washr an R-package to facilitate FAIR data publishing

FAIR Data Publishing

Lars Schöbitz

Nov 28, 2024

Data

✋ hands up ✋

Who brought data?

✋ hands up ✋

Who has got “A one line description of the data”?

✋ hands up ✋

Who has got “A one paragraph, three sentence description of the data”?

✋ hands up ✋

Who has got a “Data dictionary, stored as dictionary.xlsx. Two columns, describing each variable in the dataset you come with.”

Metadata

Metadata: data about data

WHAT!?

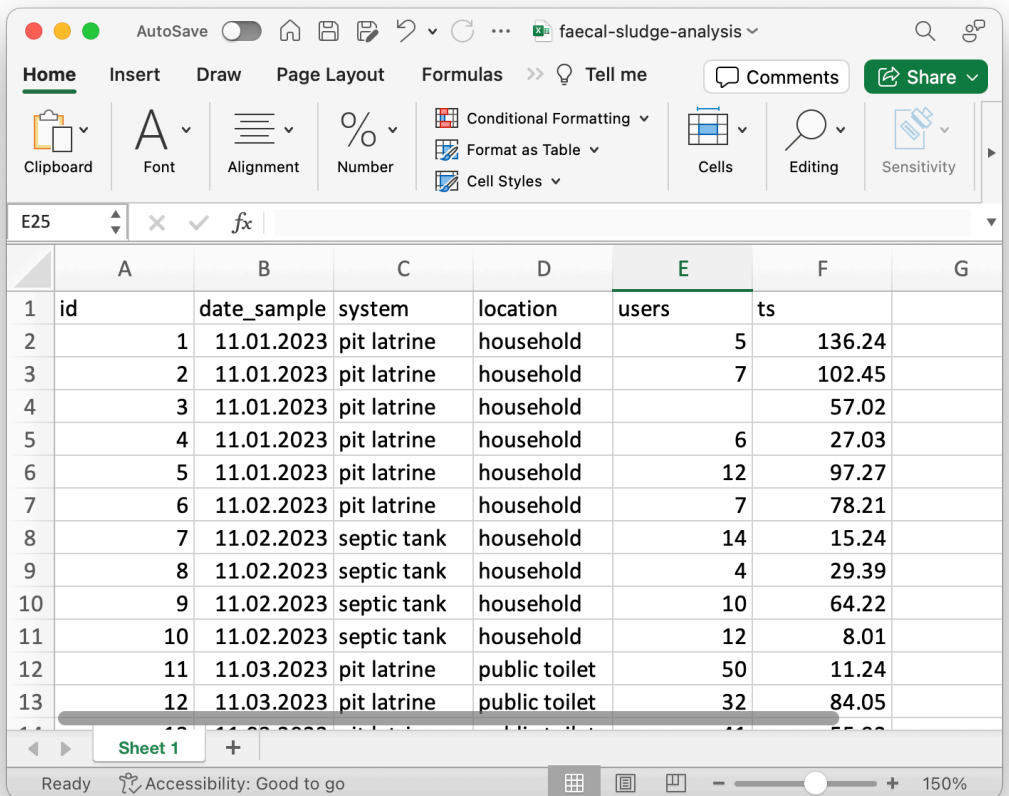
Faecal sludge samples

Imagine:

- you are new to WASH research and you have never heard of feacal sludge management.
- you are interested in learning more about the topic and you want to find some data to play with.
- you find a publication with a dataset on faecal sludge characteristics.

Faecal sludge samples

You download the XLSX file that contains the data and you open it in Excel. You see the following:

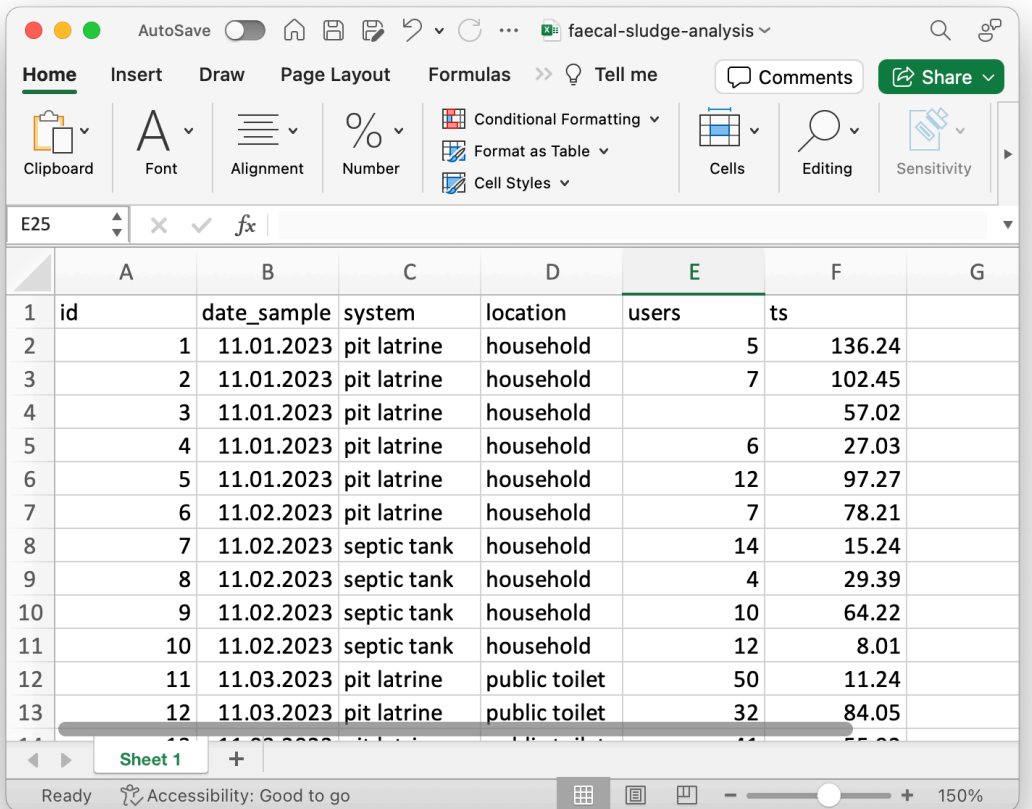


	A	B	C	D	E	F	G
1	id	date_sample	system	location	users	ts	
2	1	11.01.2023	pit latrine	household	5	136.24	
3	2	11.01.2023	pit latrine	household	7	102.45	
4	3	11.01.2023	pit latrine	household		57.02	
5	4	11.01.2023	pit latrine	household	6	27.03	
6	5	11.01.2023	pit latrine	household	12	97.27	
7	6	11.02.2023	pit latrine	household	7	78.21	
8	7	11.02.2023	septic tank	household	14	15.24	
9	8	11.02.2023	septic tank	household	4	29.39	
10	9	11.02.2023	septic tank	household	10	64.22	
11	10	11.02.2023	septic tank	household	12	8.01	
12	11	11.03.2023	pit latrine	public toilet	50	11.24	
13	12	11.03.2023	pit latrine	public toilet	32	84.05	

Faecal sludge samples

Open questions:

- What unit does `users` refer to?
- What does `ts` stand for?
- The `date` of what?
- Where was this data collected?
- Which method was used to collect the samples?



	A	B	C	D	E	F	G
1	id	date_sample	system	location	users	ts	
2	1	11.01.2023	pit latrine	household	5	136.24	
3	2	11.01.2023	pit latrine	household	7	102.45	
4	3	11.01.2023	pit latrine	household		57.02	
5	4	11.01.2023	pit latrine	household	6	27.03	
6	5	11.01.2023	pit latrine	household	12	97.27	
7	6	11.02.2023	pit latrine	household	7	78.21	
8	7	11.02.2023	septic tank	household	14	15.24	
9	8	11.02.2023	septic tank	household	4	29.39	
10	9	11.02.2023	septic tank	household	10	64.22	
11	10	11.02.2023	septic tank	household	12	8.01	
12	11	11.03.2023	pit latrine	public toilet	50	11.24	
13	12	11.03.2023	pit latrine	public toilet	32	84.05	

Questions that only the original author may have the answers to.

You as an author

have the chance to document your data properly once to make it easier for everyone else to know what it contains.

Documentation

Goes into a separate README file

- General information (authors, title, date, geographic location, etc.)
- Sharing / access information (license, links to publications, citation)
- Methodological information (sampling, analysis, etc.)

Data dictionary

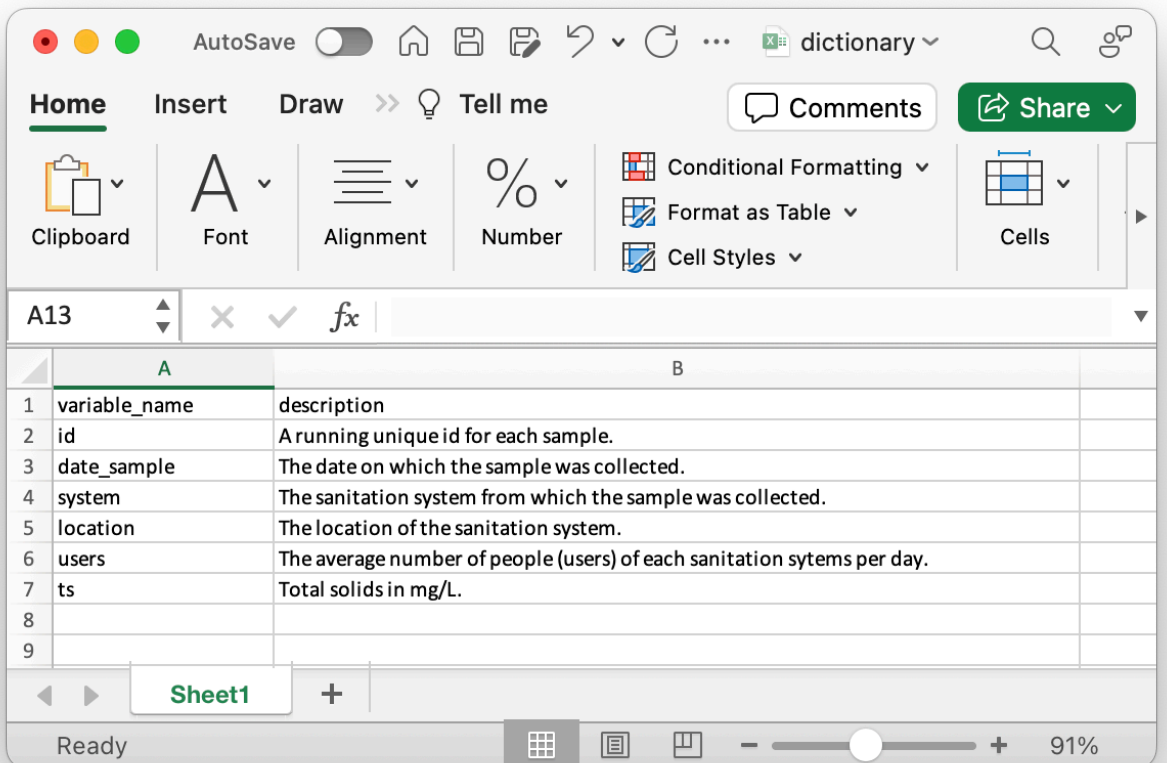
Goes into a separate file (`dictionary.csv`).

Minimum required information

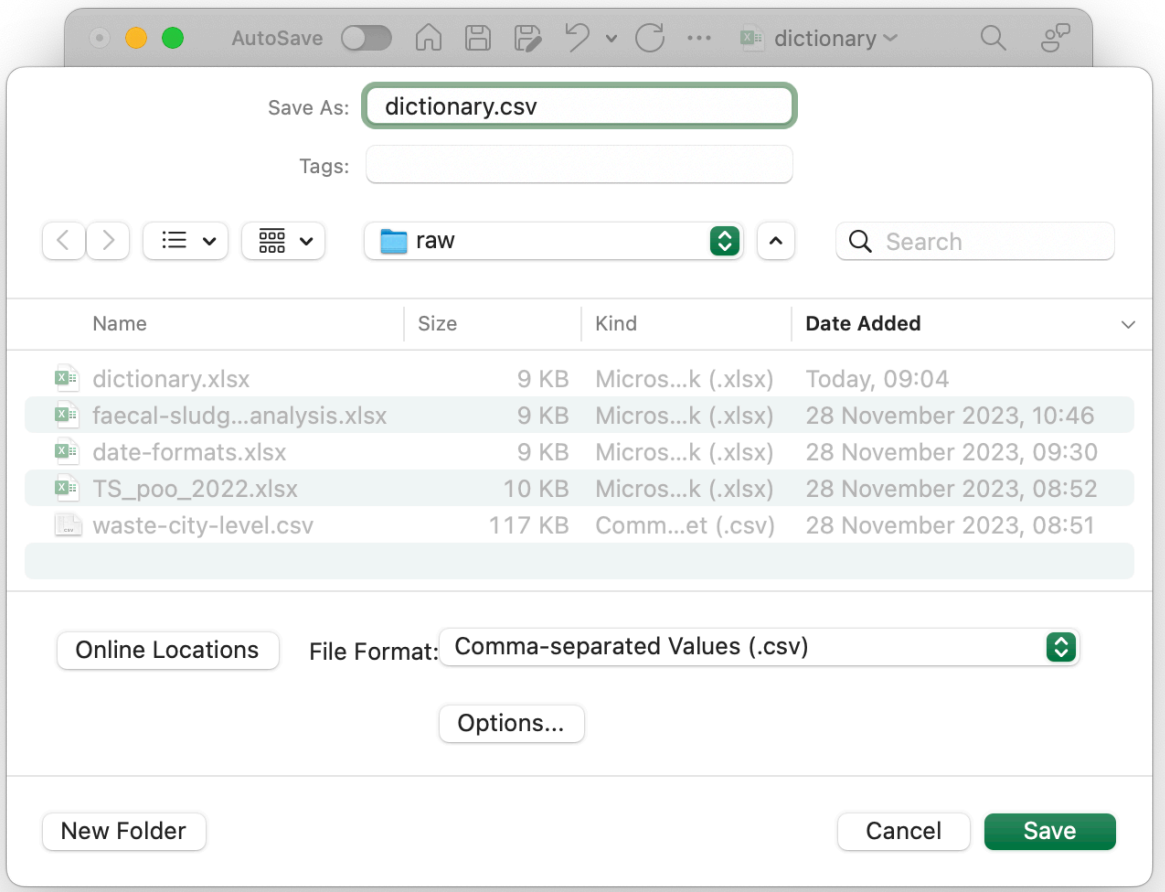
- Variable name
- Variable description

Data dictionary for faecal sludge samples

- Edit in spreadsheet software (e.g. MS Excel)



Data dictionary for faecal sludge samples



- Save as CSV file

Directory tree of an R data package

Capstone project of Rainbow Train: https://github.com/openwashdata/washopenresearch

```
.
├── CITATION.cff
├── DESCRIPTION
├── LICENSE.md
├── NAMESPACE
├── R
├── README.Rmd
├── README.md
├── _pkgdown.yml
├── data
├── data-raw
├── docs
├── inst
├── man
├── vignettes
└── washopenresearch.Rproj
```

Directory tree of an R data package

- `R` folder: R scripts with documentation of data resources
- `data-raw` folder: raw data, contains `data_processing.R` script to prepare processed, analysis-ready data
- `data` folder: documented analysis-ready data in R-internal `.rda` format
- `README.Rmd`: R data package documentation, displayed on GitHub and website landing page

Inside the `data-raw` folder

```
data-raw/
├── data_processing.R
├── dictionary.csv
├── unc-article-url-manual-
collection.csv
└── washdev.csv
```

- `unc-article-url-manual-collection.csv`: one raw data resource
- `washdev.csv`: another raw data resource
- `data_processing.R`: where the work happens
- `dictionary.csv`: data dictionary describing the variables in the processed, analysis-ready data

My turn: A tour of washopenresearch

Sit back and enjoy!

fairdatapub-washcentre.github.io/website/

05:00

Your turn: Create a GitHub repository for your data

1. Navigate to the GitHub organisation for this course: github.com/fairdatapub-washcentre/
2. Click on the green “New” button to create a new repository.
3. Chose a name for your repository (all small letters, no spaces or dashes)
4. Make the repository public.
5. Keep everything else unchanged.
6. Click on green “Create repository” button.
7. Keep the tab open for the next steps.

Our turn: Create a project on Posit Cloud

1. Navigate to the course workspace on posit.cloud/spaces/588816/content/.
2. If you haven't yet, bookmark the page.
3. Follow my instructions on the screen.

Take a break

Enjoy your lunch! Let your emails rest in peace.

fairdatapub-washcentre.github.io/website/

05:00

Our turn: Work through Data Publishing Guide

For the rest of the day we will work through

fairdatapub-washcentre.github.io/website/

Your turn: Prepare a dataset on your own

1. Navigate to the course website: fairdatapub-washcentre.github.io/website/
2. In the left-hand menu, click on Module 3, then select am-03: Data Organization
3. Follow the instructions
4. Place a yellow sticky note on your laptop when you have completed the assignment

Module 3 documentation

fairdatapub-washcentre.github.io/website/modules/md-03.html

Wrap-up

Participate in our hackathon tomorrow

- All tools welcome
- Win up to 750 CHF (~ 850 USD)
- https://openwashdata.org/pages/events/2024-11-18-hackathon/
- Registration is still open: https://forms.office.com/pages/responsep
  id=7KY0lmaio0WrFHTEIR_FgjqEYzwAXr5PiYenM5_ZB_lURTU2UklP

Your turn: Sign up for the openwashdata newsletter!

fairdatapub-washcentre.github.io/website/

05:00

What's next?

Holiday break! I will be off from December 9th to January 15th.

fairdatapub-washcentre.github.io/website/

What's next?

Happy New Year! I will be at UKZN on Thursday, 16th January. I will leave South Africa on Thursday, 13th February.

Thanks! 🌻

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/

Access slides as PDF on GitHub