Publish with Ease

A Low-Cost, Open-Source Solution for Research Data Sharing

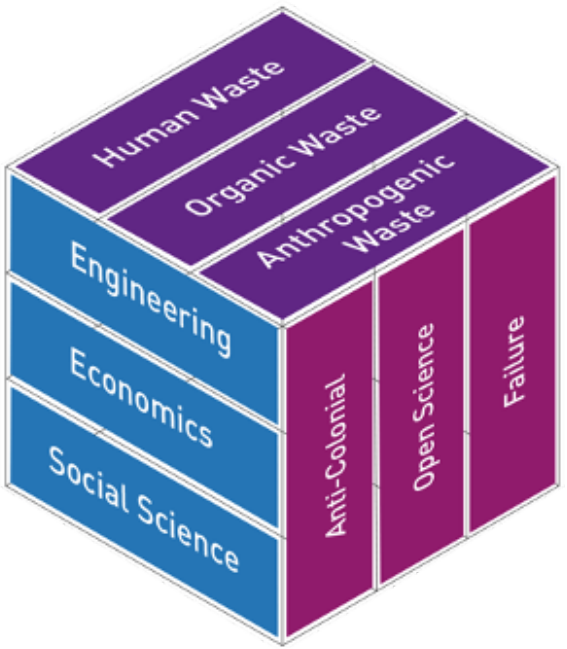
Colin Walder

Global Health Engineering - ETH Zurich

October 9, 2024

Introduction

- Data Steward @ Global Health Engineering (GHE), ETH Zürich
- *“Addressing the determinants of health as a function of engineered interventions and systems”.*
- One major component of health is WASH – **wa**ter, **s**anitation and **h**ygiene



One major problem

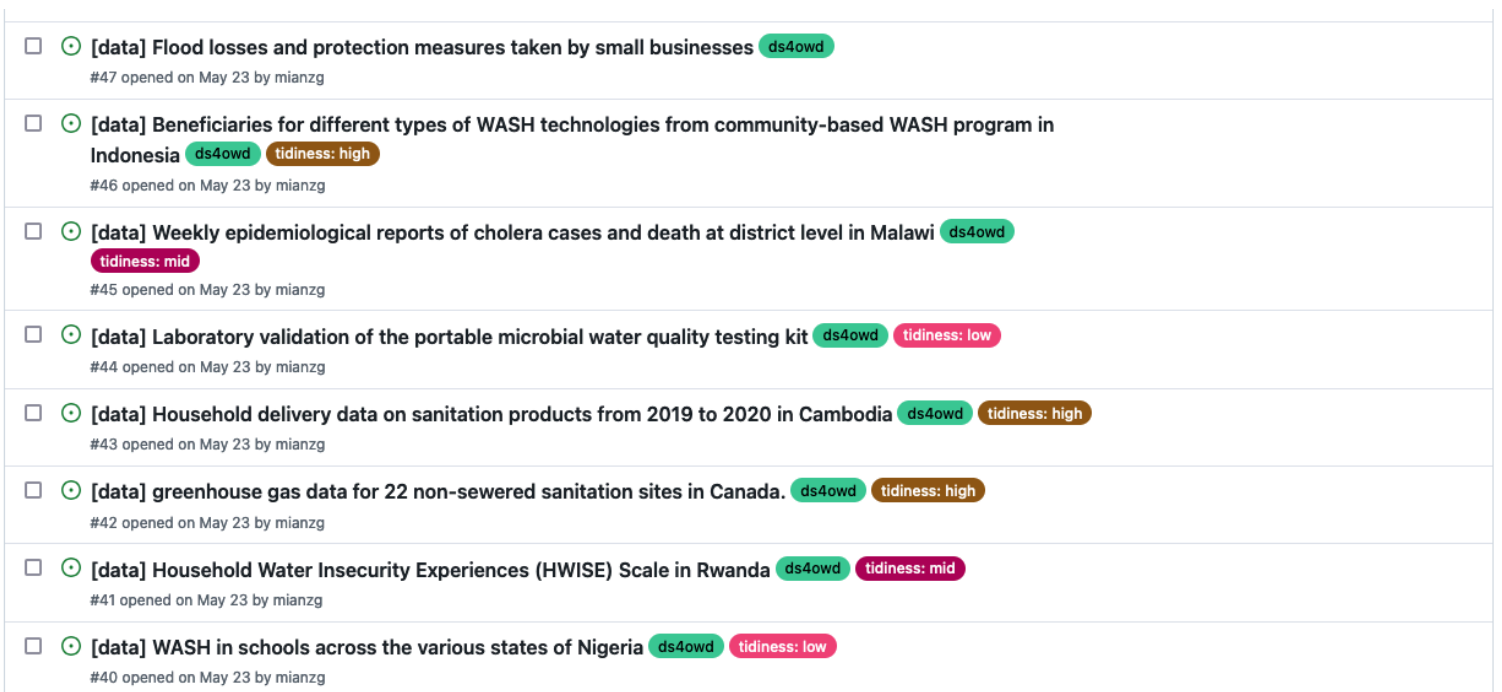
WASH practitioners often lack skills in computational data management

openwashdata

- Established in 2021, applying FAIR principles (Wilkinson et al. 2016) to data generated in the WASH sector
- **Empower WASH professionals to engage with tools and workflows for open data and code**
- Support other organizations with their data management, data events (e.g., hackathon), free data science course
- Core team: Two data stewards and one intern at GHE; many collaborators

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



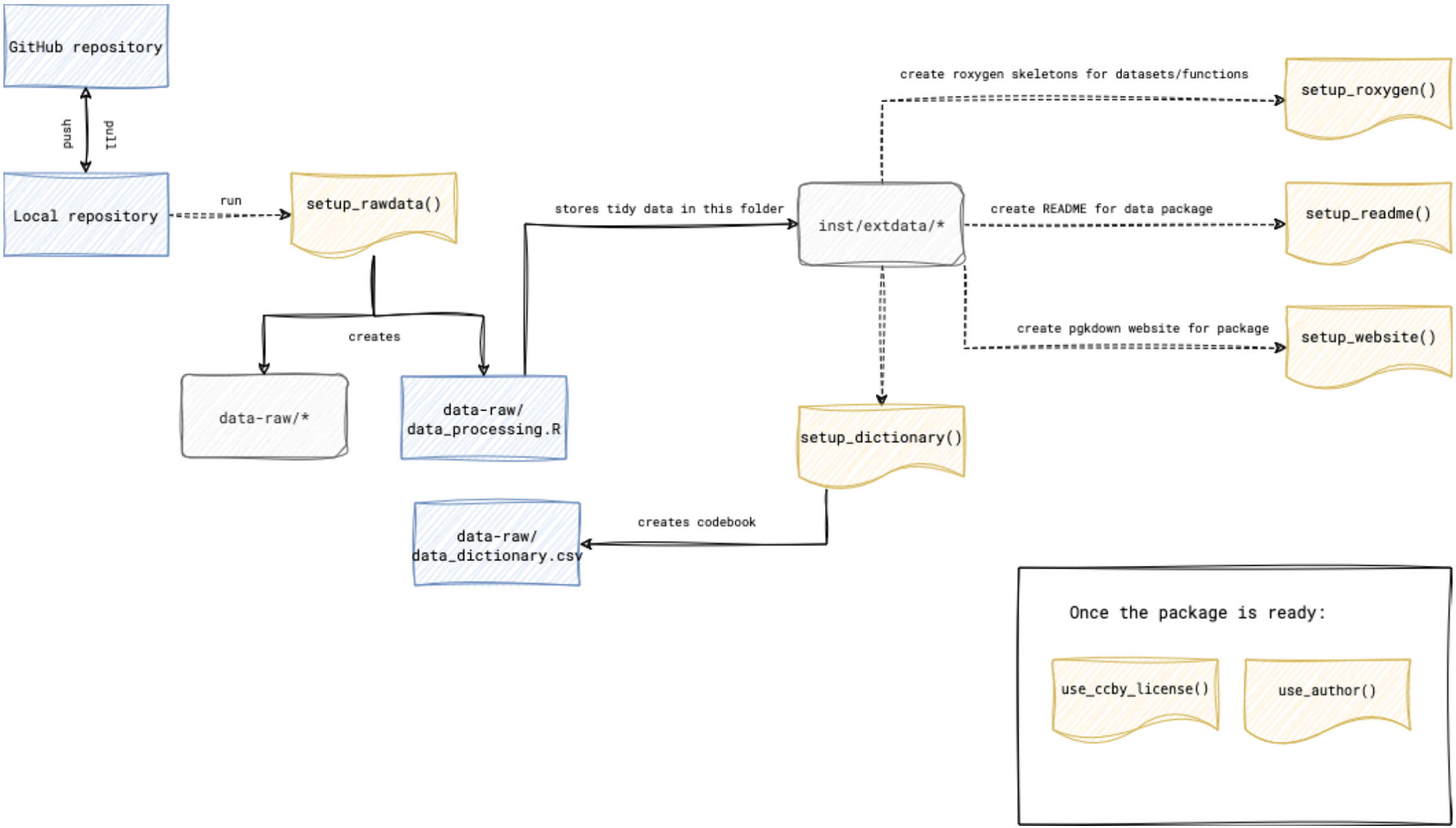
How can we streamline the data publishing procedure?

`washr`

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

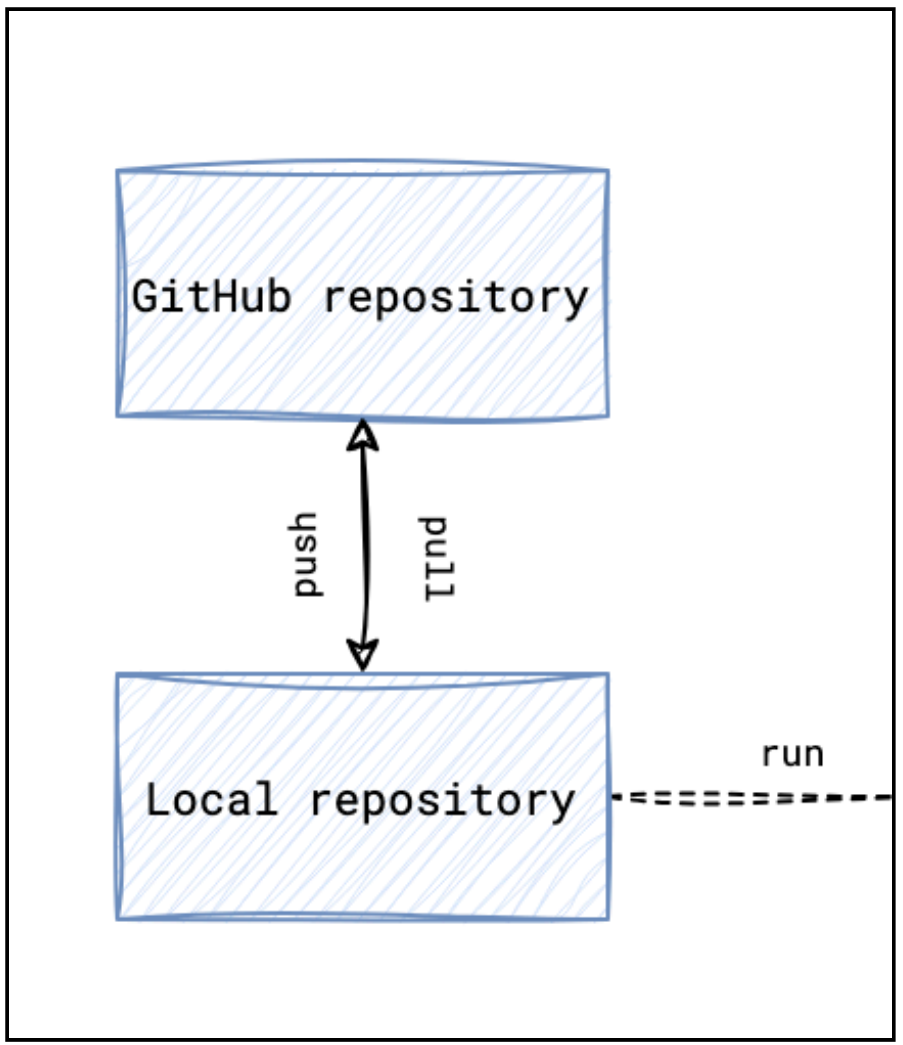
So far:

- 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()

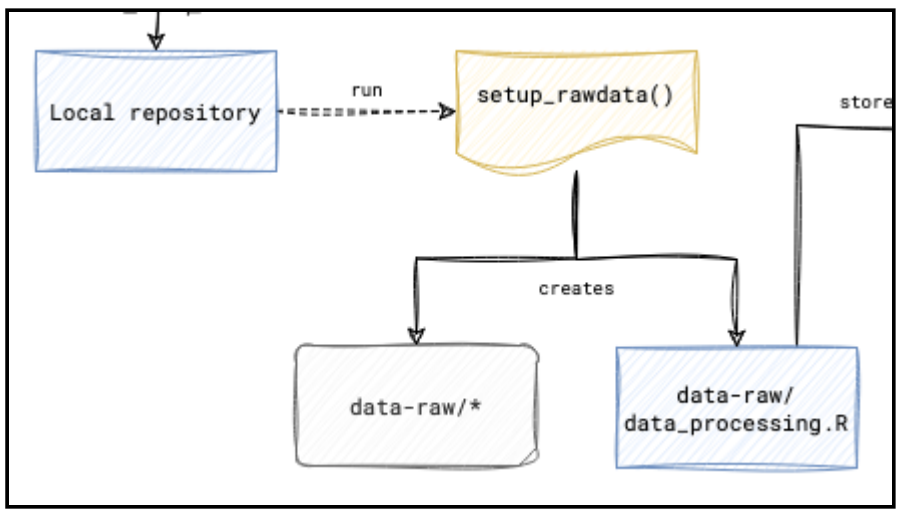
Preparing the data

- Start a local version-controlled folder, connect it to GitHub



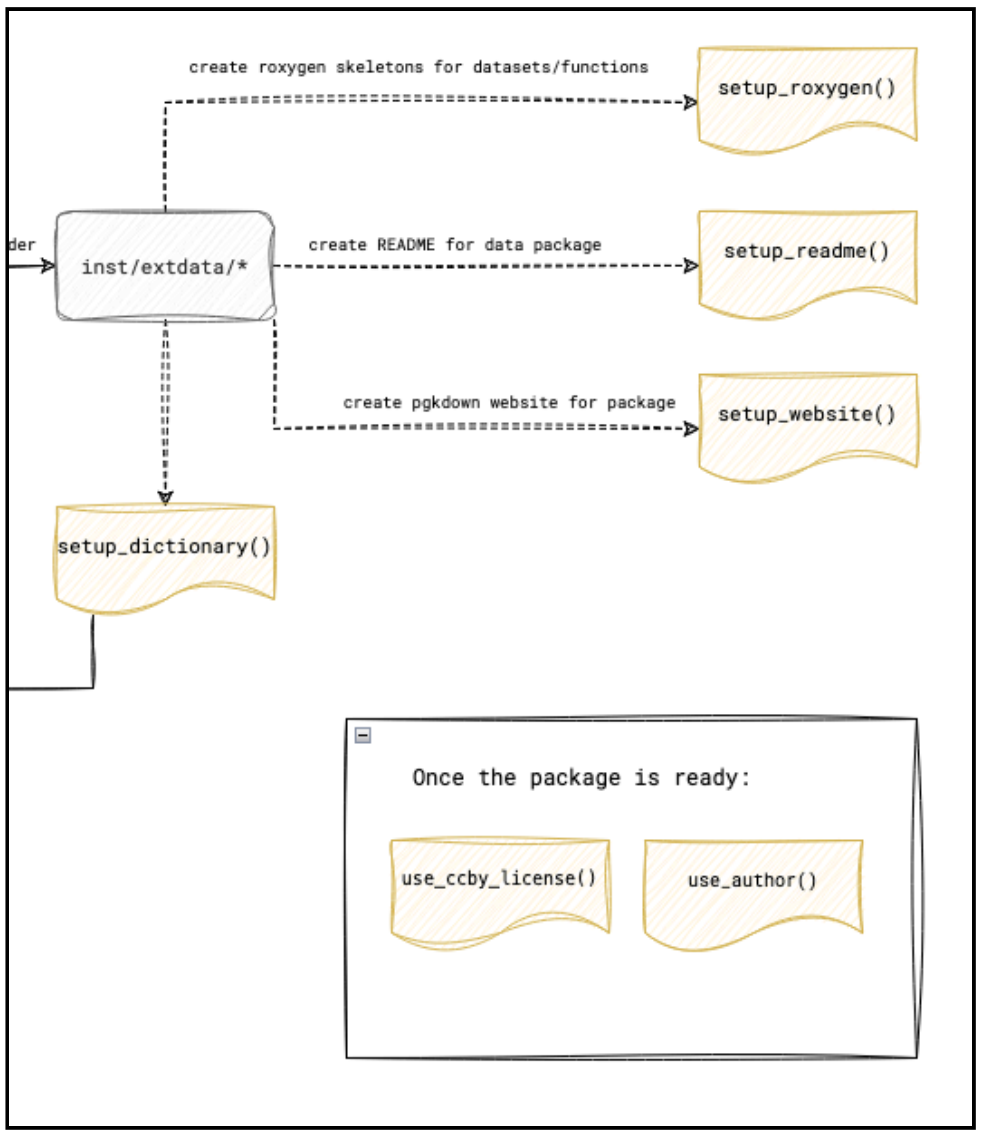
Preparing the data

- `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)



Publishing the data

washinvestments

FAIR principles from a *data perspective*

- **Findable** by having an ORCID
- **Accessible** by utilizing open-source software for both preparing and publishing data
- **Interoperable** by exporting harmonized data as tidy data
- **Reproducible** by ensuring transparency in the data cleaning process (see `data-raw/data-processing.R`)

FAIR principles from a *user perspective*

- **Findable** by publishing tools on accessible platforms (e.g., CRAN, group wiki)
- **Accessible** by thoroughly documenting functions and use cases
- **Interoperable** by maximizing generalizability
- **Reproducible** by consistently updating and revising

What now?

- Continue expanding `washr` R package (more functions, tests)
- Publish `washr` on CRAN to make it more accessible
- Create meta package to easily download data published on openwashdata.org
- Enhance the guide for building and preparing data packages using `washr`
- Expand the package's functionality to be applicable in diverse contexts

Thanks!

Links and Downloads

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

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

Slides: https://ghe-open.ch/slides

openwashdata: https://openwashdata.org

Global Health Engineering: https://ghe.ethz.ch/

Sign up for the openwashdata newsletter!