

Tidy and messy data

FRES seminar 26.06.25



Content

- **Why**
- **Data structure:** *Tidy* and structured data
- **Data format**
- **Data storage**



If it works, why care?



*~80% of data analysis is spent cleaning and preparing data**

- Future you may thank you
- Facilitate collaboration and sharing
- **FAIR**
- Future & present IMR and publishing requirements



* Wickham, H. (2014) J. Stat. Soft. 59. doi:10.18637/jss.v059.i10

Structure: The Anna Karenina principle

- *All happy families are alike; each unhappy family is unhappy in its own way — Leo Tolstoy (1877)*
- ***Tidy datasets*** are all alike, but every messy dataset is messy in its own way — Hadley Wickham (2000s)



Structure: tidy data

country	year	cases	population
Afghanistan	1999	745	1980071
Afghanistan	2000	666	2059360
Brazil	1999	3737	17200362
Brazil	2000	488	17450398
China	1999	21258	127201272
China	2000	1766	128042583

variables

country	year	cases	population
Afghanistan	1999	745	1980071
Afghanistan	2000	666	2059360
Brazil	1999	3737	17200362
Brazil	2000	488	17450398
China	1999	21258	127201272
China	2000	1766	128042583

observations

country	year	cases	population
Afghanistan	99	745	1980071
Afghanistan	00	666	2059360
Brazil	99	3737	17200362
Brazil	00	488	17450398
China	99	21258	127201272
China	00	1766	128042583

values

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table



Structure: Messy example: nested columns

Prøve	Type	Dato	Kommentar	Aluminium			Jern		
				# Partikler	np_konsentr asjon	np_deteksjonsgre nse	# Partikler	np_konsentr asjon	np_deteksjonsgre nse
instrument_blank_1	blank	07/07/2021	bare MQ-	0	0.01	40	4	0.09	10
standard_100ng/L	standard	07/07/2021	i 0.1% HNO3	0	100.3	40	2	98.4	10
blåskjell_23_Q3	prøve	07/07/2021	ekstra homogenise rt	422	14.02	40	242	500.2	10

```
1 #Extracts isotope information from the headers of the masshunter file.
2 #NOTE, FIRST COLUMN STARTS WITH # !
3 #NOTE Sample names must be UNIQUE!
4 #NOTE MUST BE FIXED FOR MASS-SHIFT/ 8380.
5 get_isotopes <- function(file) {
6   read_lines(file, n_max = 1) %>%
7   str_replace_all(c("#" = "", " " = " "))
8   '197' -> '197 Au',
9   '197' -> '197 Au',
10  '(?<=\\d){1,3}(\\d{2,3})(?<=\\s|\\s{2})' = '') %>%
11  str_extract_all("(\\d{1,3}|\\d{2,3}|\\d{4})") %>% #match digit or space x 3
12  unlist() %>% #then word char. OR
13  set_names(seq_along(.) - 1)
14 }
15
16 reshape_masshunter_df <- function(file, delim = ",", decimal_mark = ".") {
17   isotopes <- get_isotopes(file)
18   read_delim(file = file, delim = delim,
19             locale = locale(decimal_mark = decimal_mark),
20             skip = 1) %>%
21   select(-matches("(^X$")) %>%
22   mutate(across(
23     starts_with("np"), ncol(),
24     ~ str_remove_all(., "[<=>"] %>%
25     as.numeric()
26   )) %>%
27   rename_with(
28     ~ str_replace_all(., c(
29       "(?<=[a-z])(?=[A-Z])" = "_",
30       " " = " ",
31       "\\." = "."
32     )) %>%
33     str_to_lower()
34   ) %>%
35   rename_with(
36     str_replace_all,
37     pattern = c("$" = "_ID"),
38     cols = starts_with("np.") & !matches("\\d{1,3}")
39   ) %>%
40   rename_with(str_replace_all, pattern = c("(?<=\\d)$" = "_")) %>%
41   pivot_longer(
42     matches("_"),
43     names_to = c("variable", "isotope"),
44     names_sep = "_",
45     values_to = "value"
46   ) %>%
47   pivot_wider(
48     names_from = "variable",
49     values_from = "value"
50   ) %>%
51   arrange(isotope) %>%
52   mutate(isotope = isotope[as.character(isotope)]) %>%
53   mutate(
54     dataset = str_extract(file, "(?<=\\s|\\s{2}).*"), # dataset identifier
55     .before = 1) %>%
56   mutate(
57     sample_name = str_replace_all(
58       sample_name,
59       pattern = "[i]", replacement = "") #fix encoding problems
60   )
61 }
```

Tidy data

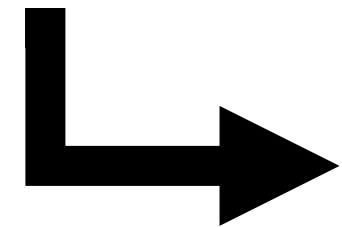


60 linjer R-kode

Structure: Messy example: nested columns

				Aluminium		Jern			
Prøve	Type	Dato	Kommentar	# Partikler	np_konsentr nse	np_deteksjonsgre nse	# Partikler	np_konsentr asjon	np_deteksjonsgrense
instrument_blank_1	blank	07/07/2021	bare MQ- vann	0	0.01	40	4	0.09	10
standard_100ng/L	standard	07/07/2021	i 0.1% HNO3	0	0.3	40	2	98.4	10
blåskjell_23_Q3	prøve	07/07/2021	ekstra homogenise rt	422	14.02	40	242	500.2	10

Messy and may require hours of coding to read



~60 lines of code

Prøve	Type	Dato	Kommentar	Element	# Partikler	np_konsentr asjon	np_deteksjonsg rense
instrument_blank_1	blank	07/07/2021	1 bare MQ-vann	Alumimium	0	0.01	40
instrument_blank_1	blank	07/07/2021		Jern	4	0.09	10
standard_100ng/L	standard	07/07/2021		Alumimium	0	100.3	40
standard_100ng/L	standard	07/07/2021		Jern	2	98.4	10
blåskjell_23_Q3	prøve	07/07/2021	1 ekstra homogenisert	Alumimium	422	14.02	40
blåskjell_23_Q3	prøve	07/07/2021	1 ekstra homogenisert	Jern	242	500.2	10

TIDY and machine-readable



Format: *dictated by purpose*

- Excel is primarily for visualizations and analysis
- CSV is recommended for storage and exchange*

.CSV

a text file

- + Open and simple
- + (more) compatible
- + Efficient
- Limited possibilities

EXCEL (xls/xlsx)

a program (source code + data)

- Proprietary and complicated
- ± Automatic formatting
- + Tools for analysis and visualisation



Microsoft Excel blamed for gene study errors

© 25 August 2016

*Error in 1/5 of studies using excel!***

... also happens e.g. for CAS nrs



*Publications Office of the European Union. (2021). Data.europa.eu data quality guidelines. Publications Office. <https://data.europa.eu/doi/10.2830/79367>.

**<https://www.bbc.com/news/technology-37176926>

Storage: name

Goal:

- Reproducibility (short and longterm)
 - Findability
 - history/lineage
- Machine readability

Action: consistent naming

- Avoid space and special characters:
 - CapitalizeMyName.csv or use_underscore.csv
- YYMMDD_docname_vYYMMDD
- 250625_pres_sommerfest.csv



Storage: location

- **File servers** (ces/delphi): rawdata, big files, media
- **Onedrive**: private (IMR do not have access)
- **Teams** (IMR is owner – powerful search, AI, collaboration)
- *Owncloud (status?)*



Takk!



Resources

- <https://data.europa.eu/apps/data-visualisation-guide/intro-to-tidy-data> (EU info and guidelines on all things data)
- <https://r4ds.hadley.nz/data-tidy.html>
- <https://www.hi.no/resources/imr/Data-policy-IMR.pdf>
- HI har ingen veileder for datalagring, men kommer? (ref NMD)
- Wickham, H. (2014) *J. Stat. Soft.* **59**. doi:[10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)

