

Simplifying Research Data Sharing with R

Adriana Clavijo Daza^a, Lars Schöbitz^a, Prof. Dr. Elizabeth Tilley^a, Colin Walder^a, Nicolo Massari^a, Yash Dubey^b, Mian Zhong

Abstract

Through collaboration with WASH researchers in resource-limited countries, we identified significant barriers to open data sharing due to unfamiliarity with FAIR practices and lack of accessible publishing tools. We developed two R packages to address this challenge: `washr` and its enhanced successor `fairenough`, which automate the complete workflow from raw data to publication-ready data packages. These tools democratize open data publishing by minimizing technical requirements while ensuring proper recognition for data contributors in the broader academic community.

Keywords: open data, open science, FAIR principles, data publishing

^aGlobal Health Engineering (GHE) ETH Zurich

^bETH Zurich

Working with WASH (water, sanitation and hygiene) researchers across multiple resource-limited countries, we observed that valuable datasets often remain underutilized due to researchers' limited familiarity with FAIR (Findable, Accessible, Interoperable, Reusable) data practices (Wilkinson et al. 2016). As part of the academic community, we recognize that research extends beyond traditional metrics like citations and publications. The demanding work of generating, collecting, and cleaning data frequently goes unrecognized, leaving many contributors without proper acknowledgment of their contributions.

Our survey data¹ reveals sub-optimal data storage practices among WASH researchers, with many still relying on local storage methods rather than collaborative platforms (Figure 1).

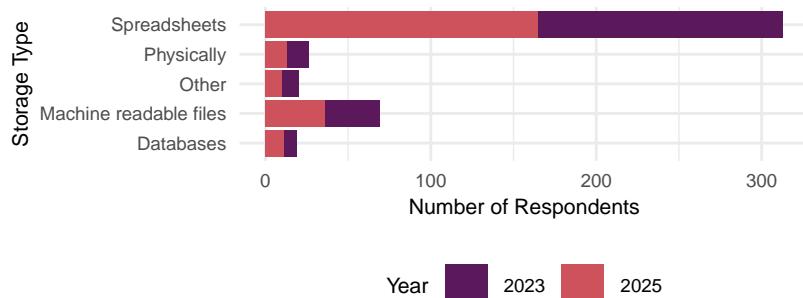


Figure 1: Data storage practices among WASH researchers

The survey data also shows varying levels of programming proficiency (Figure 2), with many researchers having limited experience with R specifically, highlighting the need for user-friendly tools that don't require extensive programming knowledge. A primary barrier is the lack of accessible tools that simplify data publication and distribution using open-source software. This challenge motivated the creation of `washr` (Zhong et al. 2024), an R package that streamlines the process of transforming raw data into publication-ready data packages using `devtools` utilities.

To complement `washr`, we developed a comprehensive data publishing guide (Walder, Schöbitz, and Dubey 2025) as an online book using R and Quarto. This resource provides step-by-step instructions for creating data packages, including automated website generation where datasets are available for

¹The data was collected as part of the registration for the course Data Science for Open WASH Data.

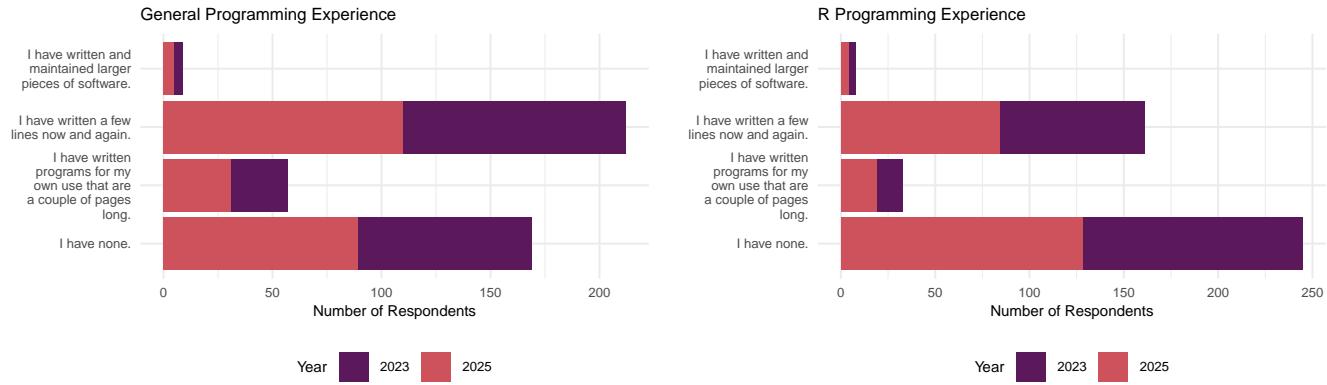


Figure 2: Programming experience among WASH researchers

download in CSV and XLS formats. The guide also covers version control with Git and GitHub, and DOI generation through Zenodo integration.

Following user feedback and recognizing the broader academic community's need for accessible open data tools, we developed `fairenough`²: an enhanced R package designed for more efficient data publishing workflows with minimal user input requirements. `fairenough` provides a complete pipeline for R data package creation with the following features:

- **`fairenough()`'s one-click pipeline:**

Complete R data package creation with a single `fairenough(chat)` call with automated workflow from tidy data to finished package.

- **Granular control options:**

Individual wrapper functions (listed below), global support for overwriting documentation and detailed messages of the process in the console.

1. `setup()` - R project setup: R package structure initialization with `usethis` and, directory and files organization (`data_raw`, `.gitignore`, etc.).
2. `process()` - automated data processing: reads all the raw data with `readr` and `readxl`, validates data structure and formats and with the argument `auto_clean = TRUE` provides automated minimal data cleaning and tidying.
3. `collect()` - interactive metadata collection: guided prompts for package metadata (title, description, authors, etc.) using `cli` and direct saving to `DESCRIPTION` file with `desc`.
4. `generate()` - LLM-powered documentation: data dictionary generation using `ellmer` chat/LLM integration, using the package's description as context and actual data samples. The argument `context` allows for additional user-supplied context.
5. `build()` - complete package infrastructure: Roxygen documentation generation with `roxygen2`, citation file creation with validation using `cffr`, README generation with `rmarkdown` and package website building ready for deployment.

As compared with `washr` this new iteration minimizes the input required from users by reusing all the information provided when possible and suggesting content, for example using LLMs with `ellmer` to automatically generate data dictionaries. We plan to similarly provide a very detailed guide to work with `fairenough` and translate it to Spanish to increase the accessibility of our packages.

The immediate impact of `fairenough` will be demonstrated through its integration into our upcoming Data Science for Open WASH Data course with over 100 participants. Throughout the course, students

²<https://github.com/openwashdata/fairenough>

learn R, Git, GitHub, and Quarto, culminating in capstone projects using their own data. As part of the curriculum, participants will work with **fairenough** to create publication-ready data packages, providing both practical training and real-world application of open data principles. By automating metadata generation, ensuring proper documentation, enabling version control, and facilitating DOI assignment through Zenodo, **fairenough** directly addresses each component of the FAIR principles—making data *Findable* through comprehensive metadata, *Accessible* via the R data package and associated website, *Interoperable* by providing data and metadata in machine-readable formats, and *Reusable* with clear licensing and attribution.

References

- Walder, Colin, Lars Schöbitz, and Yash Dubey. 2025. “ghedatapublishing.” <https://doi.org/10.5281/zenodo.1234>.
- 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): 160018. <https://doi.org/10.1038/sdata.2016.18>.
- Zhong, Mian, Margaux Götschmann, Colin Walder, and Lars Schöbitz. 2024. *Washr: Publication Toolkit for Water, Sanitation and Hygiene (WASH) Data*. <https://doi.org/10.32614/CRAN.package.washr>.