

Data for social good, yes and then? A year at openwashdata as a data scientist

openwashdata.

Experience and reflections from Mian Zhong

Mian Zhong

Global Health Engineering, ETH Zurich

2024-10-16

Note: This blog post is a re-published version from the [Global Health Engineering website](#). For attribution, please see the citation information at the end of this post.

About one year ago, I joined [the Chair of Global Health Engineering \(GHE\)](#) as a new graduate with a Data Science degree. At GHE, the research is focused on the systems and technologies that can help improve all aspects of health in over-exploited countries and populations. Given this research spectrum, you might wonder how I fit in. The answer is, openwashdata. In a nutshell, openwashdata aims to promote [open code and data](#) and empower a community around it in the Water, Sanitation and Hygiene (WASH) sector. Despite WASH being a field impacting 2 billion people's basic daily needs, if you put "data water sanitation and hygiene" in the search engine, you won't find much, let alone a data community. So I said "yes" to the job offer thinking, "This will be my professional start to do data for social good."

Being part of openwashdata is awesome. I work as a coder, designer, teaching assistant, researcher, and sometimes a marketing person. Most of the time, I happily scrape and clean the hidden-gem datasets from the PDF documents, the spreadsheets, and the web pages behind my laptop. Searching for insights and visualizing them in charts, graphs, and maps adds more joy. In addition, I was fortunate to attend [Stockholm World Water Week](#) and co-hosted a panel at [UNC Water and Health Conference](#). Yet, this job is also challenging. The messiest dataset in the data science classroom cannot beat a real-world WASH dataset. I also had to face the reality of standing there as the only data scientist at conferences. Looking back, I want to unfold my year into three parts:

Data processing

Understanding the reality of data generated from a high-impact multidisciplinary sector was my first lesson. Academics, non-government organizations, the private sector, and the government publish data with different meaningful purposes and thus store data in many ways. Valuable datasets may only be available in textual reports. Sometimes I was able to write code to retrieve these datasets into tabular format. But other times, I had to be prepared for some manual data processing. For example, in our ["washopenresearch"](#) project, we wanted to investigate research data accessibility from scientific publications related to WASH. However, I found that the "Data

Availability Statement” section was not formatted consistently on the websites within and across the journals. I wrote code to extract the data availability statement for several journal websites and manually collected the rest together with our research assistant.



Figure 1: Figure 1: An illustrative code snippet about extracting tables embedded in a pdf file.

Community mobilizing

Outreach is as important as developing tasks for a newborn data community. After all, we want the data packages and resources to be reused. My second lesson addresses how to get more professionals to interact with openwashdata and establish new partnerships. Getting out of the coder's comfort zone, I attended Stockholm World Water Week 2023 as one of the twenty Young Professionals accepted. As one of the biggest conferences in the WASH sector, the trip was eye-opening for me to exchange diverse insights and perspectives. In October, I co-hosted a roundtable discussion at the side event “Data: A Key to Unlocking Quality in WASH programmes” at the UNC Water and Health Conference. What strikes me most is that data sharing is more complicated than I had thought. Challenges like over-surveying, ownership, and validation need more discussion and solutions. Both conferences featured data as a key point to improve for WASH, and participants were enthusiastic about learning and bringing changes in open data. This enthusiasm was also shown from our first online academy [“data science for openwashdata”](#). I served as a teaching assistant to facilitate course modules and review projects. Online learning allows a global audience to access hands-on data and programming skills. We received registrations from 80+ countries, and this spring, [20 graduates completed a capstone project](#) with skills learned in the academy. I see huge potential and motivation to improve the data landscape of WASH. Many participants accomplished their capstone projects with limited internet access and hardware. Their dedication and efforts allowed unique WASH data about the Global South spread to everyone.



Figure 2: Figure 2: Mian Zhong (Left) at the side event “Data: A Key to Unlocking Quality in WASH programmes” at the UNC Water and Health Conference 2023.

Tools building

To bridge data science and WASH, updating data tools and making them accessible needs more attention and practice. Throughout the year, our team has tested out [a workflow for developing and publishing data packages](#). My recent work has focused on creating [an R package “washr”](#) to automate this workflow. My vision is to streamline the development of data packages in an efficient and consistent way, so developers can focus more on data cleaning and visualization. To test this package, I organized the first openwashdata hackathon in which each of the six participants used it to deliver a data package within 8 hours. The successful hackathon made a step forward to the vision. Now I am working on submitting the washr package to the Comprehensive R Archive Network (CRAN) to attract and benefit a broader community by publishing data in a low-barrier way.



Figure 3: Figure 3: Participants at the openwashdata hackathon to develop data packages.

My year at GHE is approaching an end. Every time I open my Kanban board and read items on the backlog: a data package about hygiene, a usage tracking dashboard for the published packages, and “text as data” for WASH, I wish I could multi-task more. It has been a journey full of learning, creating and sharing. I hope that we are heading toward where open data is not just a need, it’s a must.