Data Organization in Spreadsheets for Social Scientists: Glossary

Key Points

Introduction	• Organizing your data tables according to tidy data principles will make them easier for you and others to use for analysis.
Formatting data tables in Spreadsheets	• Never modify your raw data. Always make a copy before making any changes. • Keep track of all of the steps you take to clean your data. • Organize your data according to tidy data principles. • Record metadata in a separate plain text file.
Formatting problems	• Avoid using multiple tables within one spreadsheet. • Avoid spreading data across multiple tabs (but do use a new tab to record data cleaning or manipulations). • Record zeros as zeros. • Use an appropriate null value to record missing data. • Don't use formatting to convey information or to make your spreadsheet look pretty. • Place comments in a separate column. • Record units in column headers. • Include only one piece of information in a cell. • Avoid spaces, numbers and special characters in column headers. • Avoid special characters in your data.
Dates as data	• Use extreme caution when working with date data. • Splitting dates into their component values can make them easier to handle.
Quality assurance	• Always copy your original spreadsheet file and work with a copy so you don't affect the raw data. • Use data validation to prevent accidentally entering invalid data.
Exporting data	• Data stored in common spreadsheet formats will often not be read correctly into data analysis software, introducing errors into your data. • Exporting data from spreadsheets to formats like CSV or TSV puts it in a format that can be used consistently by most programs.

Digi SYSKA session 1:
Get your sheet together

Glossary

cleaned data
: data that has been manipulated post-collection to remove errors or inaccuracies, introduce desired formatting changes, or otherwise prepare the data for analysis

conditional formatting
: formatting that is applied to a specific cell or range of cells depending on a set of criteria

CSV (comma separated values) format
: a plain text file format in which values are separated by commas

factor
: a variable that takes on a limited number of possible values (i.e. categorical data)

metadata
: data which describes other data

null value
: a value used to record observations missing from a dataset

observation
: a single measurement or record of the object being recorded (e.g. the weight of a particular mouse)

plain text
: unformatted text

quality assurance
: any process which checks data for validity during entry

quality control
: any process which removes problematic data from a dataset

raw data
: data that has not been manipulated and represents actual recorded values

rich text
: formatted text (e.g. text that appears bolded, colored or italicized)

string
: a collection of characters (e.g. "thisisastring")

TSV (tab separated values) format
: a plain text file format in which values are separated by tabs

variable
: a category of data being collected on the object being recorded (e.g. a mouse's weight)