SURF Technical Consultation Meeting

25 September 2023, Utrecht

Next Generation Research Data Repositories

Bringing Computation to Data for Exploratory Data Analysis and Visualisation

UNIVERSITY OF TWENTE.

ITC

TU Delft

Research data repositories allow open access to research data, but they **have limitations reducing FAIRness**

- **Metadata on individual data files do not exist** and only a basic preview of their content can be displayed* that **reduces Findability**.
- **Folders are not supported**, resulting in researchers to create archive files to preserve their dataset structure that **reduces Accessibility**.
- **Random access to data files is not supported** that especially for cloud-native data **reduces Interoperability**.
- Datasets **lack material to support their further use**, such as example notebooks, that **reduces Reusability**.

In practice, the only way to access data is to **download it – even it is big**.

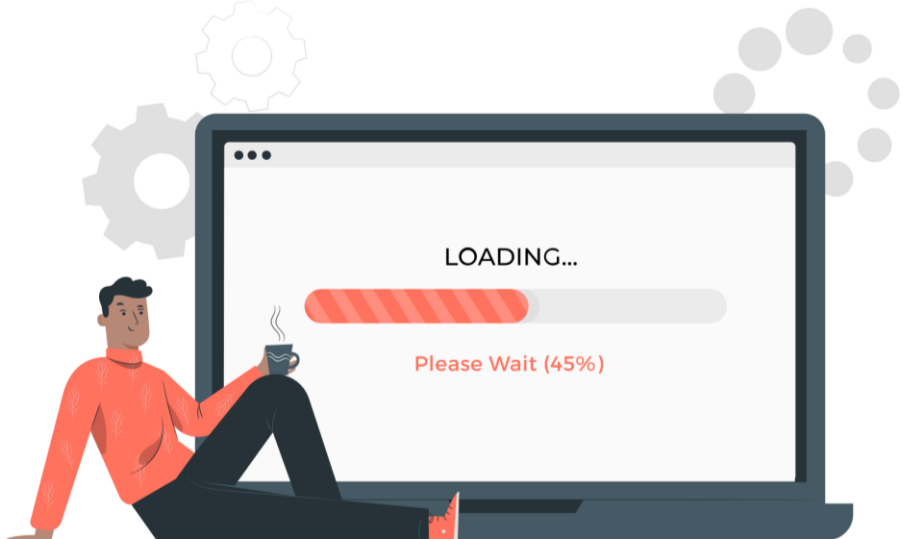


Illustration by Storyset.com

The researchers also
need to find a way to explore data.

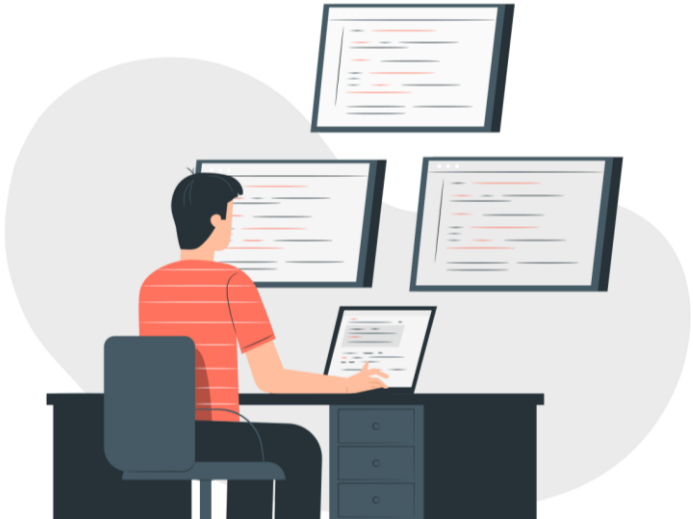


Illustration by Storyset.com

We aim to **facilitate** interactive access to open research data

- Resolve the need for a **separate environment to explore** research data
- Reduce the **time to explore** research data
- Reduce the **amount of data** that is transferred but not effectively utilized

Illustration by Storyset.com

We are developing a platform to provide **analysis-ready exploratory research environment** with data

- Development of an **open-source software** to create and manage interactive computing environments with analysis-ready data.
- Development of **template interactive notebooks** to facilitate rapid exploratory data analysis.
- Operationalizing of a **prototype platform**.
- Development of the user **documentation and training material**.
- Organizing a **training workshop**.
- **Feasibility and benchmarking study** to use the **SURF** infrastructure.

Illustration by Storyset.com

Our **project team** brings together different skills and experience

Dr. Serkan Girgin [a]

Dr. Pablo Castellanos Nash [b]

Néstor De la Paz Ruiz [c]

Alexandru Matcov [a]

Core Team

Dr. Wendy van Ginkel [c]

Madeleine de Smaele [b]

Dr. Alice Nikuze [a, c]

Roel Janssen [b]

In-kind Contribution

[a] University of Twente, Faculty of Geo-information Science and Earth Observation
[b] TU Delft, Digital Competence Center / 4TU.ResearchData
[c] University of Twente, Digital Competence Center *

We utilize and **extend well-known tools** and technologies

- The platform is based on **fairly**, **JupyterHub, JupyterLab, Docker, and FastAPI**.
- Datasets are **directly accessible** without download and extraction by the user.
  Our target is zero waiting time for popular and new research datasets through active monitoring and smart caching.
- An interactive IDE and core exploratory **data analysis tools and packages** will be available for various languages (e.g., Python, R).
- Template **interactive notebooks tailored** to the dataset will be available in various languages (e.g., Python, R).
- We do not aim to provide a full-fledged computing environment.
  In the future, it might be possible to support compute-intensive tasks.
- The service will be **easily integrated** to data repositories.
  Inclusion of a simple JavaScript snippet to the data page will be sufficient.

Illustration by Storyset.com

fairly provides an API to clone and manage research datasets

- Open-source software to facilitate **research data management** and publishing through a **3-tier approach**: API, CLI, JupyterLab extension.
- Quick **retrieval of metadata and data** files by using URL address, DOI, or record identifier.
- Automatic **extraction of archived data** files (e.g. .zip, .tar.gz).
- **Unattended downloading** of a high number of files, including large ones.
- **Smart synchronization** by automatic identification of added, removed, or modified files.

Funded by the NWO Open Science Fund 2021, File No. 203.001.114

More information is available at https://github.com/ITC-CRIB/fairly

```
import fairly

# Open a remote dataset
dataset = fairly.dataset("

# Get dataset information
dataset.id
{'id': '21588096', 'versio

dataset.url
'https://data.4tu.nl/artic

dataset.size
33339

len(dataset.files)
6

dataset.metadata
Metadata({'keywords': ['Ea
'online_date': '2022-11-24

# Update metadata
dataset.metadata["keywords
dataset.save()
```

Smart caching allows efficient storage and access to datasets

- Published research datasets are **static**, hence suitable for caching.
- Smart caching by **proactively monitoring** popular and new datasets.
- Caching **on-demand** by user request.
- Smart cache management by **using operational data**.
- Effective caching by **maximizing storage** utilization (~90%).
- Effective caching with **compression at file system** level (ZFS LZ4).
- Custom server application to **cache and manage datasets**.

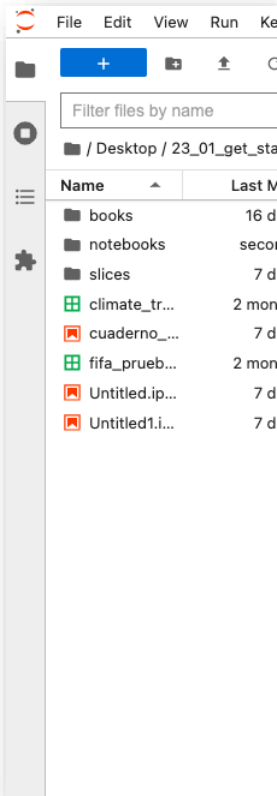
Illustration by Storyset.com

JupyterHub manages interactive analysis environments

- **Lightweight integration** with JupyterHub with minimum code change.
- Cookie-based **custom authenticator** for users.
- Access to **multiple datasets** concurrently by the same user through **named user servers**.
- **Docker-based** operation to provide containerized user servers.
- Datasets are made available by **read-only volume binding** concurrently.
- Template notebooks are made available by **read-write volume binding**.
- **FastAPI-based** custom server application to manage the named servers.

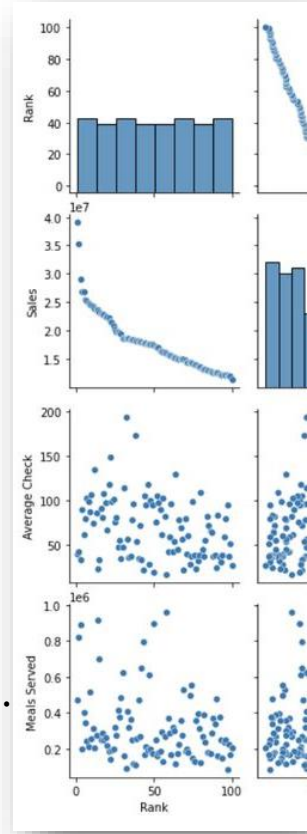
Interactive analysis environments are based on **JupyterLab**

- Customized JupyterLab environment based on **Jupyter Docker stack**.
- Additional **useful packages and extensions** are provided, including AI support.
- Becomes **available immediately** even if the dataset is not readily available.
- Data folder is read-only only in the container, i.e. data can be copied externally.
- The analysis environment is available **for a limited duration**, e.g. one hour*.
- The user can **extend the duration** if necessary*.
- **Custom extension** to display progress indicator for dataset availability, the count-down timer and to extend duration.
- Custom extension will also allow **rating** of the dataset and provided service*.



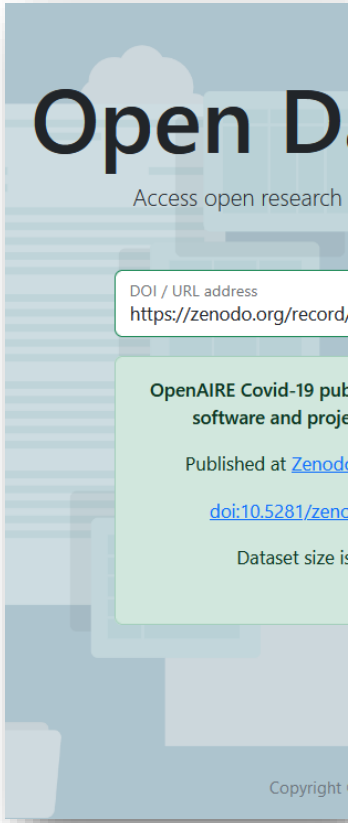
Template notebooks facilitate exploratory data analysis

- Template notebooks will be available for **different file types**.
- Based on the file types of the dataset, only **relevant templates** will be provided.
- They will be **pre-processed** to make them tailor-made for the dataset.
- **Dedicated notebooks** for specific datasets will also be supported*.
- **AI-based support** will be available for the users for creating small scripts.
- Template notebooks will have a **public code repository**.
- We expect that the **community** can be a template notebook provider in time.
- It might be possible to let users **submit notebooks** they created on the platform*.
- It might be possible to pre-calculate and **cache data analysis reports***.



The **platform portal** will provide a single point of entry

- It will allow access to datasets by **DOI and URL address**.
- It will **validate** requests and provide information about **dataset availability**, including estimated data preparation time.
- It will **trigger** dataset cloning once a dataset is searched – no waiting for access request.
- It will support common **public research data repositories**.
- Research data repositories will be able to provide **customized-links** to the platform by using a tiny JavaScript snippet.
- It will provide data cache and platform **usage statistics**.
- It will **highlight** most popular and newly cached datasets.
- It may highlight **community developed notebooks** in the future*.



We are **progressing** in a collaborative manner

- The design of the **system architecture** is finished.
- **Development** of the system components are on-going.
- **Literature review** on notebooks for exploratory data analysis is finished.
- Development of **template notebooks** are on-going.
  We are also considering to provide pre-calculated exploratory data analyses.
- The **infrastructure** for the operational prototype is ready.
  200 TB storage, 1.5 TB memory, 176 vCPU.
- Next **milestone** in October:
  A prototype platform available at **https://opendataexplorer.org**

Illustration by Storyset.com

Open Data Explorer

Access open research data interactively without downloading!

DOI / URL address
https://zenodo.org/record/8221703

OpenAIRE Covid-19 publications, datasets, software and projects metadata.

Published at Zenodo on 16/7/2023

doi:10.5281/zenodo.8221703

Dataset size is 1.38 GB

Available!

You can access it immediately.

Requested **2** times.

Access Now

Let's demonstrate!

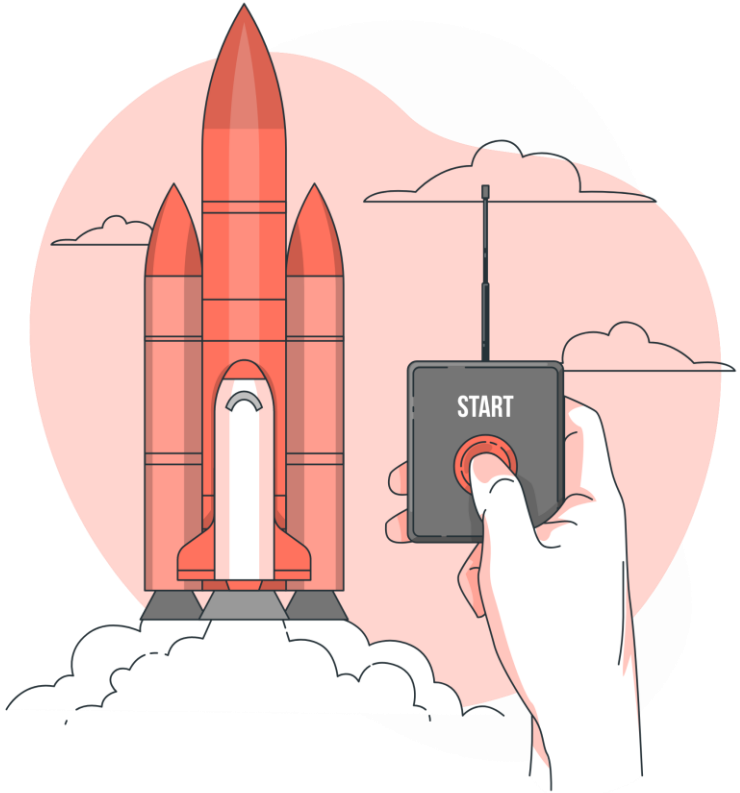


Illustration by Storyset.com

Collaboration with SURF is important for **sustainability**

- For short-term we have the required infrastructure, but for **long-term** the use of SURF infrastructure could be a better option for **sustainability and improved user experience**.
- Moreover, SURF infrastructure can allow the researchers to go beyond exploratory data analysis and do **more intensive co-located computing** with data.
- It is vital to **promote** the platform so that we can **reach the researchers**. SURF can facilitate this through its communication channels (e.g., web portal, social media)

Illustration by Storyset.com

Let's discuss together!

Illustration by Storyset.com

How can we **collaborate to enable interactive access** to research data?

- How can we make use of the SURF infrastructure to **cache research datasets**?
- How can we make use of the SURF infrastructure to **provide lightweight research environments** for interactive exploratory data analysis?
- How can we make use of the SURF infrastructure to **provide powerful interactive research environments** with analysis ready data?
- …?

Illustration by Storyset.com