Universiteit Utrecht

Application form for Fostering Open Science Practice Fund

Closing date: 18 January 2023

The Open Science Fund is an opportunity for **Utrecht University** and **University Medical Centre Utrecht** employees to access small grants with which they can apply Open Science principles into their research. This funding amounts to € 10.000 (minimum) - € 15.000 (maximum) per application.

Contact and information
If you are considering an application and you would like to discuss this with a member of the Open Science Programme team, please send a mail to openscience@uu.nl or contact Judith de Haan, programme director.

More information, such as selection criteria, who can apply and the selection process, can be found on the fund website.

Names	Gerko Vink[1], Odilia Laceulle[1], Sanne Nijhof[1], Heidi Lesscher[1], Anne Hoefnagels[2], Anne Margit Reitsema[2]		
Position/role	[1]Associate Professor, [2]PostDoc		
Department	Methodology and Statistics / Developmental Psychology / UMCU (Pediatrics)		
Faculty	Social and Behavioural Sciences / UMCU		
Email address	g.vink@uu.nl		
Telephone number	0624111054		
Title of proposed project	Making DoY data more easily accessible through data synthesis methods: a case study on Thriving and Healthy Youth data.		
Project start date	Sept 2023	**Project end date**	Aug 2024
WBS number	SA.130402.101		

THIS PROPOSAL HAS AN OPEN DEVELOPMENT REPOSITORY:
https://github.com/gerkovink/OSF2023

Universiteit Utrecht

Please provide a summary of your project (max. 100 words):
(to describe the project on our website)

Data sharing of cross-sectional and longitudinal research faces unique challenges, due to the size and complexity of these data sets, making it a difficult process to automate. Moreover, the potential identifiability of participants is a concern given the myriad ways that information can be combined, even more so for highly sensitive data (e.g., medical). Our project addresses both concerns through developing (1) a script that automates the merging of data from multiple longitudinal studies, and (2) a framework to generate synthetic data sets that are indistinguishable from the original data but which do not disclose any private participant information.

Please outline the proposed project, including the *purpose* of Open Science Practice, the specific *topic* *it addresses*, the *approach* being taken and the *links* to research' (max. 500 words):

Universiteit Utrecht

We propose to develop a framework that makes sharing of (longitudinal) data at Utrecht University and UMC Utrecht more timely and easier. Longitudinal data sets are often large and structurally complex. Merging data from multiple assessments is a time intensive process and automating parts of this process could increase reliability and lower the barrier for open science practices. Medical data is labelled as special personal data, for which stricter conditions apply (AVG), and this data is difficult to anonymize, resulting in restrictions for open access.

In order to develop a widely applicable template for both the UU and the UMC Utrecht , we focus on one research project in particular, the PROactive Cohort Study[1], which can serve as an example for other future projects. This study monitors the psychosocial wellbeing of children with various chronic diseases over the course of their lifespan. This cohort also offers the possibility to link psychosocial well-being to medical data such as diagnosis, disease duration, disease activity. Adding this medical data increases the aforementioned problems with privacy issues.

However, anonymizing data is often not sufficient to ensure confidentiality with longitudinal research. Our second aim is therefore to examine the validity and feasibility of using synthetic data sets. Researchers at Utrecht University demonstrated the possibility of generating synthetic variants of real-life data that could – without loss of generality - be analysed and yield the same conclusions as the original data. This methodology is built on **<mice>**[2], the de facto standard software and methodology in the statistical analysis of incomplete data[3].

With synthetic data, there is no need for distribution of the true data sets, preserving participants' privacy. This will also facilitate easier and faster data sharing. We will study the properties and validity of this methodology on real life data, using data from the PROactive study.

Specifically, we will

1. Build syntaxes using SQL for merging longitudinal data from different sources.
2. Use the PROactive data to develop, validate, and demonstrate a framework for synthetic data analysis.
3. Highlight both methodologies in scientific publications.
4. Disseminate these methodologies in well-documented and open software packages.

We write the first of the SMART objectives as follows, where the remaining objectives are considered in the following box.

- Specific: Two R packages, one with a syntax for longitudinal data merging, and one with synthetic PROactive data sets and functions to synthesize new data sets. Two scientific manuscripts that detail the underlying frameworks and highlight the opportunities and pitfalls of open science practices with longitudinal research.

Universiteit Utrecht

How will you evaluate the progress, outcomes and impact of your project? How will these results be shared? (max. 300 words)

We continue with the remaining SMART objectives:

- Measurable: By the end of this project there will be two open-source software initiatives with corresponding public development repositories, R packages, instructional videos, development instructions and package vignettes. The project will operate under a GNU GPL-3 license, preventing closed source distribution. We can measure the impact of this project by CRAN downloads, GitHub forks and stars, development contributions by other scientists and scientific referencing (long-term)

- Achievable: We work with manageable deliverables (D) that build up to milestones (M):
    - D1.1: Open GitHub repository for development of both projects
    - D1.2: Manuscript that outlines a study into longitudinal data merge approaches and proposes a fguidelines for merging longitudinal data
    - D1.3: R-Package with longitudinal data merge functions PROactive data sets and synthesis functions [M1]
    - D2.1: Manuscript that outlines the data synthesis approaches and proposes a framework for synthetic data analysis of PROactive data.
    - D2.2: R-Package for the synthetic data generation and evaluation of PROactive data [M2]
    - D2.3: Extend the package with synthetic PROactive data sets
    - D2.4: Instructional vignette with corresponding website
    - D2.5: Publish both R-packages to CRAN [M3]

- Relevant: Data sharing and publication is often omitted because the private nature of the data does not allow for open dissemination. Our project may greatly advance data sharing in academia and across and may fortify the position of our research into data synthesis. We also aim to serve as a practical example for transparent and achievable data synthesis efforts in academia. We believe that this would fit well into the UU-wide ambitions on remaining a progressive leader in open science.

- Timebound: Start and end dates are clear (Sept 2023-Aug 2024). Open Science Festival or Open Science Community Utrecht presentation proposal as midway point (March/April 2024).

[1] Nap-van der Vlist, M. M., Hoefnagels, J. W., Dalmeijer, G. W., Moopen, N., van der Ent, C. K., Swart, J. F., ... & Nijhof, S. L. (2022). The PROactive cohort study: rationale, design, and study procedures. *European Journal of Epidemiology*, *37*(9), 993-1002.

[2] Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67.

[3] For reference, in 2022 *mice* has been downloaded 1,063,611 times from CRAN, the largest repository for R.

Universiteit Utrecht

Please describe the potential for learning and/or development for researchers (max. 150 words):

This project has potential for enormous impact and may revolutionize the standard of private data dissemination. The developed software will facilitate applied researchers in transitioning towards open data dissemination, by lowering the threshold for analysing and sharing other's private data. We provide a safe and easy 'steppingstone' for those who are not able to open up their data by providing a proxy of the data that does not carry over any privacy disclosure risk. With that, we provide a learning opportunity for colleagues and collaborators. Moreover, the output generated with this project may result in a wider use of available data sets with less need for costly new data collection efforts. Finally, this open-source project offers learning opportunities for the interdisciplinary research team by centring FAIR software principles from day one. The project may ultimately become a showcase for open data initiatives at Utrecht University and abroad.	

Please detail the amount of funding applied for and justify the costs requested:

The total budget requested is €15.000, which would facilitate

75h - Gerko Vink
75h – Anne Hoefnagels
200h – Student Assistant capacity for 8h/wk for 25 weeks.

Contribution in kind: Odilia Laceulle, Sanne Nijhof, Heidi Lesscher, Anne Margit Reitsema

GV will act as project leader, core developer for the packages and website maintainer. SA, AMR and AH will act as data and analysis leads under supervision of AH, SN and OL. GV, OL and SN will create and streamline documentation and invite other researchers to contribute their analyses. All involved will work together on multiple deliverables. SA will focus mostly on coding and simulation studies under supervision of GV, OL, AH. Regular meetings will be held with the whole team.

Funding would enable us to reach the projects potential: i.e. study the application and utility of synthetic data for large data collection efforts like DoY, reach a larger audience and thereby generate more impact and advance the transparency of analysing and sharing private data by synthetic proxies.

Please send the completed application form to openscience@uu.nl by 20 May 2022.