.13 2013.II.27.  $27\frac{1}{2}$ -13 2013.158904109

MMXIII-II-XXVII MMXIII  $\frac{\text{LVII}}{\text{CCCLXV}}$  1330300800

$((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$  2013  Hissss

10/11011/1101 02/27/20/13  $\begin{array}{ccccc} 2 & 3 & 1 & 4 & \\ 0 & 1 & 2 & 3 & 7 \\ 5 & & 6 & 7 & 8 \end{array}$

# **Basic quality assurance and control**

# **Exporting data from spreadsheets**