10.5281/zenodo.5053596

Spreadsheets

Esther Plomp
@toothFAIRy@scholar.social
e.plomp@tudelft.nl

Data Organization in Spreadsheets for Social Scientists

10.5281/zenodo.5053596

Lesson outline

- Good **data entry** practices
- **Formatting data tables** in spreadsheets
  - Avoid common mistakes
  - Handling dates
- **Basic quality control** and data manipulation in spreadsheets
- **Importing & exporting** data from spreadsheets

You will learn how to think about **data organisation** to effectively wrangle data later

10.5281/zenodo.5053596

Software needed

Microsoft Excel

Libre Office

Slides

Presentation is shared on a **data repository**!

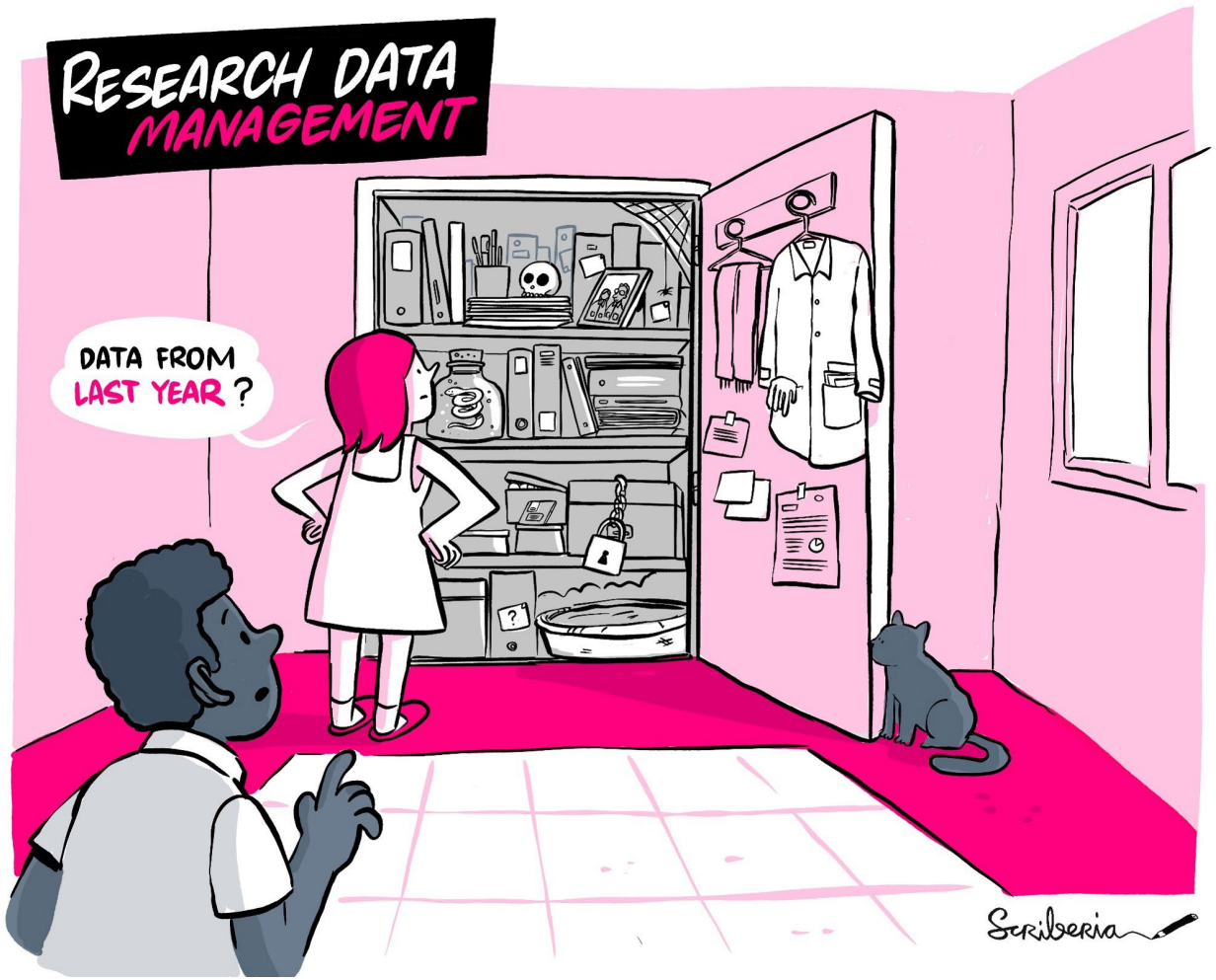
https://doi.org/10.5281/zenodo.5053596

Question

How many people have used spreadsheets in their research?

Spreadsheet management

10.5281/zenodo.5053596



10.5281/zenodo.5053596

Spreadsheets

- Data entry
- Tables for publications
- Generate summary statistics
- Figures

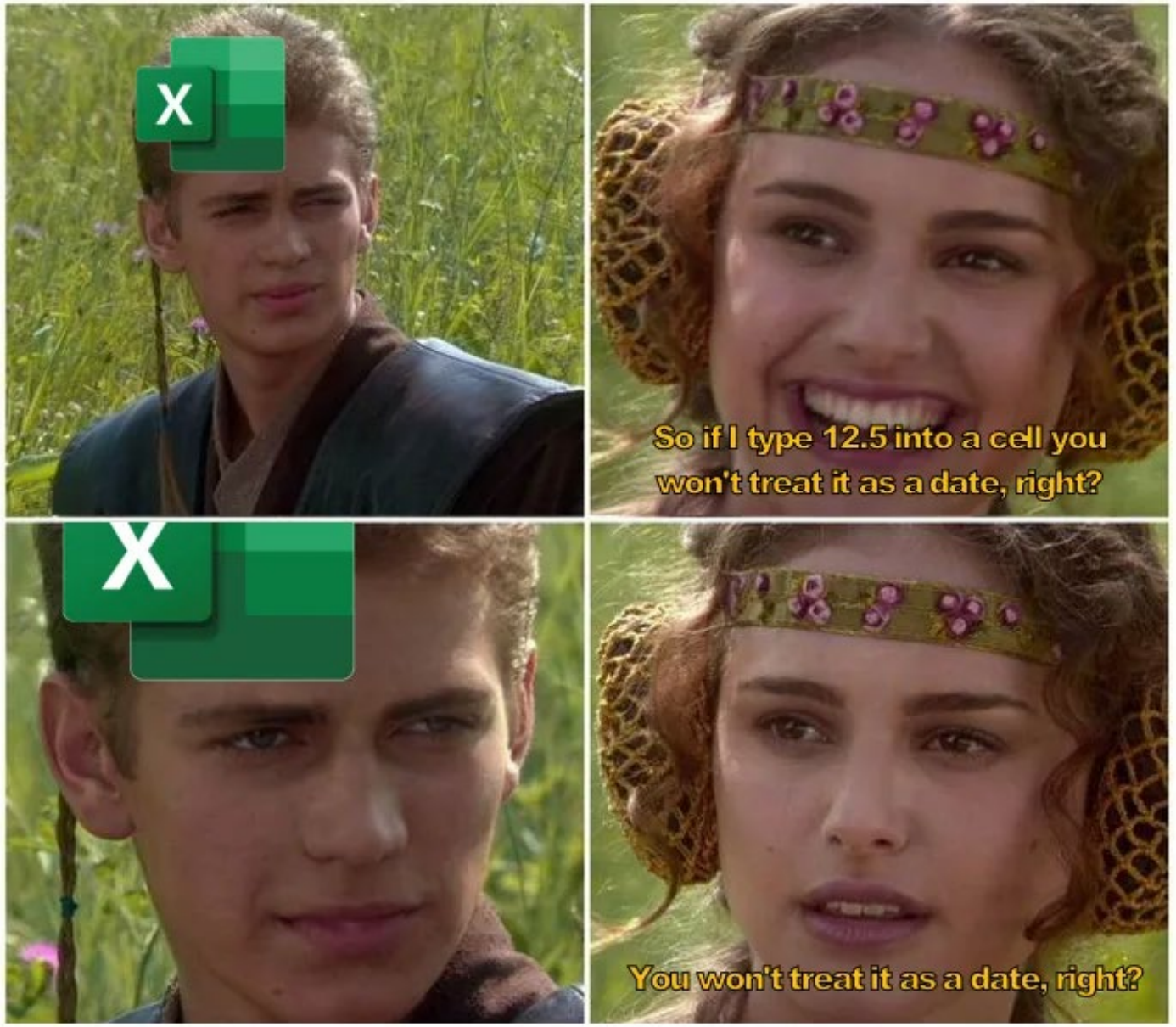
But…

- Manual:
  - Easy to make mistakes
  - Copy pasting the wrong cells
  - Messing up formula
  - Accidental deletion
  - Difficult to reproduce
- Not Machine readable:

  Software can't process certain information:
  - Notes in the margin
  - Spatial layout of data
  - Field formatting

10.5281/zenodo.5053596

Excel & Dates



https://9gag.com/gag/ayMQeKM

10.5281/zenodo.5053596

Excel & Dates

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

Presentation by Mandhri Abeysooriya on 'Spreadsheet Problems In Genomic Research'

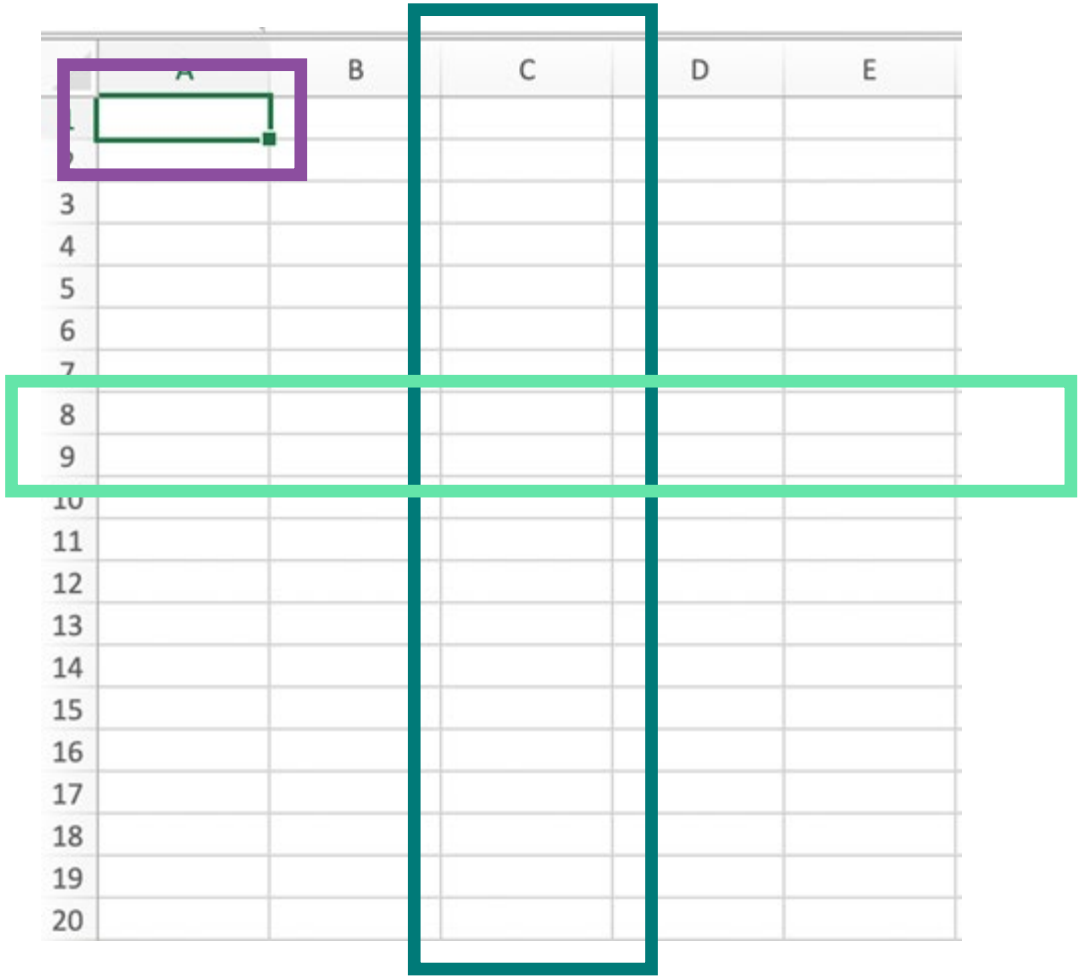
10.5281/zenodo.5053596

Cells, columns and rows

Each box is a **cell**

Each **column** is a specific data feature

Each **row** is an individual item/observation



10.5281/zenodo.5053596

Spreadsheets that are less error-prone, easier for computers to process, more accessible, and easier to share

- Be **consistent**
- Write dates like YYYYMMDD
  - Always check when there are dates in imported data and consider to split the YYYY, MM, and DD in separate cells
- Do not leave empty values (use NA)
- Put as **few information possible in a single cell** and one observation per row
- Create a **data dictionary** that describes the spreadsheet and any cleaning steps you took
- **Leave the raw data alone!**
- **Avoid formatting** (colours, font, bolding)
- Use **data validation** to avoid errors (OpenRefine)

Extended version

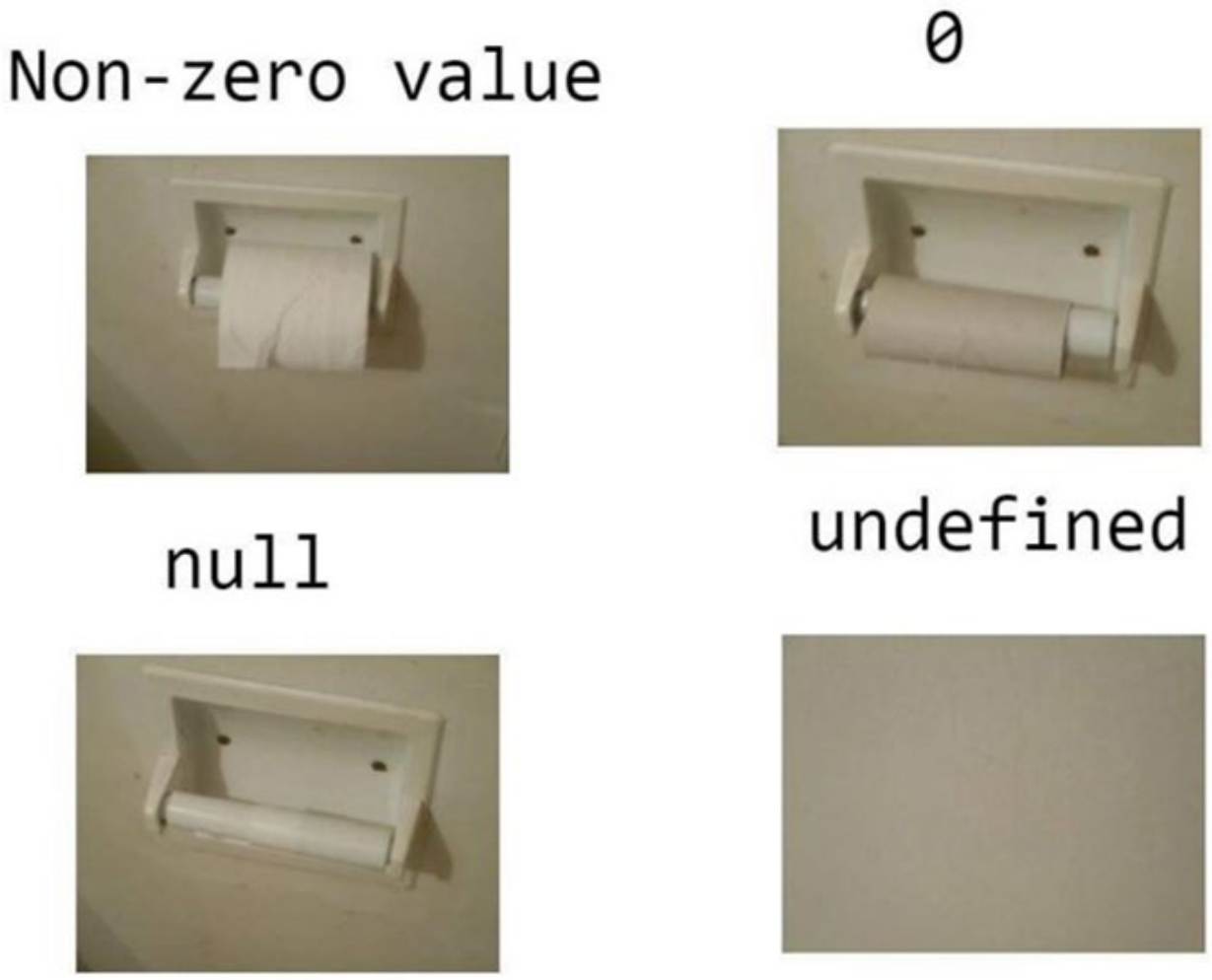
Broman and Woo 2018

Carpentries Spreadsheets for Ecologists or Social Scientists

10.5281/zenodo.5053596

Why use Not Available / NA?

- Difficult to know if a value is missing
- Blanks can be confusing when spaces/tabs are used as delimiter
- Use a consistent null value and indicate it in the metadata/README



Kelly Davis/Heather Froehlich - Introduction to Spreadsheets with Excel

10.5281/zenodo.5053596

Tidy Data (R)

Tidy Data Paper by Wickham (see here for the CRAN code heavy version)

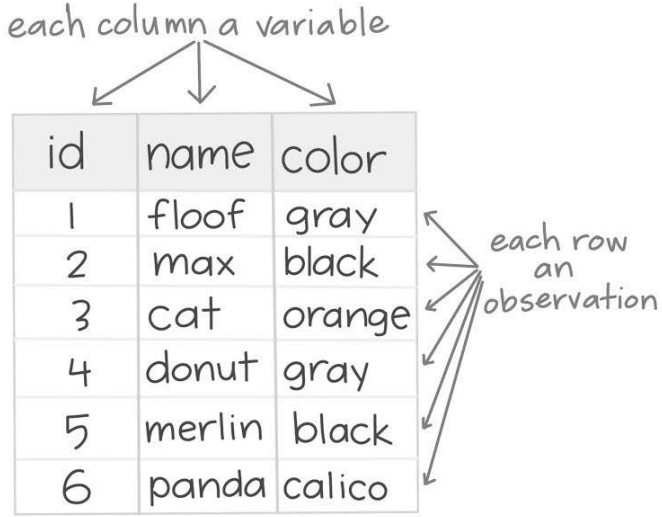
Welcome to the Tidyverse

The Tidyverse style guide

> "TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."
>
> –HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

https://www.openscapes.org/blog/2020/10/12/tidy-data/ Illustration by Allison Horst (CC-BY)

10.5281/zenodo.5053596

Workshop Dataset

Studying African Farmer-led Irrigation (SAFI) Dataset

- Interviews of farmers in two countries in eastern sub-Saharan Africa (Mozambique and Tanzania, conducted between November 2016 and June 2017)
- Topics:
  - household features (construction materials used, number of household members)
  - agricultural practices (water usage)
  - assets (number and types of livestock)
- This is a simplified version of a real dataset!

Exercise
7 min

1. Download the messy data
2. Open up the data in a spreadsheet program
3. Notice that there are **two tabs.** Two researchers conducted the interviews, one in Mozambique and the other in Tanzania. They both structured their data tables in a different way. Now, you're the person in charge of this project and **you want to be able to start analysing the data**!
4. **With the person next to you, identify what is wrong with this spreadsheet.** Discuss the steps you would need to take to clean up the two tabs, and start cleaning the spreadsheet

! Do not forget
to create a new file (or tab) for the cleaned data &
never modify your original (raw) data

Metadata

10.5281/zenodo.5053596

Data about data

Information to make the data understandable

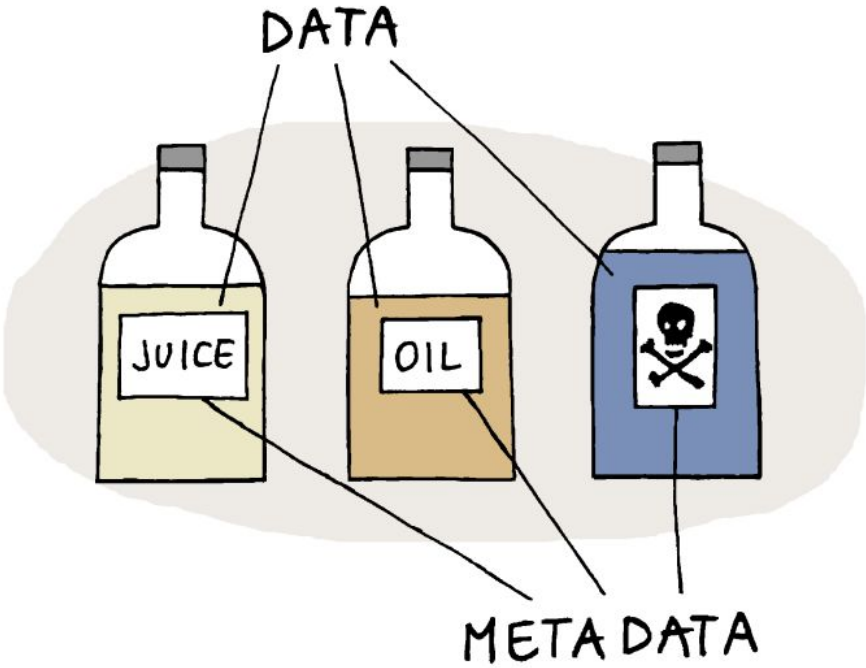
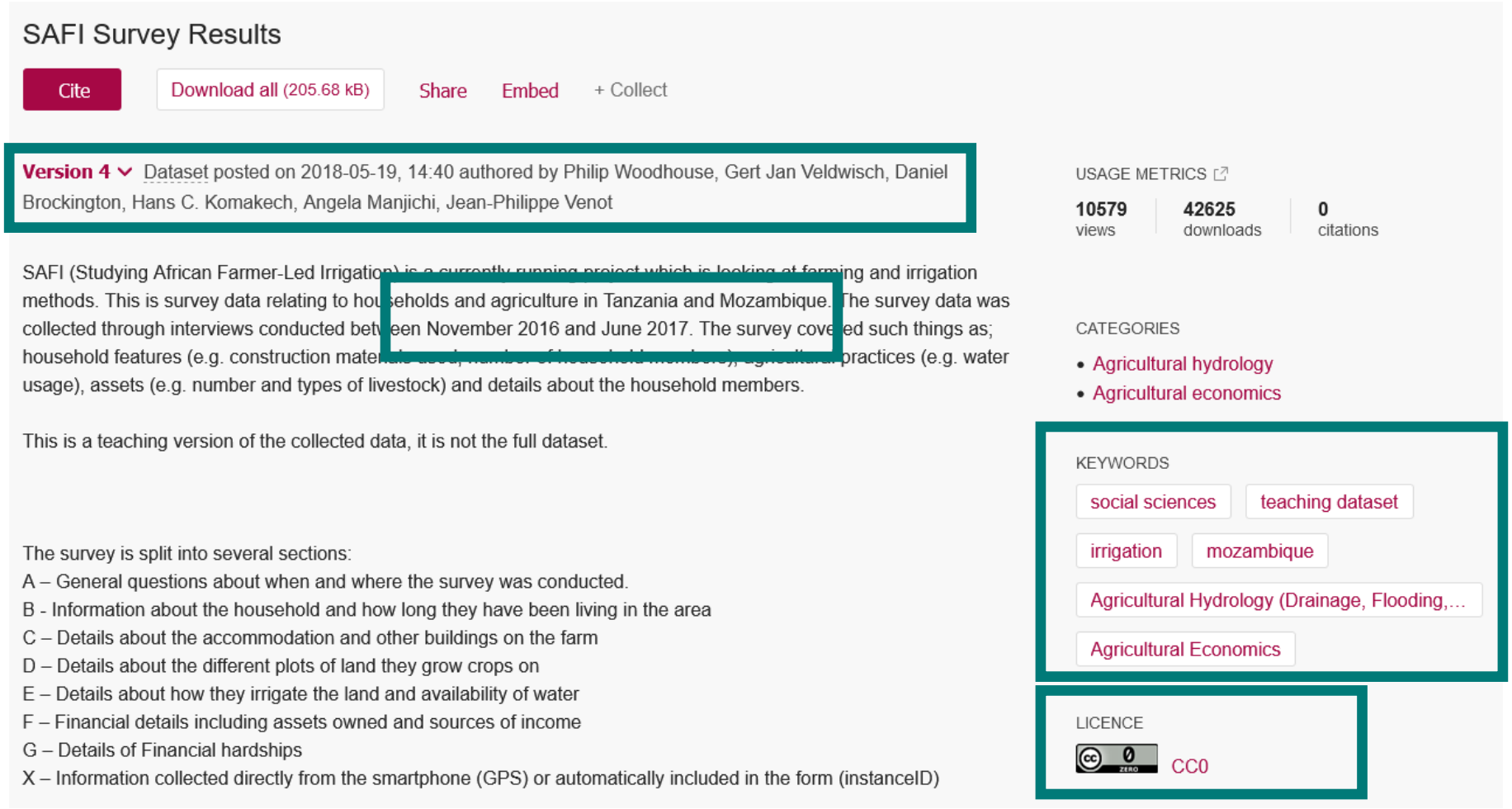


Image by Piotr Kononow

10.5281/zenodo.5053596

SAFI metadata



SAFI Survey Results

Cite | Download all (205.68 kB) | Share | Embed | + Collect

Version 4 Dataset posted on 2018-05-19, 14:40 authored by Philip Woodhouse, Gert Jan Veldwisch, Daniel Brockington, Hans C. Komakech, Angela Manjichi, Jean-Philippe Venot

USAGE METRICS

10579	42625	0
views	downloads	citations

SAFI (Studying African Farmer-Led Irrigation) is a currently running project which is looking at farming and irrigation methods. This is survey data relating to households and agriculture in Tanzania and Mozambique. The survey data was collected through interviews conducted between November 2016 and June 2017. The survey covered such things as; household features (e.g. construction materials used, number of household members), agricultural practices (e.g. water usage), assets (e.g. number and types of livestock) and details about the household members.

This is a teaching version of the collected data, it is not the full dataset.

CATEGORIES

- Agricultural hydrology
- Agricultural economics

KEYWORDS

social sciences | teaching dataset | irrigation | mozambique | Agricultural Hydrology (Drainage, Flooding,... | Agricultural Economics

The survey is split into several sections:
A – General questions about when and where the survey was conducted.
B - Information about the household and how long they have been living in the area
C – Details about the accommodation and other buildings on the farm
D – Details about the different plots of land they grow crops on
E – Details about how they irrigate the land and availability of water
F – Financial details including assets owned and sources of income
G – Details of Financial hardships
X – Information collected directly from the smartphone (GPS) or automatically included in the form (instanceID)

LICENCE

CC0

Studying African Farmer-led Irrigation (SAFI) Dataset

10.5281/zenodo.5053596

Metadata Standards

- Ensures **interoperability** and machine readability
- Example: YYYYMMDD = RFC3339
- Use FAIRsharing to look for standards in your field

Slide adapted from Christine Staiger https://doi.org/10.5281/zenodo.2585691

Vocabulary: how to say Female

18-day pregnant females
2 yr old female
400 yr. Old female
Adult female
Asexual female
Castrate female
Cf.female
Cystocarpic female
Dikaryon
Dioecious female
Diploid female
F
Famale
Femail
Female
Female – worker
Female (alate sexual)
Female (calf)
Female (lactating)
Female (pregnant)
Female (outbred)
Female parent
Female plant
Female with eggs
Female worker
Female, 6-8 weeks old
Female, virgin
Female, worker
Female(gynoecious)
Femele
Female, pooled
Femalen
Females
Females only
Gynoecious
Healthy female
Probably female
Individual female
Igb*cc females
Mare
Female (worker)
Monosex female
Ovigerous female
Oviparous sexual females
Worker bee
Female enriched
Pseudohermaprhoditic female
Remale
Semi-engorged female
Sexual oviparous female
Sterile female worker
Strictly female
Tetraploid female
Thelytoky
Female (gynoecious)
Worker caste 'female'
Sex female
Female, other
Female child
Femal
3 female
Female (phenotype)
Female mice
Female, spayed
Femlale
Metafemale
Sterile female
Normal female
Sf
Vitellogic replate female
Worker
Hexaploid female
Female (f-o)
Hen

10.5281/zenodo.5053596

Documentation

- README files
- (Electronic) Labbooks
- Guide for data documentation
- Data Dictionary
- Code Book
  - See the dataset info page

https://twitter.com/thecrobe/status/1373590641012322306

Exercise
5 min

1. **Download** the clean version of the SAFI dataset
   - **Excel**: Open a new/empty workbook in Excel
   - *Libre: select the right options in Text Import and click Ok*
2. Select '**Data**/**Gegevens**' on the ribbon, and then '**Gegevens ophalen**' -> '**From Text**/**CSV**'
3. Select the SAFI dataset and click **Import**/**Laden**
4. **Earlier Excel versions only**: In the **Text import wizard,** ensure the **'Delimited'** option (step 1), Tick **'Comma'** (step 2) and in step 3, keep **'General'** ticked and '**Finish'**

Follow along with screenshots or *Libre: Importing and Exporting CSV Files*

Exercise
5 min

- Check out the clean SAFI dataset
  - This data has more variables than the messy spreadsheet and is formatted according to **tidy data principles**

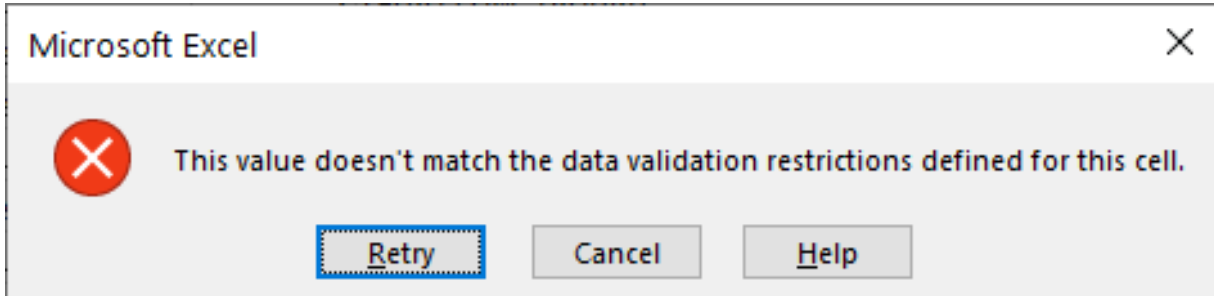
- **Discuss this data with a partner** and make a list of some of the types of metadata that should be recorded about this dataset
  - What is not immediately obvious to me about this data?
  - What questions would I need to know the answers to in order to analyze and interpret this data?

10.5281/zenodo.5053596

Quality Assurance - data validation

Excel allows us to specify a variety of **data validations** to be applied to cell contents.

If the validation fails, **an error is raised** and the data we entered does not go into the particular cell.



Excel support page on data validation

10.5281/zenodo.5053596

Data validation - restricting data to a numeric range

- column no_membrs = number of people in the household
  - positive integer
  - reasonable maximum value

Clean version of the SAFI dataset

Exercise

3 min

Data validation numeric range

1. **Use** the clean version of the SAFI dataset
2. Select the **no_membrs** column
3. In **Data/Gegevens tab** select **Data Tools** and then **Data Validation/Gegevensvalidatie** *(Libre: Data menu, select Validity…)*
4. Select **'Whole number/Geheel getal'** from the **Allow/Toestaan** drop down options *(Libre: Allow: Whole Numbers and then Data: valid range)*
5. **Minimum** and **Maximum** boxes appear (1-30, click **Ok**)
6. Enter a message in the **Input Message/Invoerbericht** tab *(Input Help)*

(You can change the error/warning in the **Style/Stijl** option on the **Error Alert/Foutmelding** tab.)

Follow along with screenshots

Data validation

10.5281/zenodo.5053596

- **Data validation rules are not applied retroactively** (data already in cells)
- **Existing data will not be flagged** with a warning
  - In some Excel versions, you can click in the Data tab on **Data Validation**/Gegevensvalidatie and then "**Circle invalid data**/Ongeldige gegevens omcircelen". This will put red circles around invalid data entries.
- **Set up data validation rules** for each column when you set up your spreadsheet (before data collection)

Clean version of the SAFI dataset

10.5281/zenodo.5053596

Data validation - restricting data to entries from a list

drop-down list of the available items

- **Typing a list** of values where only a few possible values exist (like "grass, muddaub, burntbricks, sunbricks, cement")

OR

- Create a **small table** (in a separate tab of the workbook), and use these cells as the source of acceptable inputs.
  - Makes the data entry process more flexible. If you add or remove contents from the table, these are immediately reflected in any new cell entries based on this source. You can also have different cells refer to the same table of acceptable inputs.

Clean version of the SAFI dataset

Exercise
3 min
Data validation lists

1. **Use** the clean version of the SAFI dataset
2. Select the **respondent_wall_type** column
3. In **Data tab** select **Data Tools** and then **Data Validation/Gegevensvalidatie** *(Libre: Data menu, select Validity…)*
4. Select **List/Lijst** from the **Allow/Toestaan** drop-down menu *(Libre: choose the List option),* to pop up a **Source/Bron** box
5. Type a list of all the values that you want to be accepted in this column, separated by a comma (without spaces!) or semicolon
   - grass; muddaub; burntbricks; sunbricks; cement
     - *Libre: enter the words on separate lines*
     - Excel: create a meaningful input message, then click **OK**

Follow along with screenshots

10.5281/zenodo.5053596

Exporting Data in an open format

Tab/Comma delimited

Comma Separated Value files (.csv) are plain text files where the columns are separated by commas

10.5281/zenodo.5053596

Open Format - Why?

Accessible

Useable by multiple types of software that are freely available (Microsoft requires a paid license)

Publishing

Data repositories, journals and funding agencies may have requirements for open formats

Life expectancy

Open formats are more suitable for long term preservation because they don't rely on a single software product

Universal format

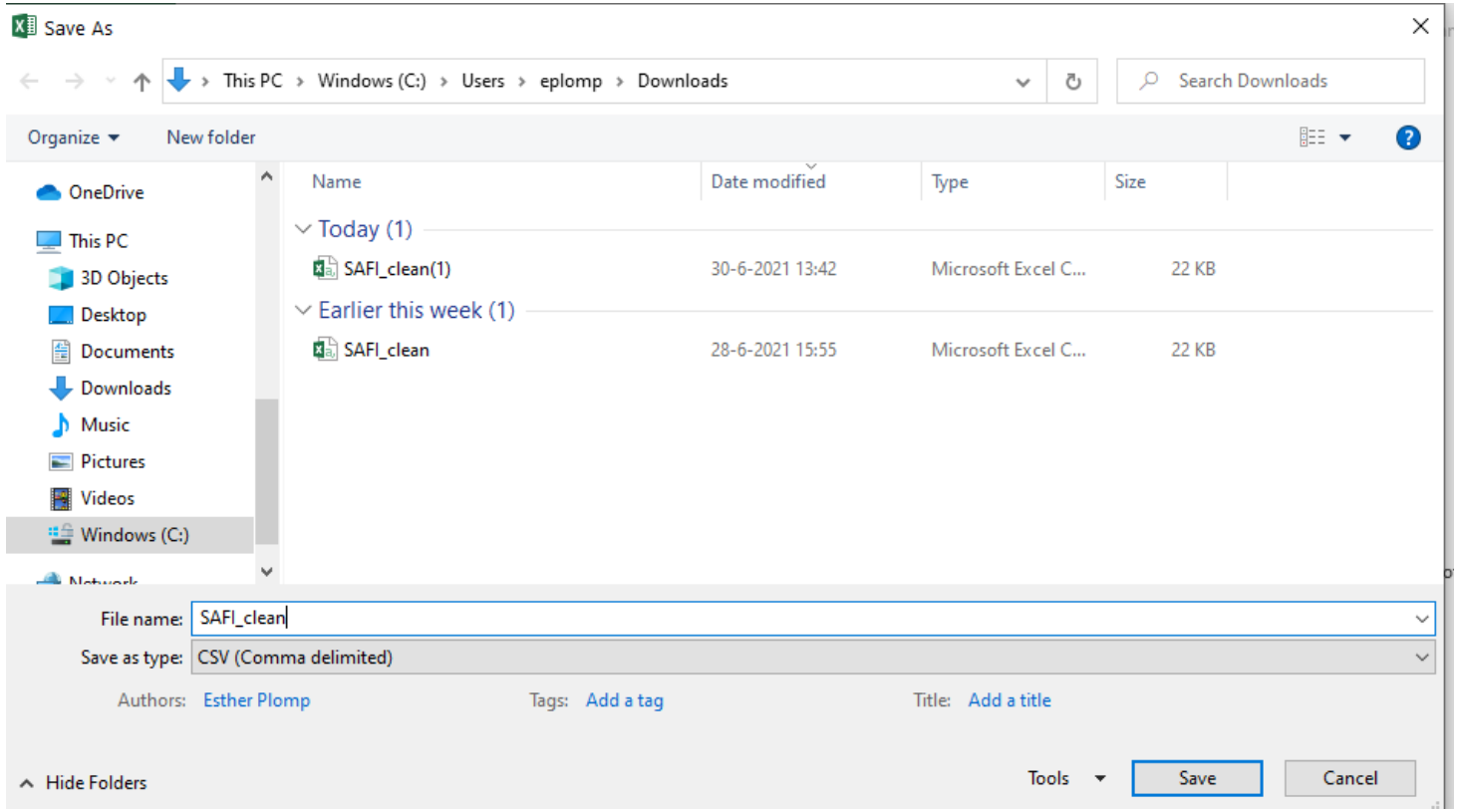
Interoperable format that produces the same results when it is imported by various software (including plain text editors and R)!

Exercise

2 min

export to .csv

1. From the top menu select **File** and **Save as**.
2. In the **Format field**, from the list, select **Comma Separated Values (*.csv)**.
3. Double check the file name and the location where you want to save it and select **Save**.



10.5281/zenodo.5053596

Thank you!

Esther Plomp
@toothFAIRy@scholar.social
e.plomp@tudelft.nl

Data Organization in Spreadsheets for Social Scientists

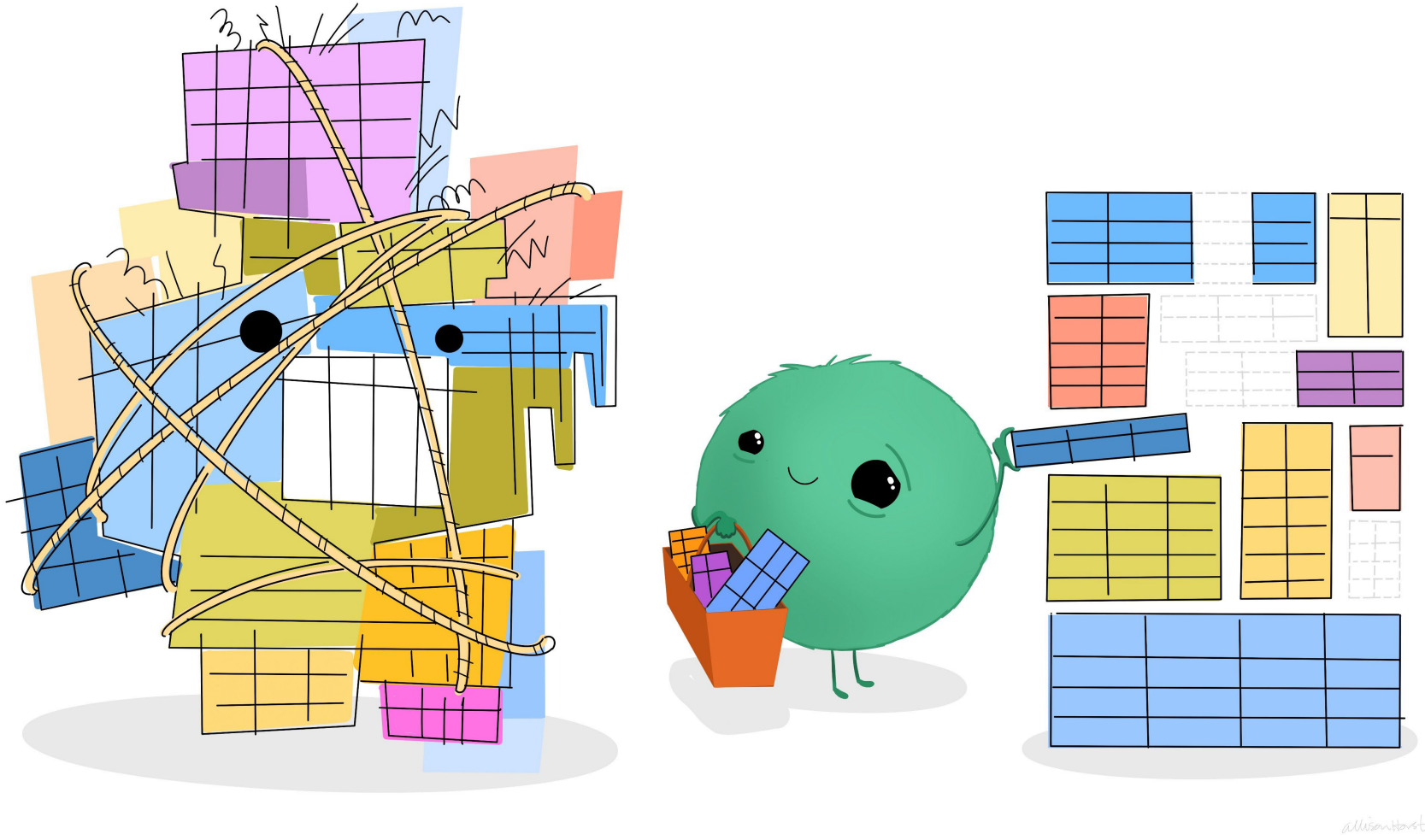


Illustration by Allison Horst (CC-BY)

10.5281/zenodo.5053596

Data with comma's

Enclose the data fields with double quotes

Double check that the file you are exporting can be read in correctly!

10.5281/zenodo.5053596

Extra resources

- **Ten Guidelines for Better Tables** by Jonathan A. Schwabish
- **Reproducible RDM workflows for tabular data** by Eirini Zormpa

https://twitter.com/blakeaburge/status/1535594636969902081