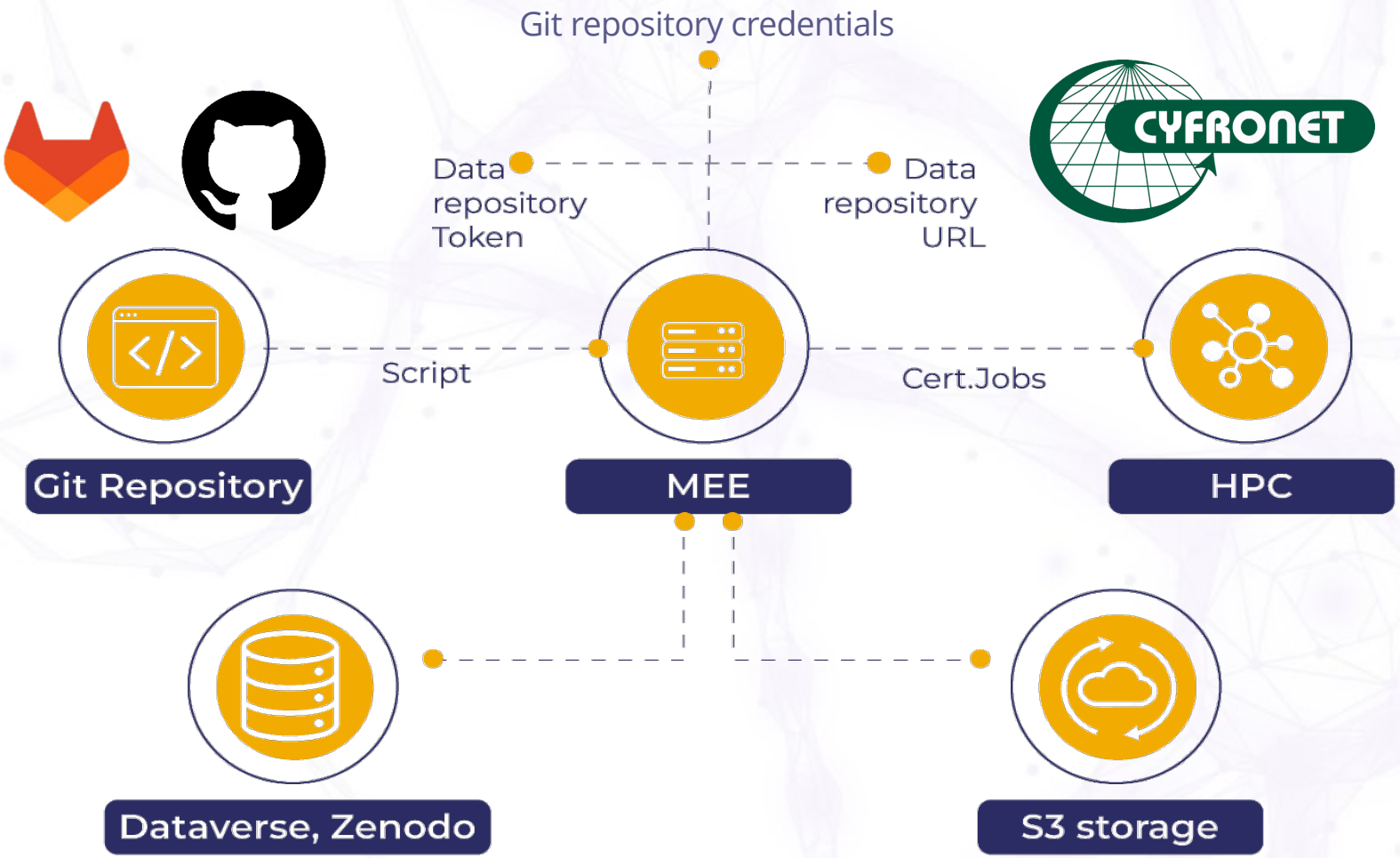
Integrating data repositories with HPC resources for execution of VHT models

Taras Zhyhulin, Karol Zając, Maciej Malawski, Jan Meizner, Piotr Nowakowski

Technical Problems in Research

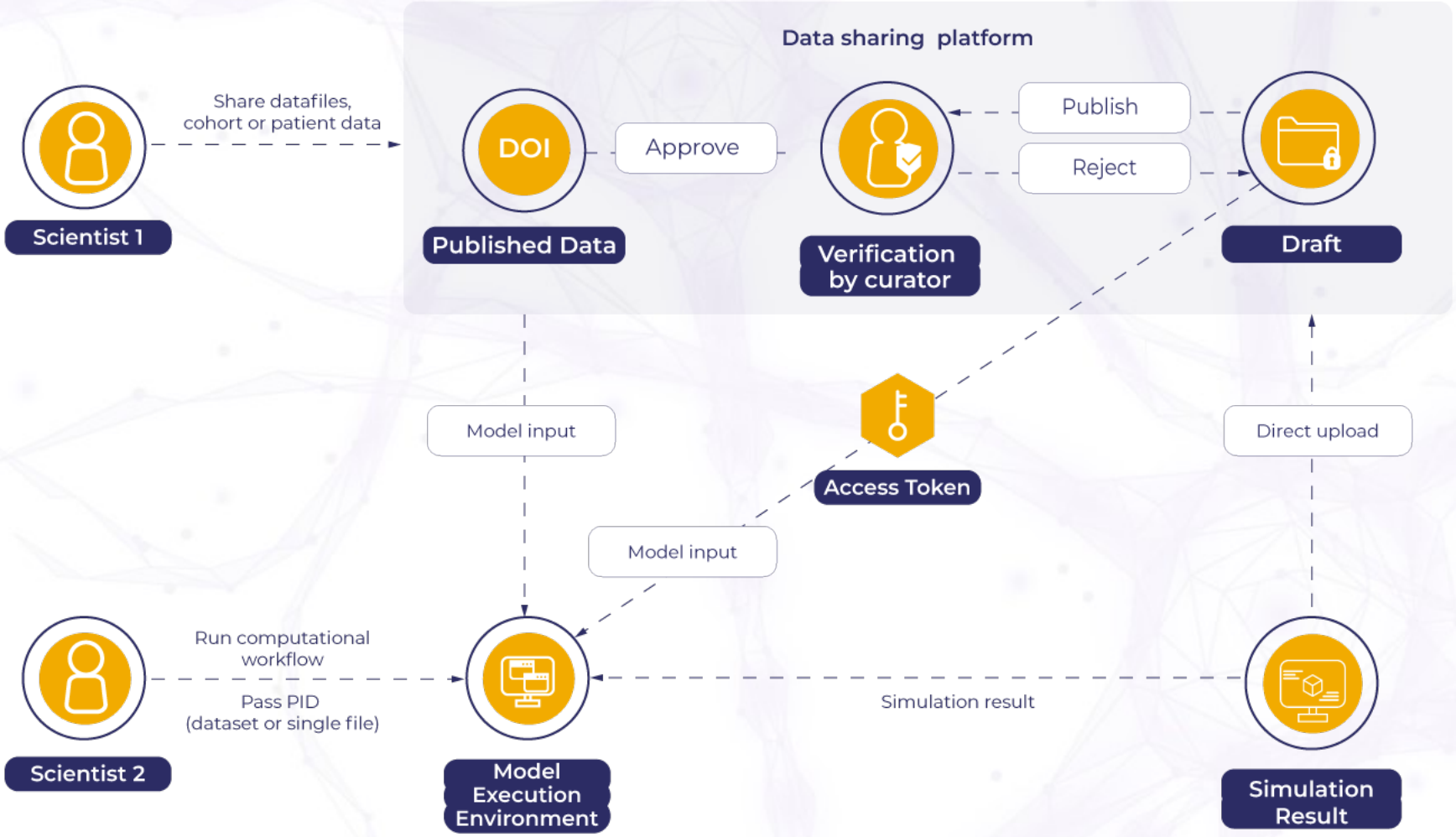
- **Virtual Human Twin** – safe and accurate experiments
- Models require computational power – HPC can help
- HPC is not easy to use for scientists
- Research data has to be **shared** between scientists
- **Publishing** results – supporting open science

Model Execution Environment – the connecting link



- MEE is developed in collaboration with the Academic Computer Centre Cyfronet
- Job execution on HPC using SSH certificates
- User-friendly interface for simulation execution
- Model versioning mechanism with GitHub and GitLab integrations
- Seamless two-way data sharing with Zenodo and Dataverse integrations

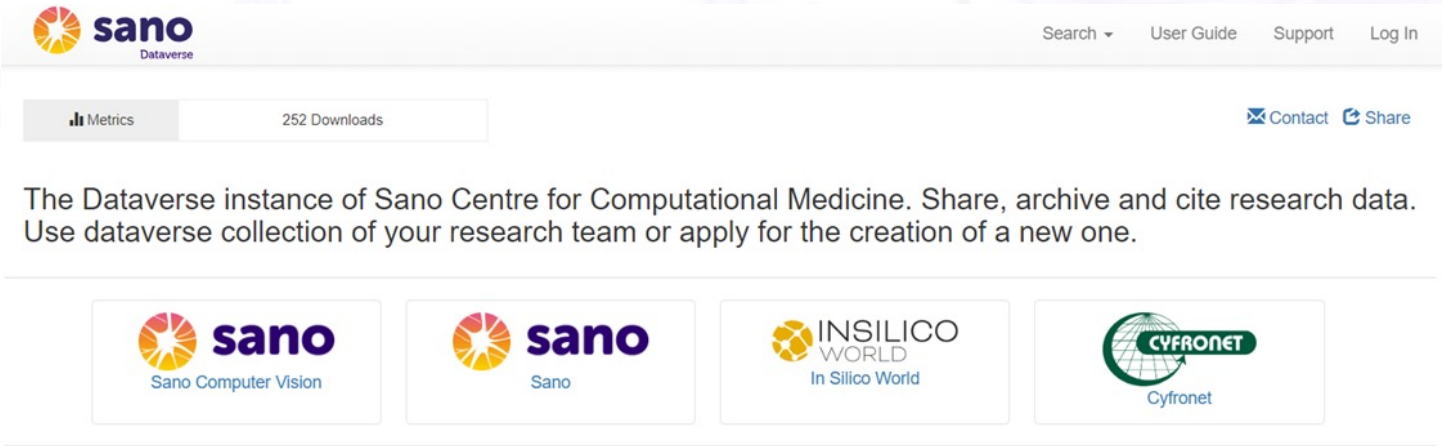
Data repositories and MEE – Example workflow



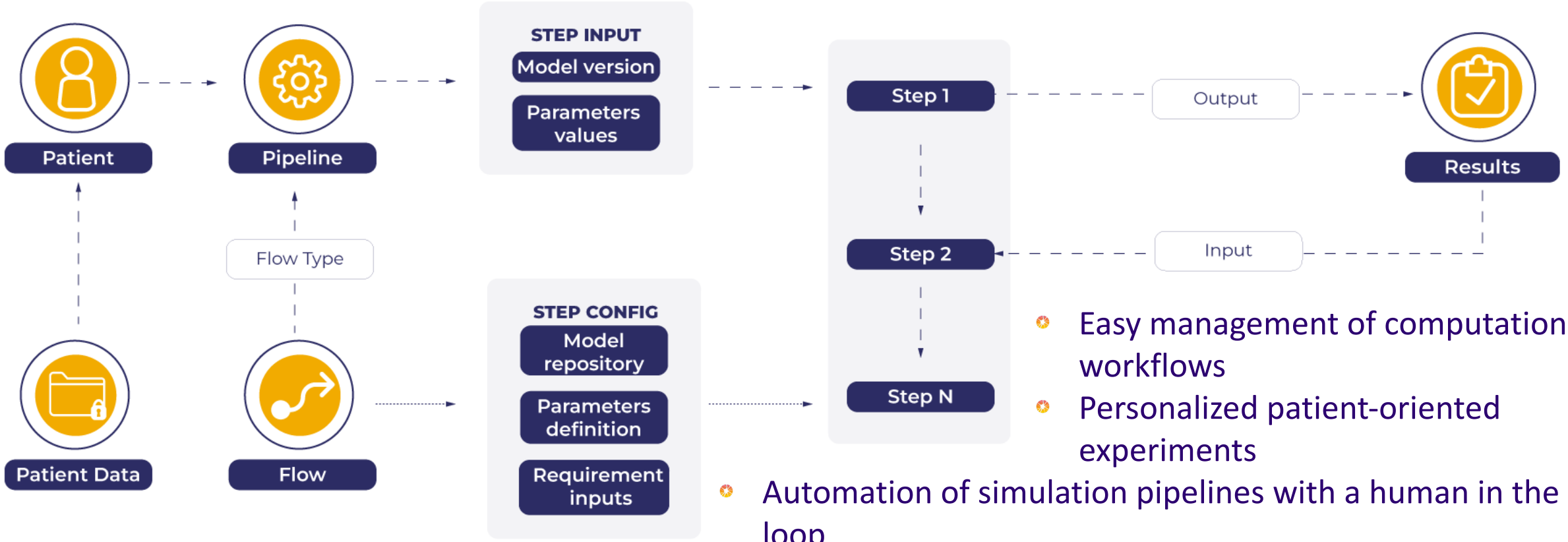
- Allows direct upload to the data-sharing platform
- Allows direct download of datasets and files as model inputs
- Datasets are verified before publication, but unpublished versions can still be used
- If a dataset is in draft form, an access token is required

Sano's Dataverse Instance

- Hosted for Sano research teams and project partners
- DataCite Fabrica account connected to register DOIs
- Joining RODBUK to disseminate datasets and ensure their reliability



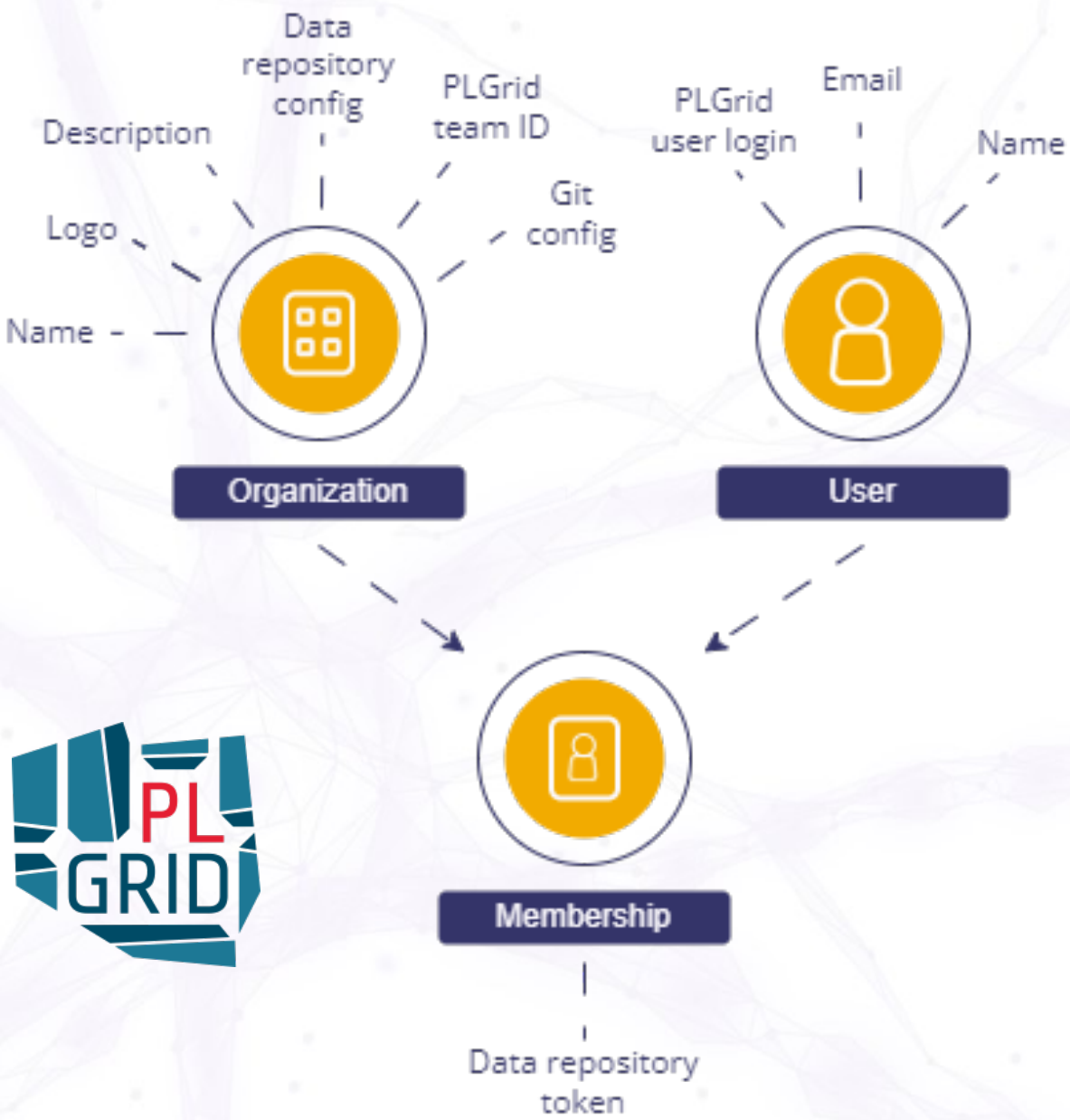
Model Execution Environment



- Easy management of computation workflows
- Personalized patient-oriented experiments
- Automation of simulation pipelines with a human in the loop
- Model versioning and pipeline history revision
- Monitoring execution status and logs of each workflow step
- Text and graphical viewers for visual comparison of results

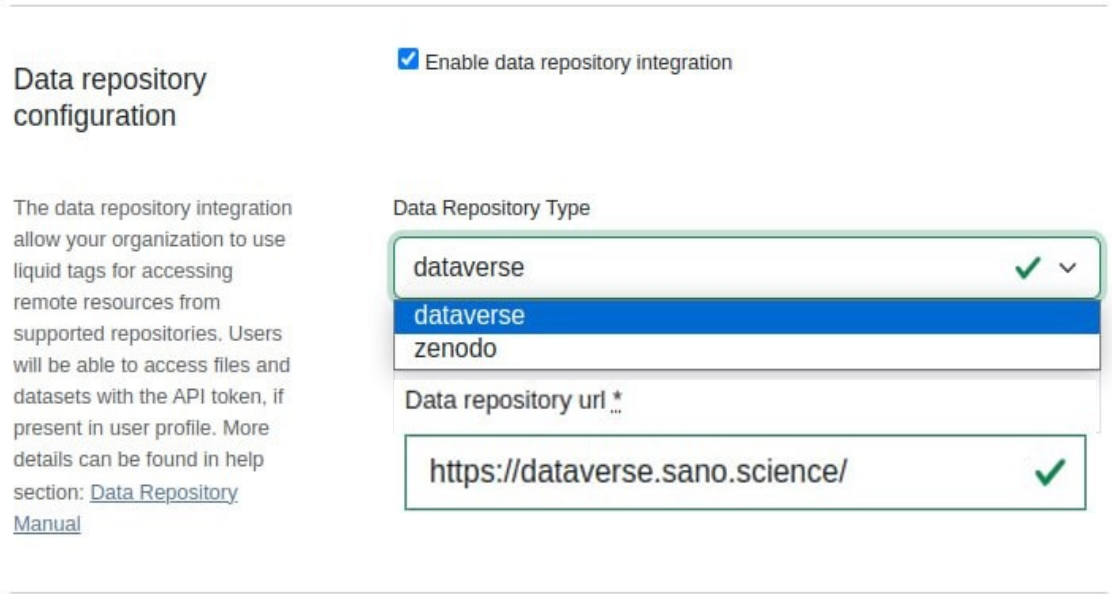
Model Execution Environment

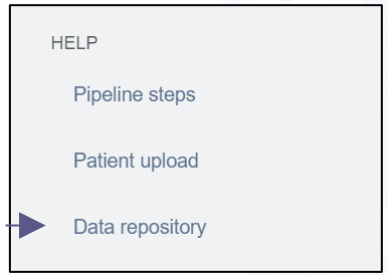
- Access delineation for HPC resources and integrated services
- PLGrid – Polish HPC managing infrastructure
- Membership – connection of a pair of a user with an organization
- Data repository token – API token from a data repository



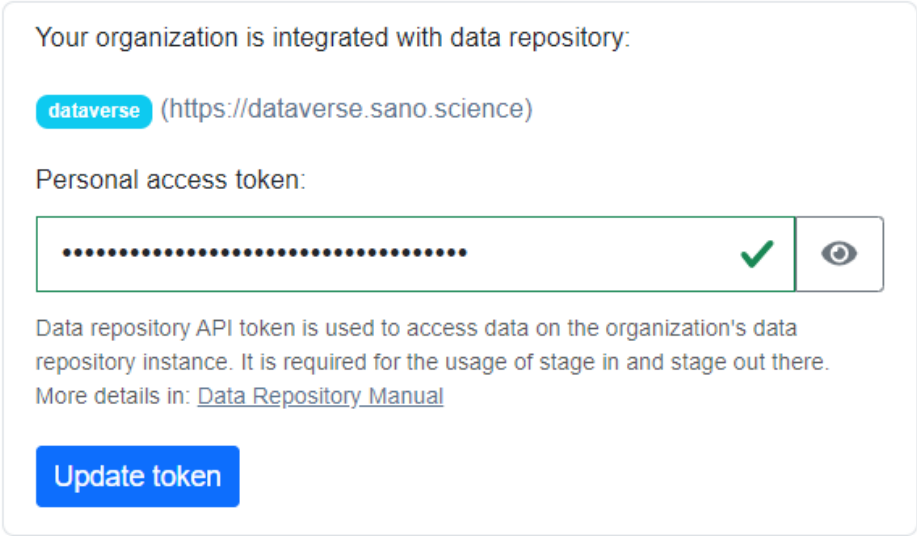
MEE and Data Repository Integration

- Data Repository Instance Configuration
  - In „Organization Details”
  - Configured by admin
- Data repository token
  - Configured only if Data Repository integration in current organization is turned on
  - Unique for every instance
  - Configured by user in each organization





User guide for the integration



MEE and Data Repository Integration

- Script features – liquid tags

```
{% dataverse_file_stage_out filePath persistentID metadata %}

{% dataverse_file_stage_in persistentID target %}

{% dataverse_dataset_stage_in persistentID target %}
```

- **filePath**: The path to the file that you want to upload to Dataverse.
- **persistentID**: The identifier for the target file or dataset in Dataverse, provided in a specific format, for example DOI (e.g., *"doi:10.12326/shoulder/23GF4H"*).
- **metadata**: An optional parameter that can be included in JSON format to provide additional file metadata. **(OPTIONAL)**
  For example:
  {"directoryLabel":"dir1/subdir","categories":["Data"], "restrict":"false"}
- **target**: The name of the file where downloaded content should be saved.(zip archive in case of dataset downloading) **(OPTIONAL)**

MEE and Data Repository Integration

- Script features – liquid tags

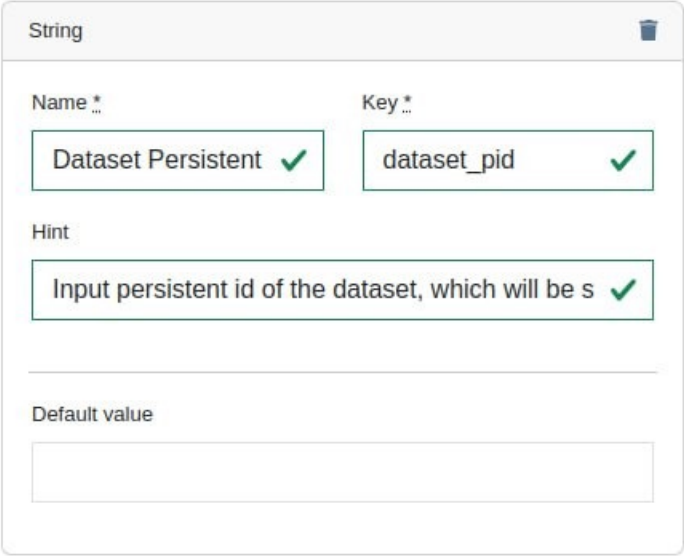
```
{% zenodo_file_stage_out filePath depositID %}

{% zenodo_file_stage_in recordID  filename  target %}

{% zenodo_dataset_stage_in recordID target %}
```

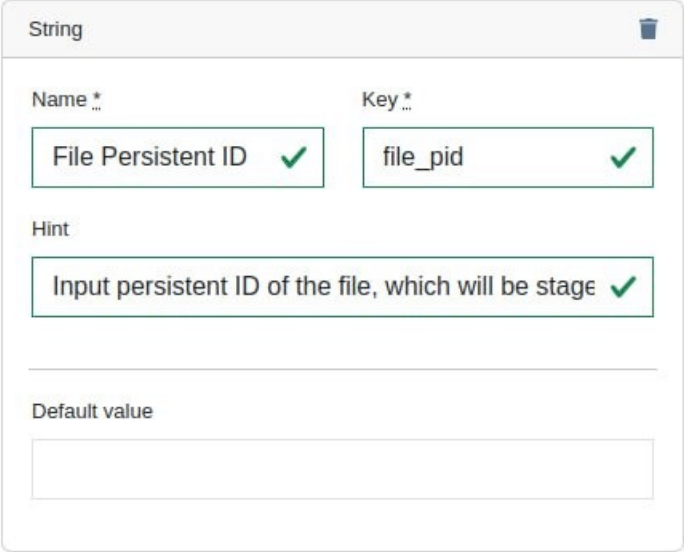
- **filePath**: The path to the file that you want to upload to Zenodo.
- **depositID**: The identifier for the upload deposition in Zenodo (may be unpublished).
- **recordID**: The identifier of the record (dataset) containing the file or a target to upload to in Zenodo .
- **filename**: The filename of the file that you want to download.
- **target**: The name of the file where downloaded content should be saved.(zip archive in case of dataset downloading) **(OPTIONAL)**

Integrated Simulation Workflow

- Example basic script and step configuration

```
echo Download dataset from dataverse
echo --------------------START----------------------------
dataset_pid={% value_of dataset_pid %}
file_pid={% value_of file_pid %}
{% dataverse_dataset_stage_in $dataset_pid %}
echo "Downloaded dataset"
{% dataverse_file_stage_in $file_pid %}
echo "Downloaded file"
echo --------------------END------------------------------

find . -type f -name "*.txt" -exec cat {} + > merged.out


echo Uploading results
echo --------------------START----------------------------
{% dataverse_file_stage_out merged.out $dataset_pid %}

{% dataverse_file_stage_out merged.out $dataset_pid {"description":"My description.",
"directoryLabel":"dataverse/subdir1","categories":["Data", "Dummy File"],
"restrict":"false", "tabIngest":"false", "jshfdbv":"sjhfd"} %}

echo --------------------END------------------------------
echo Finish
```





Integrated Simulation Workflow

- Example input and simulation result

Citation Metadata

Persistent Identifier	doi:10.82726/sano/GE7NAW
Publication Date	2023-07-26
Title	UISS-COVID19 Datafiles
Author	Zajac, Karol (Sano) - ORCID: 0000-0003-1393-8236
Point of Contact	
Description	
Subject	
Keyword	
Depositor	
Deposit Date	

dataverse_step

dataverse_step finished successfully, results stored in the outputs directory.

Computation details

Start time	27 Feb 15:45
Site	Ares
Revision	3f95751e3c0f937566fca9cfa62d9f0cc9354a2d
Execution time	00h 00m 19s
Outputs	stdout, stderr
Status	Finished

Saved parameter values

Model version	main
Dataset Persistent ID	doi:10.82726/sano/GE7NAW
File Persistent ID	doi:10.82726/sano/WOGMAR/8F9V3H
Grant	plgsano4-cpu

File Metadata

Preview

doi:10.82726/sano/WOGMAR/8F9V3H

Use the Download URL in a Wget command or a download manager to avoid interrupted downloads, time outs or other failures. User Guide - Downloading via URL

`https://dataverse.sano.science/api/access/datafile/:persistentId?persistentId=doi:10.82726/sano/WOGMAR/8F9V3H`

b8fe52d04d06845f2af9107f0b15b44d

2023-11-13

ate 2023-11-13

2023-11-13

1.0 KB

XML

Integrated Simulation Workflow

- Outcome on the repository

1 to 3 of 3 Files | Edit Files | Download

datafile_test_mild_moderate
Plain Text - 4.5 KB
Published Jul 26, 2023
42 Downloads
MD5: 569...273
Simply mild moderate

datafile_test_mild_moderate+47D11_therapeutic
Plain Text - 4.5 KB
Published Jul 26, 2023
13 Downloads
MD5: 826...34c
47D11 Treatment

datafile_vaccine+challenge
Plain Text - 4.7 KB
Published Jul 26, 2023
10 Downloads
MD5: 85a...749
Vaccine and challange

1 to 5 of 5 Files | Edit Files | Download

datafile_test_mild_moderate
Plain Text - 4.5 KB
Published Jul 26, 2023
42 Downloads
MD5: 569...273
Simply mild moderate

datafile_test_mild_moderate+47D11_therapeutic
Plain Text - 4.5 KB
Published Jul 26, 2023
13 Downloads
MD5: 826...34c
47D11 Treatment

datafile_vaccine+challenge
Plain Text - 4.7 KB
Published Jul 26, 2023
10 Downloads
MD5: 85a...749
Vaccine and challange

merged.out
Unknown - 0 B
Deposited Feb 27, 2024
MD5: d41...27e

merged.out
dataverse/subdir1/
Unknown - 0 B
Deposited Feb 27, 2024
MD5: d41...27e
My description.
Data | Dummy File

Current and Future Goals

- Large amounts of simulation data can be managed through external data repository integration
- Extreme-scale simulations are performed on MEE within the InSilicoWorld project
- Planning to make Sano's Dataverse one of the institution's primary storages
- Aiming to provide long-term storage services for both internal and partner datasets

Taras Zhyhulin
t.zhyhulin@sanoscience.org

This publication is (partly) supported by the European Union′ s Horizon 2020 research and innovation programme under grant agreement ISW No 101016503. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016227.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533 and from the International Research Agendas Programme of the Foundation for Polish Science No MAB PLUS/2019/13.

European Funds Smart Growth
Republic of Poland
FNP Foundation for Polish Science
European Union European Regional Development Fund

Minister of National Education Republic of Poland

The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796.

Sano Centre for Computational Medicine, Krakow, Poland
www.sano.science