

Better Interoperability Between Research Data Repositories and Digital Research Environments

fairly and JupyterFAIR



Dr. Serkan Girgin

s.girgin@utwente.nl

Manuel Garcia Alvarez

m.g.garciaalvarez@tudelft.nl

José Carlos Urra Llanusa

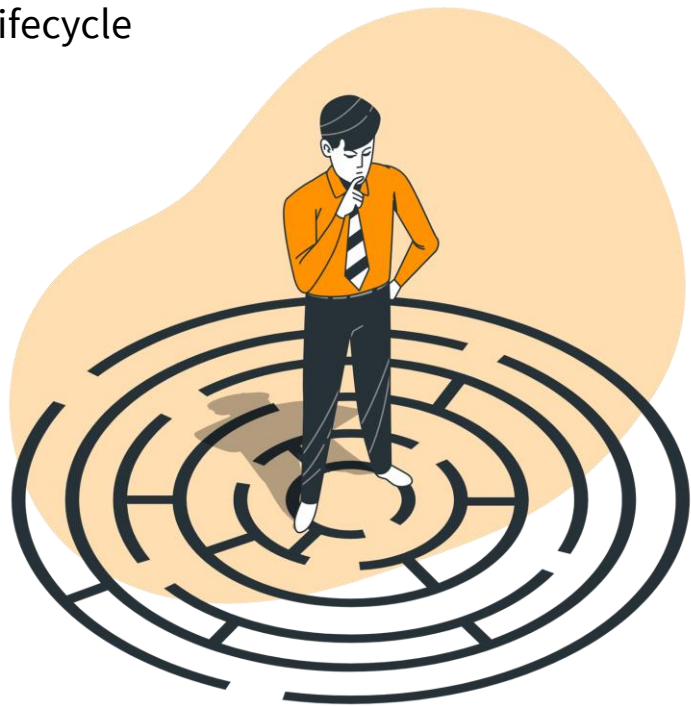
j.c.urrallanusa@tudelft.nl

Dr. Kees den Heijer

c.denheijer@tudelft.nl

The problems

- Research data are produced during the whole research lifecycle
- Data publication and sharing happens mostly at the end
- Lacks important metadata
- **Manual data publication takes time and effort!**
- Virtual Research Environments (VREs) facilitate research and research data production
- They are well connected to some infrastructure, e.g. code repositories
- **VREs are not well connected to data repositories!**



The NWO Open Science Fund project

Development of a [JupyterLab](#) extension that allows taking snapshot of research data and publishing it in a [Data Repository](#)

- A methodology to integrate VREs to research data repositories
- [JupyterFAIR](#): A modular and expandable open-source software that implements the methodology
- Demonstration at [4TU.ResearchData](#) and [ITC Geospatial Computing Platform](#)
- Technical and practical documentation and end-user training

For more information:

[Girgin et al., 2022, Integration of interactive research environments to data repositories to facilitate FAIR data management practices: JupyterFAIR](#)

DOI [10.5281/zenodo.6026285](https://doi.org/10.5281/zenodo.6026285)

The project is funded by the [NWO Open Science Fund](#), File No. 203.001.114

Our team

Kees den Heijer
Senior Researcher*
4TU.ResearchData

Researchers

User community Geospatial
Computing Platform (GCP)
& 4TU.ResearchData

Serkan Girgin
Senior Researcher
U/Twente ITC-CRIB

José Carlos Urrea Llanusa
Research Software Engineer
TU Delft DCC

Data Supporters
4TU.ResearchData
technical team, data
stewards & other orgs!

Manuel Garcia Alvarez
Research Software Engineer
TU Delft DCC



Design Features

- 3-Tier Approach
 - **Python package:** [fairly](#)
 - Provides an API to manage and publish research data
 - Allows further development by other interested parties
 - **Command line interface:** [fairly CLI](#)
 - Research data management on the command line
 - **JupyterLab extension:** [JupyterFAIR](#)
 - Research data management by using a graphical interface
- Support for multiple research data repository platforms
 - [Figshare](#)
 - [Zenodo](#)
 - and more... (object-oriented extensible design)



Product Features

- Quick research dataset cloning
 - **One-command retrieval of metadata and all data files** by using URL address, DOI, or record identifier
 - **Automatic extraction of archived data files** (e.g. .zip, .tar.gz)
- Local metadata management
 - Creation and **editing of metadata locally** by using **your favorite text editor** or API methods
- Quick dataset publication
 - **One-command** creation of **research data records** on online data repositories **in a unified way**
- Unattended large dataset uploading
 - Easy uploading of **high number of data files and folders**, including **large files**
 - **Automatic creation of archive files** (e.g. .zip, .tar.gz) if folders are not supported by the data repository
- Smart dataset synchronization
 - **Automatic identification of added, removed, or updated data files** and **uploading only if necessary**
 - Easy **versioning of datasets in a unified way** considering the repository rules

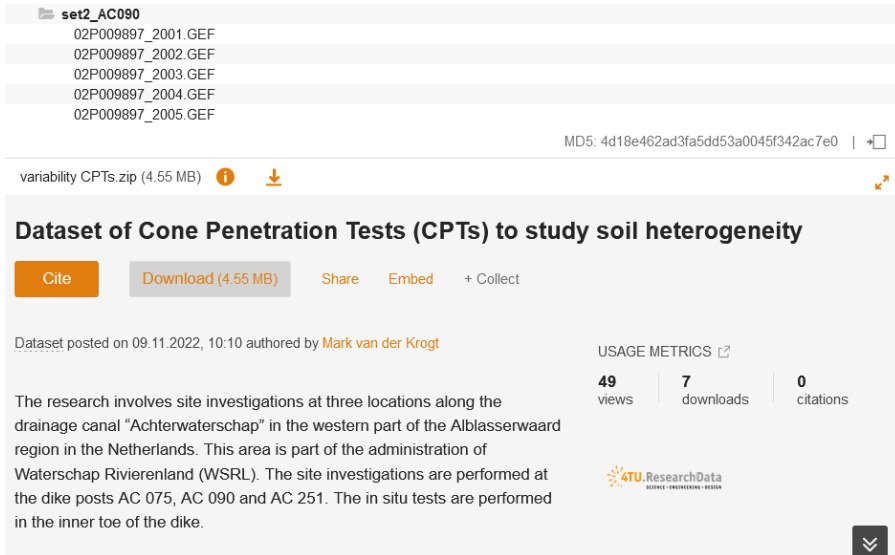
fairly package

```
1 import fairly
2
3 # Create a local dataset
4 dataset = fairly.create_dataset('/path/dataset')
5
6 # Set metadata
7 dataset.set_metadata({
8     "title": "My wonderful dataset",
9     "license": "CC BY 4.0",
10    "keywords": ["FAIR", "data"],
11    "authors": [
12        "0000-0002-0156-185X",
13        {
14            "name": "John",
15            "surname": "Doe",
16            "role": "contributor",
17        },
18    ],
19 })
```

```
20
21 # Add data files and folders
22 dataset.add_files([
23     "README.txt",
24     "*.csv",
25     "train/*.jpg",
26     "test/*.jpg"
27 ])
28
29 # Upload to the remote data repository
30 remote_dataset = dataset.upload("4tu")
31
32 # Change metadata
33 dataset.metadata["license"] = "MIT"
34
35 # Synchronize the remote dataset with the local dataset
36 dataset.synchronize()
```

Demonstration

```
import fairly
4tu_dataset = fairly.dataset("https://data.4tu.nl/Dataset_of_Cone_Penetration_Tests/21510030")
local_dataset = 4tu_dataset.store("/data/cone-penetration-tests", extract=True)
zenodo_dataset = local_dataset.upload("zenodo")
```



The screenshot shows a file explorer view of a dataset named 'set2_AC090' containing five files: '02P009897_2001.GEF', '02P009897_2002.GEF', '02P009897_2003.GEF', '02P009897_2004.GEF', and '02P009897_2005.GEF'. Below this is a download button for 'variability CPTs.zip (4.55 MB)'. The main content area features the title 'Dataset of Cone Penetration Tests (CPTs) to study soil heterogeneity' with buttons for 'Cite', 'Download (4.55 MB)', 'Share', 'Embed', and '+ Collect'. The dataset was posted on 09.11.2022, 10:10 by Mark van der Krogt. Usage metrics show 49 views, 7 downloads, and 0 citations. A description of the research is provided, along with the 4TU ResearchData logo.

MD5: 4d18e462ad3fa5dd53a0045f342ac7e0

variability CPTs.zip (4.55 MB)

Dataset of Cone Penetration Tests (CPTs) to study soil heterogeneity

Cite Download (4.55 MB) Share Embed + Collect

Dataset posted on 09.11.2022, 10:10 authored by [Mark van der Krogt](#)

USAGE METRICS

49 views	7 downloads	0 citations
----------	-------------	-------------

The research involves site investigations at three locations along the drainage canal "Achterwaterschap" in the western part of the Alblasserwaard region in the Netherlands. This area is part of the administration of Waterschap Rivierenland (WSRL). The site investigations are performed at the dike posts AC 075, AC 090 and AC 251. The in situ tests are performed in the inner toe of the dike.

4TU ResearchData

Development

Open-source Code Repository



<https://github.com/ITC-CRIB/fairly>

<https://github.com/ITC-CRIB/JupyterFAIR>

User Documentation



<https://jupyterfair.readthedocs.io/en/latest/>

fairly Package



<https://pypi.org/project/fairly/>

Follow us on Twitter for updates!



<https://twitter.com/JupyterFAIR>

