Application form NWO – Open Science Fund – 2020/2021

Section 1 – Details of the main applicant	
Name	dr. ing. Serkan Girgin MSc
Affiliation	University of Twente, Faculty of Geo-information Science and Earth Observation (ITC)
Position	Assistant Professor, Head of the Center of Expertise in Big Geodata Science
End date of contract	Permanent (starting from 1 June 2021)
E-mail address	s.girgin@utwente.nl
ORCID ID	https://orcid.org/0000-0002-0156-185X

Section 2 - Public summary

English public summary

Many researchers use virtual research environments, such as JupyterLab, where substantial data produced during the whole research lifecycle. However, data publishing and sharing typically happen only at the end of the research and shared data often lack important metadata, mainly due to the need of manual inputs. This project aims to develop and operationalize a tool (JupyterFAIR) for 'one-click' and seamless integration of research environments and data repositories, including metadata transfer and data quality checks. The tool will significantly decrease manual intervention needed to archive research data and promote more frequent data sharing in line with FAIR principles.

Word count (max 100): 98

Dutch public summary

Veel onderzoekers gebruiken virtuele onderzoeksomgevingen, zoals JupyterLab, waarin substantiële gegevens worden geproduceerd gedurende de hele onderzoekslevenscyclus. Het publiceren en delen van gegevens gebeurt echter meestal pas aan het einde van het onderzoek waarbij de gedeelde gegevens vaak belangrijke metadata missen, voornamelijk vanwege de vereiste handmatige invoer. Dit project beoogt de ontwikkeling en operationalisering van een 'one-click' tool (JupyterFAIR) voor naadloze integratie van onderzoeksomgevingen en datarepositories, inclusief metadata-overdracht en datakwaliteitscontroles. De tool zal de handmatige tussenkomst die nodig is om onderzoeksgegevens te archiveren aanzienlijk verminderen en het vaker delen van gegevens bevorderen in overeenstemming met de FAIR-principes.

Word count (max 100): 96

Section 3 – Project proposal

3.1 The details of proposal

Proposed project title and acronym	Integration of interactive research environments to data repositories to facilitate FAIR data management practices: JupyterFAIR	
Project duration (in months)	9 months	
The project will mainly address (choose a maximum of two):	☐ Rewards and incentives	☐ Open scholarly communication
	☒ FAIR outputs and standards	☐ Culture change towards Open Science
	☒ Open tools and platforms	
Relevance for a specific discipline:	☒ All disciplines	☐ Specific discipline(s), namely:

3.2 The vision for your project (Criterion: Quality of the project proposal)

Modern research is moving towards the use of virtual research environments, where a wide range of research data are produced during the whole research lifecycle by using research software developed by researchers [1, 2]. In accordance with the FAIR principles, produced data should be ideally published and shared not only at the end, but also at different key milestones during the research, because intermediate data might be as valuable as the final

data in terms of re-usability. In fact, current practices for sharing research software tend to follow this practice by using code repositories that are well integrated to research environments and supporting versioning that allows tracking of the development of research workflows [3]. Whilst most data repositories also offer similar versioning capabilities, they are rarely used because they often require the manual process of uploading data and adding metadata. Since these manual processes are laborious and time-consuming, they hinder the adoption of the FAIR principles among researchers.

This project aims to facilitate more frequent data publishing and sharing practices throughout the entire research lifecycle via a direct and seamless integration of research environments and data repositories [4]. For this purpose, a 'one-click' open-source software tool - JupyterFAIR - will be developed to allow researchers to take a 'snapshot' of their research data including auto-generated metadata and make them FAIR by publishing them in a data repository quickly and easily.

The intended outputs of the project are a) a generic methodology to integrate interactive research environments to research data repositories to facilitate FAIR practices, b) an open-source software tool to implement the proposed methodology in a modular and extendable manner, c) demonstration of the tool by operationalizing it on an existing research environment and a data repository, d) comprehensive user documentation with tutorials to support wide-spread use, e) training workshops to facilitate community engagement, and f) technical documentation to enable additional collaborators for long time sustainability beyond the project lifetime.

The target audience is all researchers who utilize interactive environments for their research and want to apply FAIR research data management principles throughout the entire research lifecycle in an easy and seamless manner.

The proposal will advance Open Science principles by lowering a barrier for the researchers by reducing manual steps necessary to move research data from the production environment to the data repositories. Automatic extraction of metadata from data and research code will not only reduce the necessary user input, but also improve the data quality as metadata otherwise left empty will become available. Easy data sharing directly from the research environment will facilitate more frequent data sharing including intermediate data throughout the research lifecycle. Overall, the proposal will support the cultural change towards Open Science and FAIR outputs.

References:

[1] Fernández, L., Andersson, R., Hagenrud, H., Korhonen, T., Laface, E., & Zupanc, B. (2016). Jupyterhub at the ESS. An interactive python computing environment for scientists and engineers. Proceedings of the 7th Int. Particle Accelerator Conf., IPAC2016. https://doi.org/10.18429/JACOW-IPAC2016-WEPOR049

[2] Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. Nature, 563(7729), 145–146. https://doi.org/10.1038/d41586-018-07196-1

[3] Lamprecht, A. L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., ... & Capella-Gutierrez, S. (2020). Towards FAIR principles for research software. Data Science, 3(1), 37-59. https://doi.org/10.3233/DS-190026

[4] Choi, Y.-D., Goodall, J. L., Sadler, J. M., Castronova, A. M., Bennett, A., Li, Z., Nijssen, B., Wang, S., Clark, M. P., Ames, D. P., Horsburgh, J. S., Yi, H., Bandaragoda, C., Seul, M., Hooper, R., & Tarboton, D. G. (2021). Toward open and reproducible environmental modeling by integrating online data repositories, computational environments, and model Application Programming Interfaces. Environmental Modelling & Software, 135, 104888. https://doi.org/10.1016/j.envsoft.2020.104888

Word count (max 450): 449

3.3 Project plan (Criterion: Feasibility of the project plan)

JupyterFAIR will be developed as an open-source extension of JupyterLab, the most common research environment used across numerous scientific disciplines. As a Minimal Viable Product, which can be further developed in time, it will aim for easy deployment and modular design with respect to integration with data repositories, which will allow researchers to connect to various data repositories. As a proof-of-concept, a full implementation will be developed between JupyterLab and 4TU.ResearchData, an international data repository for technical science disciplines that supports external API access. The tool will be tested and validated on ITC's Geospatial Computing Platform (ITC GCP), a multi-user research environment based on JupyterHub that currently serves more than 300 researchers.

Workplan:

1. *System Analysis* phase will include investigation of the existing Jupyter and research data repository landscapes with a special focus on APIs that enable interoperability. System requirements will be gathered, use cases will be documented, and the scope of the tool will be defined in detail.

2. *System Design* phase will include definition of the generic methodology that will be utilized to integrate research environments and data repositories in a seamless manner, including automated data uploading, extraction of metadata from research code, and data quality checks. The architecture, communication method, backend, frontend (i.e., user interface), and security measures will be designed in detail with prototyping, whenever necessary.

3. *Software Development* phase will include actual writing of the tool by the project team in a collaborative manner. While the generic methodology will be fully implemented as designed in the previous phase, a specific data repository access module will also be developed for 4TU.ResearchData as a proof-of-concept. A source code management system will be used to keep track changes to the code, which will be made publicly available once the code reaches a certain maturity level. Technical documentation will be written in parallel to the code development by following best practices in source code documentation.

4. *System Deployment and Testing* phase will include deployment of the tool on ITC GCP and functional testing of the system components systematically by following test-driven development principles. ITC GCP and 4TU.ResearchData users will be invited to participate in the testing phase, which will provide direct user feedback to correct potential problems and improve available features for a better user experience.

4. *Dissemination* phase will include community outreach activities that are key to the success of the project, which aims to build not only a user community around the tool, but also an active development community to further improve it. Comprehensive user's and developer's guides will be developed for this purpose. Besides step-by-step tutorials, training workshops will also be organized to enable wider use of the tool. To promote further adaptation and dissemination, blog posts will be published at popular sites (e.g., Medium) and the team members will attend and present the tool at conferences.

Timeline:

Project Phase	*M1*	*M2*	*M3*	*M4*	*M5*	*M6*	*M7*	*M8*	*M9*
1. System Analysis									
2. System Design									
3. Software Development									
4. System Deployment and Testing									
5. Dissemination									

Word count (max 500): 499

3.4 Team members (Criterion: Feasibility of the project plan)

dr. Connie Clare (4TU.Research Data, *Community Manager*): dr. Clare is an expert on open science and research data community management. She will take part in a) requirements analysis, b) system testing, c) development of user's guide, tutorials, and training material, and d) community outreach activities. https://www.linkedin.com/in/connie-clare/

Jose Urra Llanusa (TU Delft DCC, *Research Software Engineer*): Mr. Llanusa is an expert on software design, software frontend development, user Interface design, open-source CI/CD, and JavaScript. He will take part in a) requirements analysis, b) system design, c) software development, d) technical documentation, and e) training activities. https://www.linkedin.com/in/josecarlosurra86

Manuel Garcia Alvarez (TU Delft DCC, *Research Software Engineer*): Mr. Garcia Alvarez is an expert on web and back-end development, geoinformatics, computer vision, and smart cities. He will take part in a) requirements analysis, b) system design, c) software development, d) technical documentation, and e) training activities. https://www.linkedin.com/in/garciaalvarez

Word count (max 250): 143

3.5 Budget table

Type of costs	Description	Costs (euros)
Personnel	Senior Scientific Employee (0.2 FTE, 9 months)	**€ 10,916**
	Non-scientific personnel HBO (0.2 FTE, 9 months)	**€ 9,771**
	Non-scientific personnel HBO (0.2 FTE, 9 months)	**€ 9,771**
	Non-scientific personnel HBO (0.2 FTE, 9 months)	**€ 9,771**
Materials	-	**-**
Travel	-	**-**
Total		**€ 40,228**

3.6 Budget clarification (Criterion: Feasibility of the project plan)

The project has three main components, which are a) system design including requirements analysis and design of system elements, e.g., architecture, modules, interfaces, b) software development and testing, and c) technical documentation and training activities. All components are human-labour intensive and will be performed by the project team, the cost of which is indicated as personnel costs. The team members will utilize their existing research and development environments (e.g., laptops); therefore, no materials costs are required. To limit travel the project team will meet virtually by using the infrastructure provided by UT and TU Delft. Training activities will be organized online and utilize the same infrastructure without additional cost. Research environment (ITC GPC) and data repository (4TU.ResearchData) that will be used for the demonstration are provided in kind by the respective organizations, which also agree to support the project by promoting it to their users.

Word count (max 200): 145

Section 4 – Open Science track record of the main applicant

Dr. Girgin established and is currently leading the activities of ITC's Center of Expertise on Big Geodata Science (CRIB), which is a horizontal facility serving 6 scientific departments for better use of cloud computing and big data technologies in education, research, and capacity development activities, including Open Science. He has developed ITC's Geospatial Computing Platform, which is a state-of-the-art interactive computing platform providing GPU-enabled and distributed computing resources through a user-friendly data analysis interface with hundreds of ready-to-use scientific software packages. Together with support services (e.g., database servers, map servers, code repository, and data repository), the platform enables more than 300 researchers to work collaboratively and share outputs in line with FAIR principles. He is one of the main collaborators of open-source Clustering Geo-Data Cubes (CGC) software package developed by ITC and the Netherlands eScience Center to perform multi-dimensional geospatial data clustering. He is also a member of the Open Science Community Twente. In the past, he designed and developed European Commission's open, collaborative database on Natech accidents (eNatech), and open, collaborative Natech risk assessment system (RAPID-N) featuring innovative "open model" concept allowing public access to all data, equations, and their interactions in a transparent manner.

Word count (max 200): 196

Section 5 – Data management

5.1 Will this project involve re-using existing research data?

☐ Yes: Are there any constraints on its re-use?
☒ No: Have you considered re-using existing data but discarded the possibility? Why?

The project does not require existing research data for the development of the methodology and the tool. But existing public research data available on 4TU.ResearchData will be studied to understand prevalent data sharing practices and to identify lacking information (e.g., important metadata) that might be provided automatically.

5.2 Will data be collected or generated that are suitable for reuse?

☒ Yes: Please answer questions 5.3 and 5.4
☐ No: Please elaborate.

User and technical documentation, training material, and tutorials will be generated.

5.3 After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties? Are there possible restrictions to data sharing or embargo reasons? Please state these here.

User and technical documentation, training material, and tutorials will be stored at 4TU.ResearchData, which is a public data repository. Source code of the documentation as well as interactive training notebooks will be made available also at a public code repository (e.g., Github). They will be published under a Creative Commons Attribution license (CC-BY), which allows redistribution and adaptation including commercial purposes by giving appropriate credit to the related parties.

5.4 Will any costs (financial and time) related to data management and sharing/preservation be incurred?

☐ Yes: Then please be sure to specify the associated expenses in the budget table of this proposal.

☒ No: All the necessary resources (financial and time) to store and prepare data for sharing/preservation are or will be available at no extra cost.

Section 6 – Software sustainability

6.1 Will software be generated during the project?

☒ Yes: Please answer questions 6.2, 6.3, 6.4 and 6.5

☐ No

6.2 How will the software be licensed and be made available for re-use?

The software will be licensed under a 3-clause BSD license unless libraries with less restrictive licenses will be required for the development. In that case, the least restrictive license will be used. Source code will be made available to third parties for re-use and further development through a public code repository (e.g., Github).

6.3 What measures are needed to make the software appropriate for long-term (re-)use by third parties?

Sphinx-based online documentation will be provided to facilitate easy, long-term use and reuse. Interactive notebooks will illustrate the use of the software. A Slack channel will be established to support a community of practice. The use of the tools will be promoted starting with 4TU.ResearchData serving 4 different Dutch universities and ITC GCP serving hundreds of researchers at UT.

6.4 How large do you expect the community that will potentially use the software to be, and do you expect outside contributors to the software?

The software is not domain specific and will be useful for all researchers using interactive research environments and data repositories. Therefore, a large community is expected from all fields. We also expect contributions from universities, research organisations, and independent developers to further improve the software, especially to support more data repository platforms, e.g., Dataverse, DansEASY.

6.5 What expertise do you expect to be needed to make the software appropriate for long-term re-use by third parties? Is this expertise available?

Proper design and development following the best practices in open-source software development are key to make the software appropriate for long-term re-use. As the developer and operator of research environments (ITC) and members of DCCs (TU Delft) we not only conduct research, but also collaborate closely with other researchers to support their activities. Therefore, we are very aware of the needs, which will help us for a better system analysis and design. Being long-time software developers experienced in web-based applications, we are also in a strong position with our technical know-how to develop the tool. We see the establishment of an active community and providing effective user support as fundamental, which will be supported by a dedicated team member experienced in community management. The DCCs and Open Science Communities of UT and TU Delft will also promote the use of the tool within their research communities, and whenever possible training on how to use the tool can be organized for scientific staff. The DCC at TU Delft counts with a group of software engineers that can support the maintenance of the project. Finally, the staff at 4TU.ResearchData will promote the tool and its further development among the Technical Universities in the Netherlands.

Section 7 – Other grant applications with overlapping content

There are no other proposals of the main applicant or the team members with overlapping content neither to NWO nor another funder.

Signature

☒ By submitting this application form, I declare that I and all the individuals involved in this proposal satisfy the nationally and internationally accepted standards for scientific conduct as stated in the Netherlands Code of Conduct for Research Integrity (The Association of Universities in the Netherlands).

☒ By submitting this application form, I declare that the team members named in this form have read and agreed with the submission of this proposal and have agreed with their role and intended contribution to the project, should this be awarded.

☒ I have completed this application form truthfully.

Name main applicant: Serkan Girgin
Place: Enschede
Date: 17/05/2021