Untitled

Lars Schöbitz

Prof. Dr. Elizabeth Tilley

Table of contents

# 1. Abstract

Decades of manual data structuring have resulted in the most comprehensive and internationally-comparable information on Water, Sanitation, and Hygiene (WASH) coverage. The databases are maintained by the WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) and data are shared in spreadhsheet-based proprietary software. This under utilizes the potential those data could have for other purposes than the national, regional and global monitoring of progress in WASH mandated to the JMP.

We will approach this unused potential by developing open-source data and software packages that follow FAIR data principles to share the data within the WASH community and beyond. In the process, we engage with the community by hosting free learning events for which we use and open-source computational tools, enabling community members to further competencies aligned with FAIR data principles.

Keywords: SDGs, monitoring, open data, open code, R package

# 2. Proposal full title

**Instructions**

Open JMP - unlocking the potential of global indicator data

Acronym: openjmp

# 3. Background and motivation

## 3.1 UNICEF/WHO Joint Monitoring Programme (JMP)

The WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) was established in 1990 to monitor global progress on drinking water, sanitation and hygiene. The JMP produces estimates for a total of 26 indicators related to water, sanitation and hygiene (WASH). Estimates are produced using linear regression models at the country, regional, and the global levels [1].

## 3.2 JMP modelled estimates (data output)

The estimates produced are the only available source of comprehensive and internationally-comparable information on WASH coverage. Data from JMP are also used in reports used to estimate the global burden of disease associated with water and sanitation and to assess the cost-effectiveness, benefit-cost ratio, global expenditure, and investment needs in global drinking water and sanitation [2]. More recently, the data have been used to quantify the service needs for 772 million onsite sanitation facilities [3].

## 3.3 JMP data input

Every two years the JMP updates its global databases to incorporate the latest available national data. [4]. The data is collected from various sources, including censuses, household surveys, administrative data, and others (e.g. research). Data formats between these sources are highly heterogeneous, and a significant amount of effort is invested into extracting data from each source and bringing it into a standardized format.

## 3.4 JMP Country Files

So-called “JMP Country Files” are used for recording these data inputs in a hidden sheet of a Microsoft Excel workbook. A proprietary statistical analysis software package (Stata 14.0) imports the data for all the 232 countries, areas, and territories for which data are available, and runs the estimation model. The resulting estimates are exported back to each country file into another hidden sheet, which in turn feeds a sheet where estimates are produced for the full set of indicators. The result is one JMP Country File for each country. While the combined modeled estimates are accessible through a MS Excel “JMP World File” for all countries, the underlying raw data for the analysis are only accessible through individual JMP Country Files. The underlying Stata scripts that perform the analysis and produce the output data for the MS Excel files are not publicly available.

## 3.5 Use cases

The JMP uses these raw data specifically to produce indicator estimates for SDG indicators 6.1 and 6.2, but there is great potential for unforeseen use cases (e.g. in research, teaching, joining with other data, etc.). These added benefits could be enabled by making the data readily accessible in a form that follows FAIR data principles [5]. There are a range of published studies, which have used the input data, but each would have gone through the effort of extracting and combining all data into a single dataframe for analysis. Further, it is difficult to identify publications that have made use of the JMP input database, because there isn’t a single DOI that could be cited, nor a clear license applied to the raw data. Providing open-source software and data in an accessible and cite-able way can lead to significant uptake within a domain. A great example is the Bioconductor project, which is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics [6]. First published in 2004, the software packages developed for the R and python environment are being used by thousands of researchers worldwide.

## 3.6 Problem Statement

We seek to address this unused potential by contributing software and data packages that address the aims of this call. Using open-source R Statistical Software, we will curate the data carefully and add valuable variables/categories to the existing data that allow for new and unexpected types of analyses. We seek to address the aim of providing a community service by organizing practical workshops and training on the developed software and data packages. These activities will allow for greater engagement with the published data on variety of levels and will provide a highly relevant dataset that also opens the opportunities for people to see “what is behind the curtain”.

In some preliminary work, we developed R scripts that extract the data from the JMP Country Files and combine them into a single dataframe [reference: washr R Package]. We have further started to develop R scripts that replicate the regression models of the JMP using documented JMP methods, but quickly realized that using the JMP Stata script for replicating the analysis would be highly advantageous. A research study is ongoing which applies alternative statistical models to the input data.

# 4. ORD project plan

## 4.1 WP1: Document and publish R data package jmpinput

**Goal:** Make data readily accessible to enable researchers and practitioners to utilize powerful data

In our preliminary work, we have already written a set of functions that extract JMP input data from the JMP Country files. While this is helpful for us, it is not yet easily accessible for others. In this work package, we will prepare detailed and concise documentation that describes the data and the structure that it is published in. We will prepare comprehensive metadata and complete codebooks. Using a collection of packages available in the R universe, we will publish everything as an R data package, including permissive licenses, a DOI, and an attractive public website that allows anyone to learn about the data. While the data is packaged as an R package, we will also ensure that guidance is provided on how it could be used by anyone using a different data analysis software. The approach is innovative because it will lead to a standardized use of the input data. The data processing activities themselves are published openly, so that each step taken to compile the resulting dataframe can be reviewed and validated.

**Activities:**

1. Write up documentation that describes the origin of data and structure that it is provided in
2. Prepare at least two use cases as “vignettes” that can be published together with the data
3. Publish data, documentation and “vignettes” as R Package using workflows described in [Ref R Packages]
4. Use packages of the R environment to publish R Package as a public website, including all documentation and vignettes

**Aims addressed**

This work package focuses on aim 1 and 2 of the call. The research potential increases significantly, as data will be made readily available for anyone to use it and build on. The currently established ORD practice of manually downloading MS Excel files will be simplified. We will further curate the data, so that novel ways of analyzing it will be feasible.

**Research questions**

* RQ1.1: What is the impact of sharing the data as a data package? Indicator: Number of citations per year
* RQ1.2: What use cases beyond using logistic regression for estimates does the WASH community establish?

## 4.2 WP2: Prepare and document R software Package jmpmodel

**Goal:** Develop open-source software to support transparent documentation of modelling activities for primary indicators

The JMP documents methods for the calculation of indicators in detail, however, the actual script which produces the estimates remains closed to users outside of the JMP. This work package envisaged covering the indicators on at least basic, limited, unimproved, and no service (water and sanitation), and basic, limited and no service (hygiene). By publishing the R scripts as open source software anyone can view, use, modify, and distribute the work for any purpose, enforced by a permissive open-source license (e.g MIT License). Further, the results will become reproducible by providing all necessary data (through WP1) and the computer code that runs the regression analysis. This will significantly increase the transparency of the process and in turn, could result in an increase in the trust and reliability of the results. Other added advantages are (1) a complete track record of the complete history of the development process, (2) facilitated collaboration and review process where useful changes and thoughtful contributions can be made to develop our project further; (3) publication of validated research and avoidance of misinformation (4) more efficiency in writing papers, thesis and reports; (5) fair credit for the work; (6) ensured continuity of the work [7].

**Activities:**

1. Work with the JMP team replicate production of the basic indicators for individual countries, areas and territories, drawing on the data inputs described in WP1
2. Prepare a set of R functions that achieve all necessary data manipulation, modelling and visualization steps
3. Write up complete documentation for developed R functions
4. Publish R functions as a complete piece of software as an R Package as described in [Ref R Packages]
5. Use packages of the R environment to publish R Package as a public website, including all documentation and vignettes

**Aims addressed**

As WP1, this work package focuses on aim 1 and 2 of the call.

**Research questions**

* RQ2.1: What are the barriers experienced for publishing the code open-source?
* RQ2.2: What are the practical difficulties of producing a set of R scripts that achieve the same results as the unpublished Stata scripts?

## 4.3 WP3: Dissemination

**Goal:** Seek active collaboration on package development to increase potential usage and long-term maintenance of data and software package.

Open Source projects live by the community that contributes to them. Open Source projects can only be successfully maintained long-term if people invest resources and time into the development. The project team has a history of working with the JMP who have confirmed to actively support this work. We will further use our established networks and communities to disseminate the developed products. The process can tightly fit into our openwashdata project (https://openwashdata.org/), funded under the Open Research Data Program of the ETH Board for Explore Projects and starting March 1st, 2023. The packages developed in this Contribute proposal will provide a fantastic opportunity to prepare novel teaching material that motivates WASH professionals to join the community and at the same time disseminates our products.

**Activities:**

1. Continuously check-in with JMP team to update on progress and share intermediate products for review
2. Host two 3-hour public participatory live coding online workshops for the openwashdata community and beyond to showcase the developed R packages from WP1 and WP2
3. Publish well documented Open Educational Resources from an online workshop for researchers in our network to be used in their own teaching activities (e.g. [NSERC WASH Canada Programme](https://onlineacademiccommunity.uvic.ca/washcanada/), [Mortenson Center in Global Engineering & Resilience](https://www.colorado.edu/center/mortenson/current-students/mortenson-center-courses), [Water, Public Health and Environmental Engineering at University of Leeds](https://eps.leeds.ac.uk/civil-engineering-water-public-health-environmental-engineering))
4. Prepare an article for the Journal of Open Source Software

**Aims addressed**

This work packages addresses aim 3 of the call. The project team will actively engage with the JMP to benefit from the established code, but the at the same time offer opportunities for the JMP team to learn how the developed software could be applied for their own benefit. As the Global Health Engineering group already is a leader in providing training for Open Science and Reproducible Research with R, we would use our expertise to transfer these competencies also to the WASH sector to increase uptake of collaborative software development products.

**Research questions**

* RQ3.1: How many hours of active collaboration (i.e. meetings, trainings, etc.) are required to complete the publication process of R Packages under WP1 and WP2
* RQ3.2: What is the cost-effectiveness (cost/hour) of publishing R Packages under WP1 and WP2 (for benchmarking future initiatives)?
* RQ3.3: Where (location), when (over time), and how (view, downloand, cite) are the R packages used? This research question can not be answered within the time constraints of the project, but will be monitored and evaluated following the close of the grant.

# 5. Impact

Please address these specific points:

* How sustainable is the proposed project inside the ETH Domain?
* To what extent may an existing or a newly formed community (be able to) engage with the ORD practice(s) built-up during the project?

Within the ETH Domain we are able to recruit and access exceptional talent that can contribute to the long-term development and sustainability of the project. This work further complements our activities on the openwashdata project, which established a community of WASH professionals who have a shared interest in data. These professionals will learn how to use the same methods and tools that we apply here and will therefore be able to carry on the activities beyond the lifecycle of the project itself.

# 6. Work Packages and milestones

The following Table 2 shows a basic gantt chart against the four work packages, including program activities and community engagement of the four defined learner personas. Column “Lead” abbreviations: LS = Lars Schöbitz. SA = Scientific Assisstant.

The following Table 3 is a list of research questions associated with each of the Work Packages and related activities in Table 2. Any publications derived from this program will be published as open access material, following ORD practices and Open Science standards for computational reproducibility and sharing of data and code under FAIR principles.

Table 2: https://docs.google.com/spreadsheets/d/1pvt08daECVK\_M-IY3dx1lNUSjcTVy-8miE0GptWAIlc/edit#gid=0

Table 3: https://docs.google.com/spreadsheets/d/1k4eOJcaWGyJDblThgGnxUCvYUnj6qTaqX3Q-0F6WVyQ/edit#gid=0

# 7. Resources (including project costs)

Table 4: https://docs.google.com/spreadsheets/d/1jQE1qrO0T88aXjeARYQaQ3sNsX8UaURJdm7B-yAvDN0/edit#gid=0

# 8. Bibliography

1. WHO/UNICEF Joint Programme for Water Supply, Sanitsation and Hygiene (JMP). JMP Methodology - 2017 Update & SDG Baselines. 2018 Mar.

2. Bartram J, Brocklehurst C, Fisher MB, Luyendijk R, Hossain R, Wardlaw T, Gordon B. [Global Monitoring of Water Supply and Sanitation: History, Methods and Future Challenges](https://doi.org/10.3390/ijerph110808137). International Journal of Environmental Research and Public Health. 2014 Aug;11(8):8137–65.

3. Greene N, Hennessy S, Rogers TW, Tsai J, de los Reyes III FL. [The role of emptying services in provision of safely managed sanitation: A classification and quantification of the needs of LMICs](https://doi.org/10.1016/j.jenvman.2021.112612). Journal of Environmental Management. 2021 Jul;290:112612.

4. WHO/UNICEF Joint Programme for Water Supply, Sanitsation and Hygiene (JMP). Metadata: SDG global indicator 6.2.1a. 2021 Dec.

5. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. [The FAIR Guiding Principles for scientific data management and stewardship](https://doi.org/10.1038/sdata.2016.18). Scientific Data. 2016 Mar;3(1):160018.

6. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. [Bioconductor: Open software development for computational biology and bioinformatics](https://doi.org/10.1186/gb-2004-5-10-r80). Genome Biology. 2004 Sep;5(10):R80.

7. Community TTW, Arnold B, Bowler L, Gibson S, Herterich P, Higman R, Krystalli A, Morley A, O’Reilly M, Whitaker K. [The Turing Way: A Handbook for Reproducible Data Science](https://doi.org/10.5281/zenodo.3233986). Zenodo; 2019.