**Application form for Fostering Open Science Practice Fund**

**Closing date: 20 May 2022**

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](mailto:openscience@uu.nl) or contact [Judith de Haan](https://www.uu.nl/medewerkers/JJdeHaan), programme director.

More information, such as selection criteria, who can apply and the selection process, can be found on the [fund website](https://www.uu.nl/onderzoek/open-science/fostering-open-science-fund).

|  |  |  |  |
| --- | --- | --- | --- |
| **Names** | Gerko Vink, Odilia Laceulle, Sanne Nijhof | | |
| **Position/role** | Associate Professor (3x) | | |
| **Department** | Methodology and Statistics /  Developmental Psychology /  UMCU (Pediatrics) | | |
| **Faculty** | Social and Behavioural Sciences / UMCU | | |
| **Email address** | [g.vink@uu.nl](mailto: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**](https://github.com/gerkovink/OSF2023)**\_THY**

**Please provide a summary of your project (max. 100 words):**

*(to describe the project on our website)*

|  |
| --- |
| Data synthesis methodology is able to generate data sets that are indistinguishable from the original data, but do not disclose any of the private information about the respondents in the data. Yet, widespread adaptation of data synthesis methods in social and medical sciences is lagging. We propose to demonstrate, validate and implement <DoYouSynth>; a synthetic data generation framework aimed at generating and distributing synthetic variants of DoY data. Our project studies the validity of inferences (conclusions) on synthetic sets and will serve as a showcase for implementing synthetic data analysis pipelines in large scale academic private data collection efforts. |

**Please outline the proposed project, including the *purpose* of Open Science Practice, the specific** [***topic***](https://www.uu.nl/en/research/open-science/topics) ***it addresses*, the *approach* being taken and the *links* to research’ (max. 500 words):**

|  |
| --- |
| We propose to explore, study and develop **<DoYouSynth>** a synthetic data framework aimed at making data sharing of highly valuable Dynamics of Youth data more timely and easier for everybody.  At Utrecht University **<mice>[[1]](#footnote-2)**, the de facto standard software and methodology in the statistical analysis of incomplete data[[2]](#footnote-3) is developed. Recently, the <mice> team has demonstrated that <mice> can be used to generate synthetic data sets that 1) exhibit the same univariate (column) properties as the original data, 2) preserve multivariate relations, 3) yield valid inferences (i.e. the same conclusions) as the original data, and 4) results in synthetic data that are indistinguishable from the original data[[3]](#footnote-4). In other words, at Utrecht University we have the means to generate synthetic variants of real-life data sets that could – without loss of generality - be analyzed instead of the original data and yield the same conclusions.  The ability to generate statistically valid synthetic data sets opens a whole new domain of uses and applications. One of the most interesting and pressing opportunities lies in protecting the privacy of respondents: When synthetic data sets can be used for analysis there is no need for distribution of the true private data sets and respondent’s privacy can be preserved by analyzing the synthetic proxy. Another great use of synthetic sets is the ability to facilitate easier and faster data sharing: If there is no risk for privacy disclosure, then data can be openly shared between research groups within and outside the university.  We realize that synthetic data sets would allow for faster data sharing with less restrictions and administrative overhead. We also realize that in order to generate the necessary synthetic data infrastructure, the properties and validity of data synthesis methodology on real life data need to be studied and demonstrated. This is exactly what we propose to do on the Dynamics of Youth data with the proposed DoYouSynth project:   1. Use the DoY data to develop, validate and demonstrate a framework for synthetic data analysis on real-life data. 2. Highlight the methodology in a scientific publication. 3. Disseminate the methodology in the well-documented and open <DoYouSynth> software package, bundled with a set of synthetic data sets.   We write the first of the SMART objectives as follows, where the remaining objectives are considered in the following box.   * Specific: An R package with synthetic DoY data sets and functions to synthesize new data sets based on one-or-more scientific models of interest. A scientific manuscript that details the proper framework of data synthesis for DoY data sets and highlights the opportunities and pitfalls of synthetic data analysis in general.   Meer gedetailleerd, maar teveel text:  Purpose of open science practice with <DoYouSynth>   * The synthetic data sets protect privacy of respondents; * The openly shareable synthetic data sets promote the core values of open science and data dissemination; * The resulting methodology will be **openly documented**, **community-driven** and implemented in the **open-source software R**, thereby allowing developers to formulate, train and explore new data analysis models on synthetic DoY data sets without the need for initial data sharing.  1. Create 3 synthetic versions of DoY data, thereby focusing on the domain *Thriving and Healthy Youth* (THY)*.*     1. A synthetic data set generated specifically for a single THY model    2. A synthetic data set generated for multiple THY analysis models    3. A synthetic data set that is model-agnostic and could apply to a broad set of data models 2. Study the utility, privacy disclosure, and validity of THY analysis models on the 3 synthetic data sets and compare those analysis models to the inference obtained on the original data sets. 3. Expand the model evaluations to new models and techniques not yet considered by THY to study the robustness of analysis models that may not have been the focus of data synthesis on the 3 sets generated in step 1. 4. Create an R package <DoYouSynth> that incorporates the above methodology to synthesize data based on a single analysis model, multiple analyses model or no analysis model, whatsoever. 5. Extend the <DoYouSynth> package with synthetic DoY data sets. |

**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 an open-source software initiative with a corresponding public development repository, R package, instructional videos, development instructions and a package vignette. The project will operate under a GNU GPL-3 license, which prevents 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 repository and website aimed at open development   + D1.2: Study of the utility and validity of synthetic DoY data sets   + D1.3: Manuscript that outlines the data synthesis approaches and proposes a framework for synthetic data analysis of DoY data. [M1]   + D2.1: DoYouSynth package for the synthetic data generation and evaluation of DoY data with mice in R   + D2.2: Extend the DoYouSynth with synthetic DoY data sets   + D2.3: Creating a package instructional vignette with corresponding website   + D2.4: Publish the DoYouSynth package to CRAN [M2] * Relevant: Data sharing and publication is often omitted because the private nature of the data does not allow for open dissemination. Synthetic data sets can overcome such hurdles as the private nature of the data is not disclosed in the synthetic sets. 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 instigate discussion about the transparency of academic research. 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. Open Science Festival or Open Science Community Utrecht presentation proposal as midway point (March/April 2024). |

**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 dissemenation. The <DoYouSynth> software will facilitate applied researchers in transitioning towards open data dissemination, by lowering the threshold for analyzing and sharing other’s private data. We provide a safe and easy ‘stepping stone’ 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 by <DoYouSynth> 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 XX/HOUR  75h – Odilia Laceulle XX/HOUR  75h – Sanne Nijhof XX/HOUR  ?h – Student Assistant capacity for ?h/wk for ? weeks.  GV will act as project leader, core developer for the package and website maintainer. OL and SN will act as data and analysis leads. GV, OL and SN will create and streamline documentation and invite other researchers to contribute their <DoYouSynth> analyses. GV, OL and SN will work together on deliverable 1. SA’s will work on deliverable 2 under supervision of OL, SN and GV. Weekly meetings will be held with GV, OL and SA’s  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 analyzing and sharing private data by synthetic proxies. |

**Please send the completed application form to** [**openscience@uu.nl**](mailto:openscience@uu.nl) **by 20 May 2022.**

1. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67. [↑](#footnote-ref-2)
2. For reference, in 2022 *mice* has been downloaded 1,063,611 times from CRAN, the largest repository for R. [↑](#footnote-ref-3)
3. Volker, T. B., & Vink, G. (2021). Anony mice d shareable data: Using mice to create and analyze multiply imputed synthetic datasets. *Psych*, *3*(4), 703-716. [↑](#footnote-ref-4)