washr an R-package to facilitate FAIR data publishing

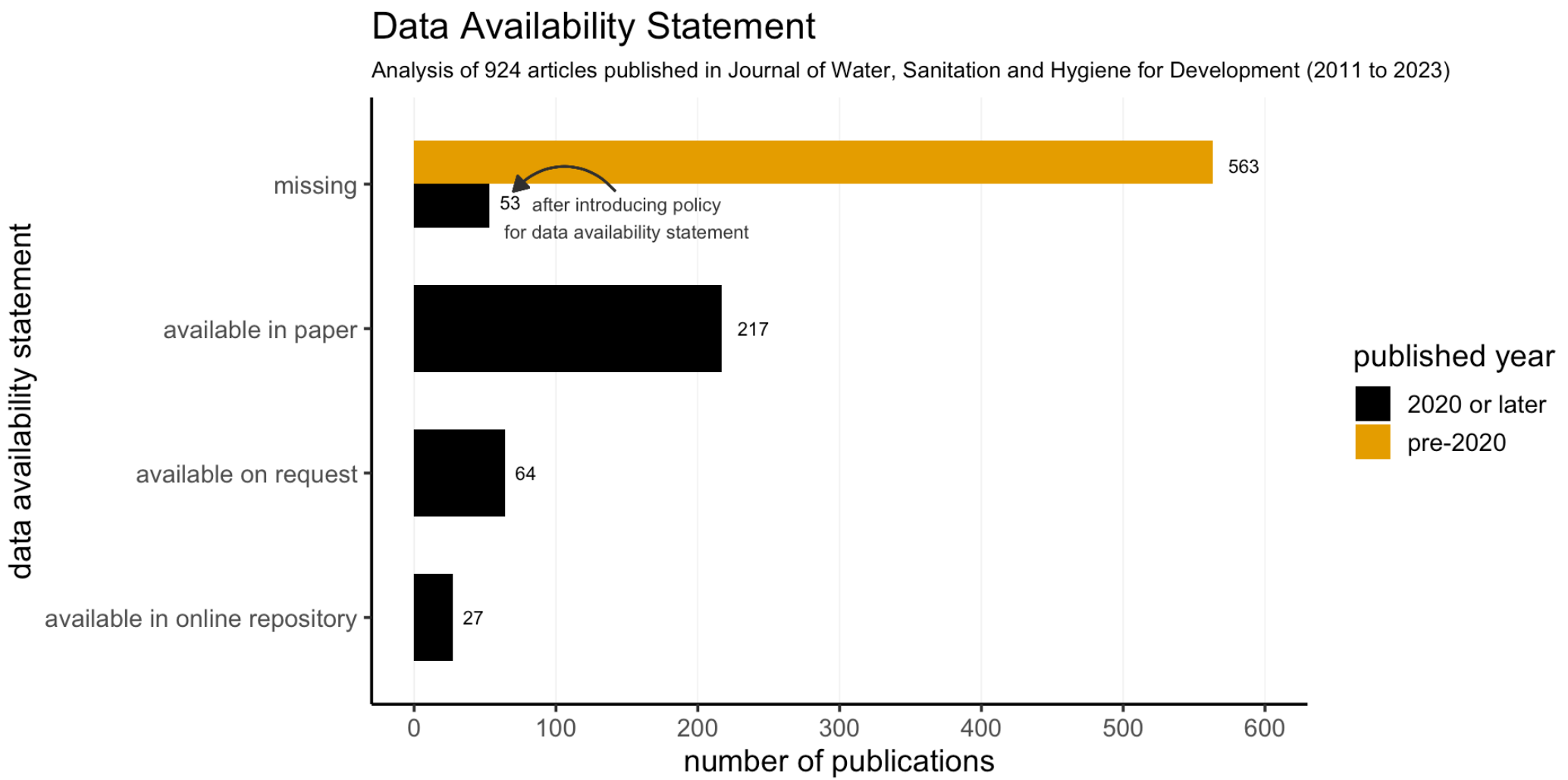
FAIR Data Publishing

Lars Schöbitz

Nov 28, 2024

The Opportunity

Journal articles



fairdatapub-washcentre.github.io/website/

Journal articles

Supplementary Material

Take-away: Not a single file is in machine-readable, non-proprietary file type format that would qualify for following FAIR principles for data sharing (Wilkinson et al. 2016).

Good practice: CSV file (comma-separated values), including a data dictionary for all variables/columns in the data

Supplementary Material

Articles published 2020 or later

file type	n[1]	%
missing	202	51.4
docx	149	37.9
xlsx	24	6.1
pdf	13	3.3
pptx	4	1.0
png	1	0.3

[1] One article can have multiple files.

openwashdata community

Vision

An active global community that applies FAIR principles (Wilkinson et al. 2016) to data generated in the greater water, sanitation, and hygiene sector.

Mission

Empower WASH professionals to engage with tools and workflows for open data and code.

openwashdata publishing

openwashdata.github.io/fsmglobal/

fsmglobal 0.0.1 Reference Articles Search for

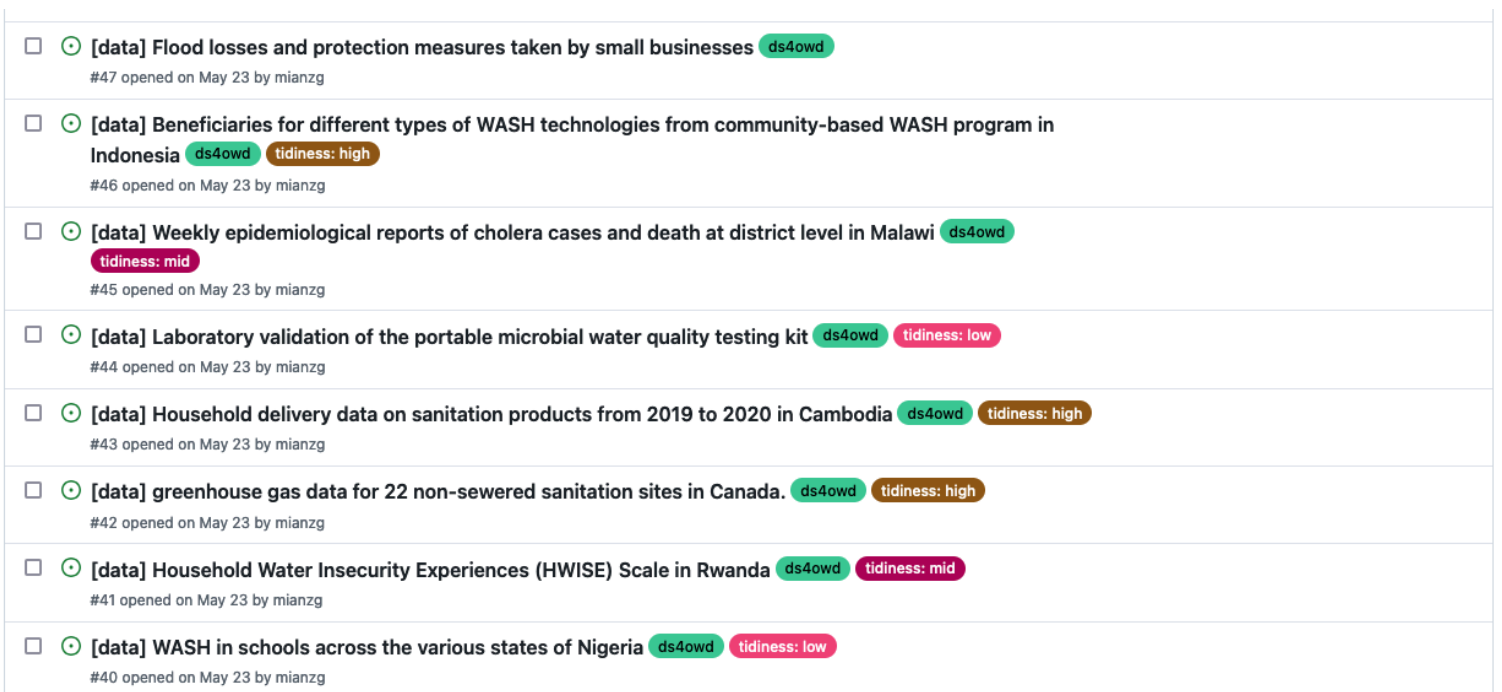
fsmglobal

This data was first published as part of a journal article by (Greene et al. 2021) and contained in the supplemental material as a table in a DOCX file. The following summary table was produced from the data and the code is shown further below.

Demand for faecal sludge emptying services		
summarised for 175 countries		
	population	**percent**
mechanized	1,030,317,694	25%
no facility	661,998,822	16%

openwashdata academy

- 10-week free data science course to empower WASH professionals to engage with tools and workflows for open data
- 200 registrations from 46 countries
- 27 datasets submitted as final projects



fairdatapub-washcentre.github.io/website/

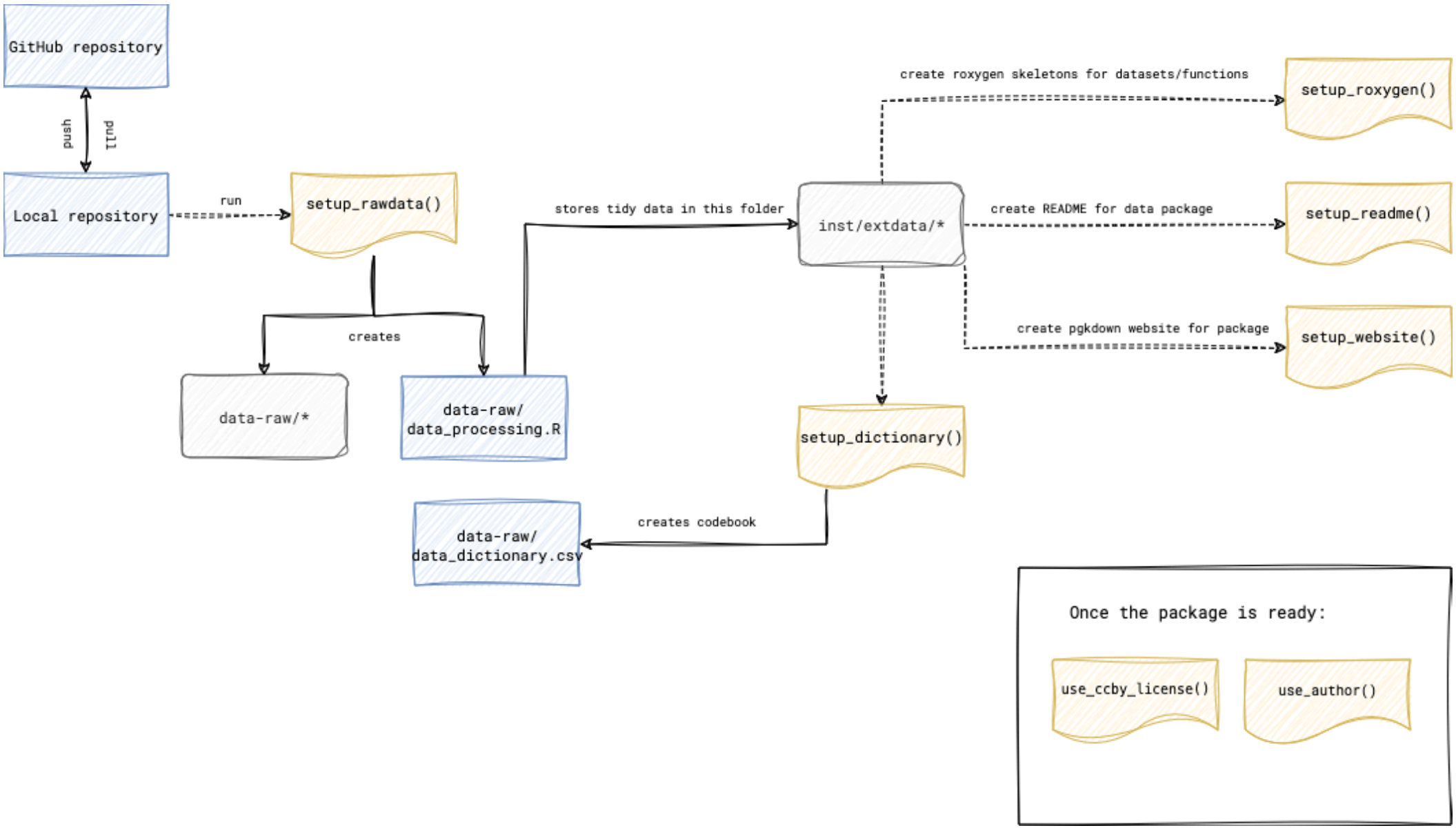
How can we streamline the data publishing procedure?

washr

- An R package designed to simplify WASH data publishing
- User-friendly functions to ensure that data adheres to FAIR principles
- Preparation of a detailed guide and workflow visualization

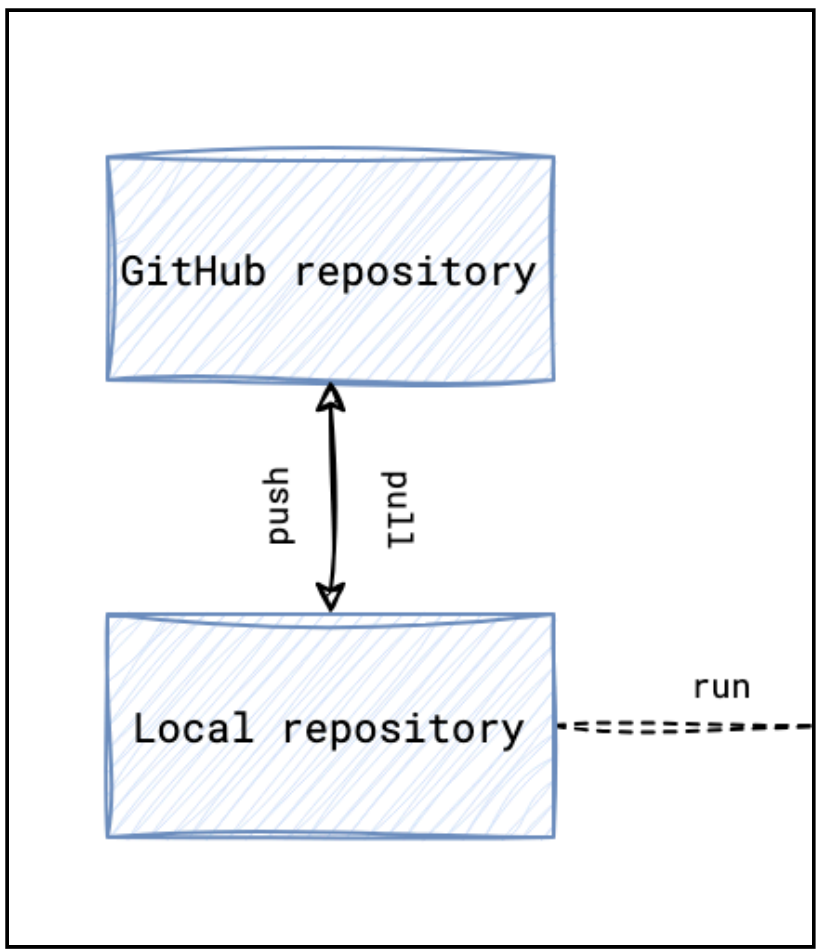
So far:

- Almost a dozen datasets published
- Requires minimal computational power
- Easily generalizable to benefit the wider community


GitHub repository
push
pull
Local repository
run
setup_rawdata()
creates
data-raw/*
data-raw/
data_processing.R
stores tidy data in this folder
inst/extdata/*
create roxygen skeletons for datasets/functions
setup_roxygen()
create README for data package
setup_readme()
create pgkdown website for package
setup_website()
setup_dictionary()
creates codebook
data-raw/
data_dictionary.csv
Once the package is ready:
use_ccby_license()
use_author()

Preparing the data

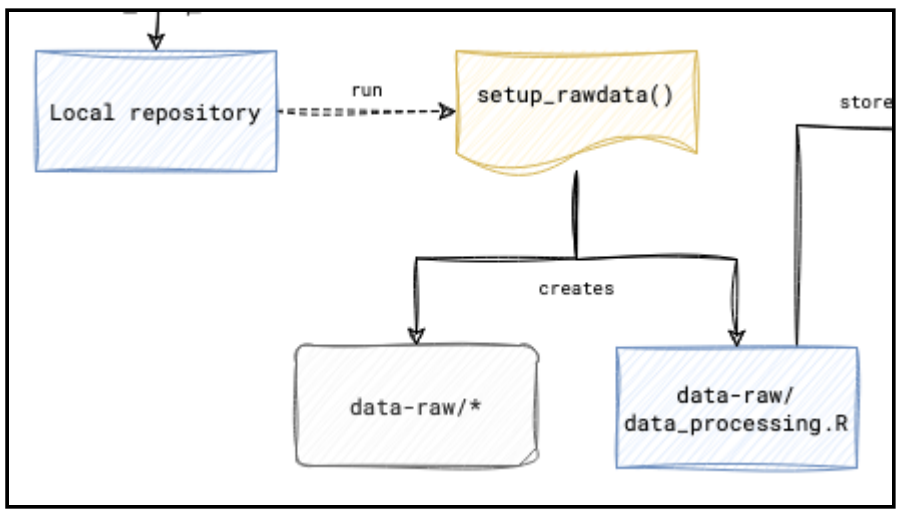
- Start a local (Posit Cloud for us) version-controlled folder, connect it to GitHub



fairdatapub-washcentre.github.io/website/

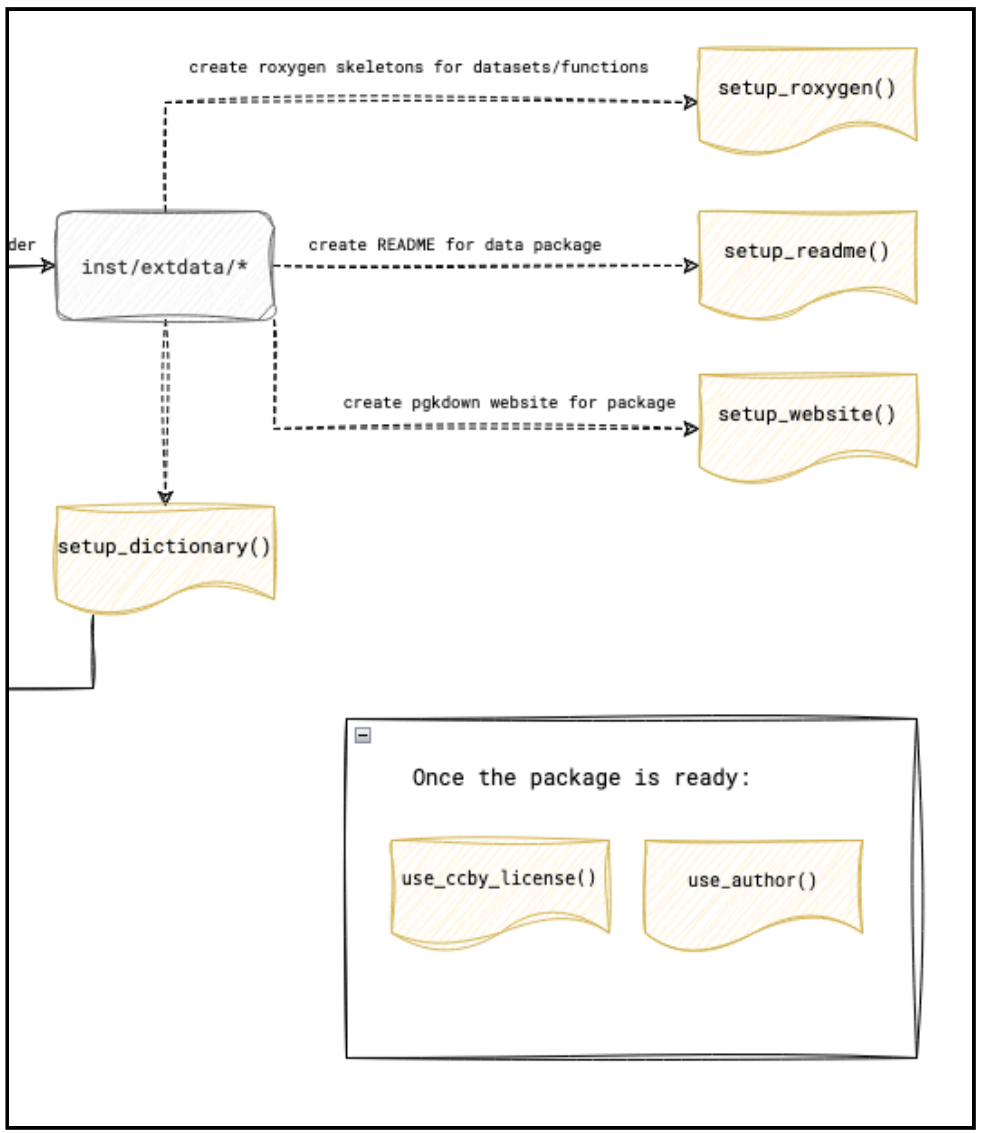
Preparing the data

- `setup_rawdata()`
  - Creates `data-raw` as suggested in `usethis` R Package[1]
  - Creates `data_processing.R` for data cleaning



Documenting the data

- Create roxygen skeletons
- Create README
- Codebook describing each variable
- Website with pkgdown R package[1]
- Add a license and author(s)



fairdatapub-washcentre.github.io/website/

Data Publishing Guide

global-health-engineering.github.io/ghedatapublishing/

Data Publishing with `washr`

Welcome

Publishing data can be challenging, especially when adhering to standards of reproducibility. Although technically available, an Excel workbook with multiple tabs, tucked away in an online archive, is far from practical. In other words, it lacks the key principles of being findable, accessible, interoperable, and reproducible — commonly known as FAIR.

This guide aims to provide a detailed walkthrough of how data can be published according to the FAIR principles. It builds on `washr`, an R package developed for swift data publication. The package emerged from the need to streamline certain steps when publishing the datasets collected during GHE's openwashdata academy.

The guide follows a chronological structure, starting from an empty repository and resulting in the publication of data as a website. 1 Creating a repository introduces version control, guiding readers through setting up both local and

Your turn: Bookmark the guide

1. Navigate to the Data Publishing Guide website: global-health-engineering.github.io/ghedatapublishing/
2. Bookmark the page and add it to your bookmarks folder.

Thanks!

fairdatapub-washcentre.github.io/website/

Links and Downloads

`washr` source code: https://github.com/openwashdata/washr

`washr` guide: https://global-health-engineering.github.io/ghedatapublishing

openwashdata: https://openwashdata.org

Download slides as PDF on GitHub

Sign up for the openwashdata newsletter!

fairdatapub-washcentre.github.io/website/

References

Greene, Nicola, Sarah Hennessy, Tate W. Rogers, Jocelyn Tsai, Francis L. de los Reyes III, and Lars Schöbitz. 2023. “Fsmglobal. Global Faecal Sludge Emptying Services Demand.” https://doi.org/10.5281/zenodo.8208293.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1). https://doi.org/10.1038/sdata.2016.18.