Untitled

Lars Schöbitz

Prof. Dr. Elizabeth Tilley

Table of contents

# 1 Call Info

## 1.1 Website

* Measure 1: Calls for Field-Specific Actions
* https://ethrat.ch/en/measure-1-calls-for-field-specific-actions/
* Contribute projects will help researchers to contribute to existing open communities built around ORD. They will validate their integration with existing infrastructures (including international ones) and focus on their potential to advance the scientific impact of the field as well as their technical quality.

## 1.2 Email

* Funding for “Contribute projects” aims at:
  + Contributing a **piece of software, data, hardware**, etc., which **simplifies** an established ORD practice, develops its **research potential**, or otherwise improves it.
  + Curating or otherwise improving available ORD, for example, by updating existing data to **adapt it to the evolution of ORD practices** or by linking different ORD datasets or by collecting and curating existing data before uploading it to an ORD platform.
  + Providing **community services**, such as the organization of a **workshop** or training, or **coordination** for the development of an ORD standard.

Funding is available for projects of up to 12 months with an overall budget that must be CHF 30’000.

## 1.3 Application guidelines (ETH Board)

* https://ethrat.ch/wp-content/uploads/2022/10/Application-guidelines\_TrackC-Contribute-projects-3rd-call-1.pdf
* Preference will be given to projects which engage and are connected to a community

## 1.4 Application guidelines (ETH Zurich)

* https://ethrat.ch/wp-content/uploads/2022/06/2022-05-00-ETH-ORD-EPFLETHZ-information-to-applicants-Contribute.pdf
* For applicants at ETHZ
* In this Contribute call, applicants at ETHZ must contribute 25% of the total budget from the resources of their lab/group/unit, either in-cash or in-kind. See the budgeting guidelines below for eligible contributions
* This means an additional 22’500 CHF for us and 7’500 CHF in-kind through my salary

## 1.5 Online submission platform

* https://epflresearchfunds.gomovein.com/locallogin/5d77bf02089cc332e062a3a2/eng
* SNSF guidelines CV
* Asbtract (max 1500 characters)
* 5 keywords

## 1.6 Template

# 2 Project Proposal

Your project proposal must not exceed 6 pages, excluding the bibliography (A4, Arial 11-point font, 1.5 line-spacing, 2-centimeter margins). Please address all points below.

## 2.1 Proposal full title (TODO)

**Instructions**

Please include a project title (max. 12 words); this title will be used for all correspondence. Proposal Acronym (mandatory)

## 2.2 Background and motivation

**Instructions**

Please explain which aim(s) of the ORD “contribute projects” you seek to address:

* Contributing a piece of software, data, hardware, etc., which simplifies an established ORD practice, develops its research potential, or otherwise improves it.
* Curating or otherwise improving available ORD. For example, by updating existing data to adapt it to the evolution of ORD practices. Or by linking different ORD datasets. Or by collecting and curating existing data before uploading it to an ORD platform.
* Providing community services, such as the organization of a workshop or training, or coordination for the development of an ORD standard.

### 2.2.1 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].

### 2.2.2 JMP modelled estimates (data output)

The produced estimates 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 has been used to quantify the service needs for 772 million onsite sanitation facilities [3].

### 2.2.3 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.

### 2.2.4 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 statistical analysis software package (Stata 14.0) imports the 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 in 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 modeled estimates are accessible through a World File for all countries, the underlying data input are only accessible through individual country files.

### 2.2.5 Use cases

The JMP uses this input 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. research, teaching, joining with other data, etc.). These 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 of their own 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 the 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 10s of 1000s of researchers worldwide.

### 2.2.6 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. This allows for engagement with the published data on variety of levels and provides a highly relevant dataset that also opens the opportunities for people to see “what is behind the curtain”.

In some preliminary work, we have started to develop R scripts that extract the data from the JMP Country Files and combine them into a single dataframe [reference: washr R Package]. The database contains xxxxx data points from xxxxx variety of sources. 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.

## 2.3 ORD project plan

* Please describe your project design, including how it addresses the aim(s) of the ORD “contribute projects” described above. Please refer to the scientific literature as well as any preliminary, unpublished results that have informed your approach.
* Which specific gap(s) in knowledge will your work address? Why is your approach innovative?
* What questions will the project answer? What specific outcomes will you focus on?
* Which strategy will you use to address the questions you seek to answer?

### 2.3.1 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 prepare detailed and concise documentation that describes the data and the structure that it is published in. We prepare comprehensive metadata and complete codebooks. Using the universe 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 to provide guidance 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 itself 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 is now readily available for anyone to use it and build on. The currently established ORD practice of manually downloading MS Excel files is simplified. We 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?

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

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

The JMP documents methods for the calculation of indicators in detail, however, the actual script which produces the estimates remains as closed access. By publishing this script as open source software anyone view, use, modify, and distribute the work for any purpose, enforced by a permissive open-source license (e.g MIT License). Further, results become reproducible by providing all necessary data (through WP1) and the computer code that runs the regression analysis. This significantly increases transparency of the process and could result in an increase of “trustability” for the actual 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 to transfer the current Stata script into an R script
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 experienced barriers for publishing the code open-source?
* RQ2.2: What are the practical difficulties of transferring a Stata script into a set of R scripts that achieve the same results?

### 2.3.3 WP3: Collaborate with the JMP to promote use and maintenance of R packages

**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 the contributes to them. They can only be successfully maintained long-term if people invest resources and time into the development. While the project team is well connected with the staff at the JMP, there is currently no defined collaboration on preparing data packages and software. While the project can still be implemented without an agreed collaboration and/or without the Stata script, it would highly benefit from active involvement of the JMP team.

**Activities:**

1. Identify JMP team member who actively contributes to WP1 and WP2
2. Work with the JMP team to transfer the current STATA script into an R script
3. Continuously check-in with JMP team member and general JMP team to update on progress and share intermediate products
4. Host a 4-hour training/workshop with key members of the JMP in Geneva and/or New York to demonstrate the potential use of the developed R Packages for JMP modeling activities

**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. Some team members already use the R Statistical Software, while others may need to be provided with targeted training. 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.

**Research questions**

* RQ3.1:
* RQ3.2:

### 2.3.4 WP4: Dissemination

**Goal:** Widely share the developed package for increased use and impact

We will use our established networks and communities to disseminate the developed products. The process can tightly fit into our openwashdata project, funded under the ETH ORD Scheme for Explore Projects and starting March 1st, 2023. The packages will provide a neat opportunity to prepare novel teaching material that motivates WASH professionals and at the same time disseminates our products.

**Activities:**

1. Host a public participatory live coding online workshop for the openwashdata community to showcase the data packages
2. Publish open source training materials from online workshop for others to use for own teaching activities (e.g. Caetano Canada WASH thingie, University of Leeds, etc.)
3. Prepare an article for the Journal of Open Source Software

**Aims addressed**

This work packages addresses aim 3 of the call.

**Research questions**

* RQ4.1:
* RQ4.2:

## 2.4 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. It is 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 therefore be able to carry on the activities beyond the lifecycle of the project itself.

## 2.5 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

# 3 Resources (including project costs)

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

## 3.1 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.