Toolkit for Managing Research Data in Medical Simulations on HPC Infrastructure

Taras Zhyhulin, Karol Zając, Piotr Nowakowski, Maciej Malawski, Jan Meizner, Marek Kasztelnik, Piotr Połeć

Contact: t.zhyhulin@sanoscience.org   Sano website: https://sano.science

Introduction

Simulations on digital twin models are computationally and memory-intensive, so they should be performed on HPC. Therefore, data management strategies and infrastructure are needed for internal sharing and transferring data to HPC for simulations. Research-related data should be published whenever possible to support the community and uphold Open Science principles. However, once data becomes accessible, researchers often use it without contributing to the original dataset, leading to stagnation in its development. To address this, the community should be encouraged to contribute data to the datasets they use if it relates to the original research. The toolkit we designed facilitates the implementation of these practices, encouraging scientists to adopt them.

Sano Dataverse and RODBUK

- The Sano Dataverse instance was configured and deployed for Sano research teams and project partners.
- A DataCite Fabrica account was obtained and connected to register DOIs for newly published datasets.
- The instance was integrated into RODBUK to disseminate datasets and ensure their reliability.
- Cyfronet guarantees the safety and long-term storage of RODBUK members' data.
- RODBUK implements advanced security measures and peer review, enhancing the trustworthiness and credibility of published data.

MEE and data repositories integration

- Integration allows the direct download of datasets and files as model inputs and the upload of results to the data-sharing platform.
- With a simple configuration of the instance URL by the administrator and API tokens by users, any instance can be integrated.
- With an API token set, users can access draft and restricted datasets, facilitating internal data sharing.
- The script can be integrated by inserting dedicated tags into it.
- The tags are detected and replaced in the script with code snippets that interact with the corresponding repository's API.
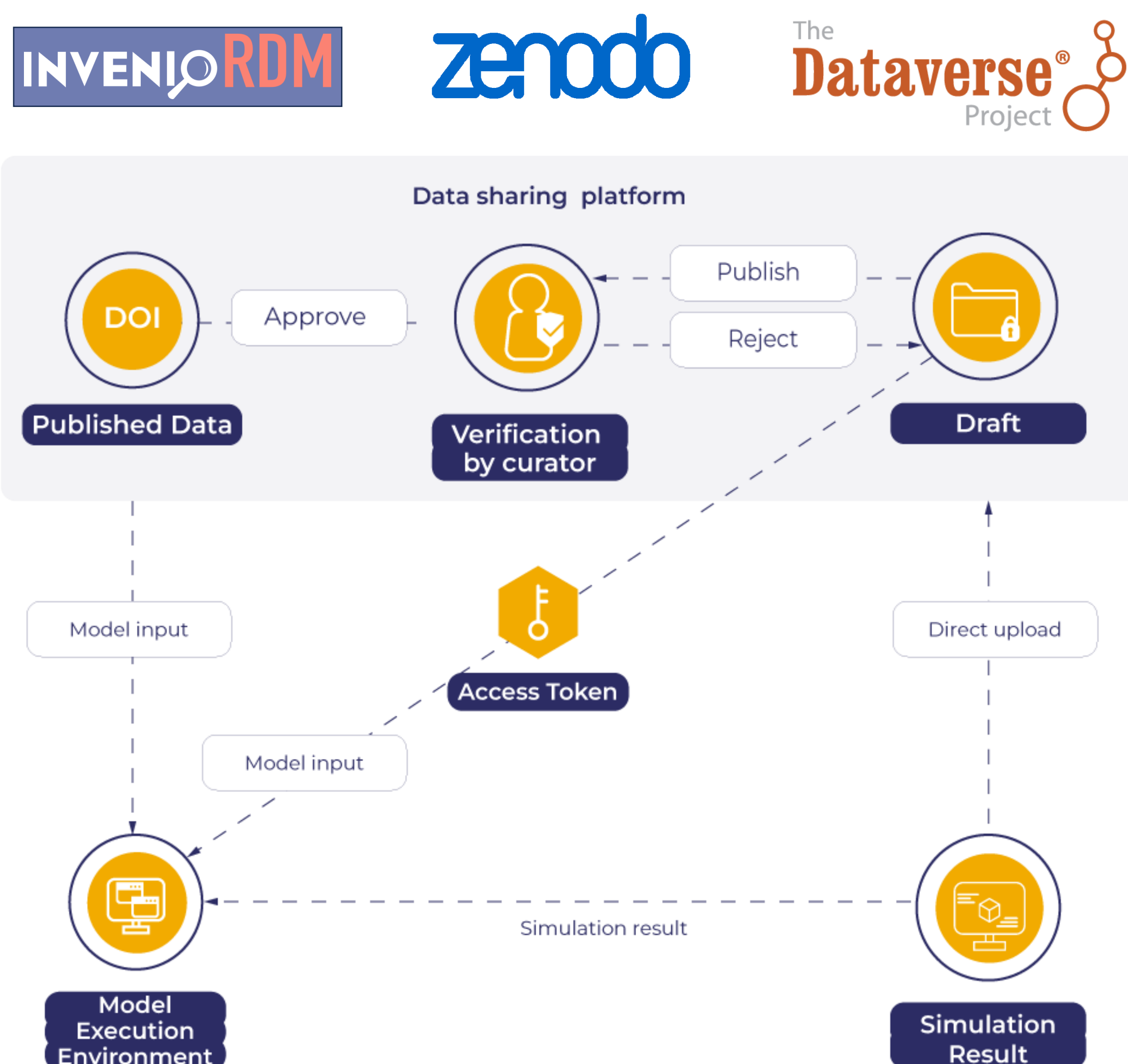
Rule-based data sharing technique

- The dataset is published under restriction or embargo.
- Interested users can request access by using the repository's contact functionality and providing a sample of their data.
- The owner reviews the sample and decides whether the contribution benefits the original dataset.
- If approved, the user is asked to provide their full contribution in exchange for a role, which provides immediate access to the whole dataset.
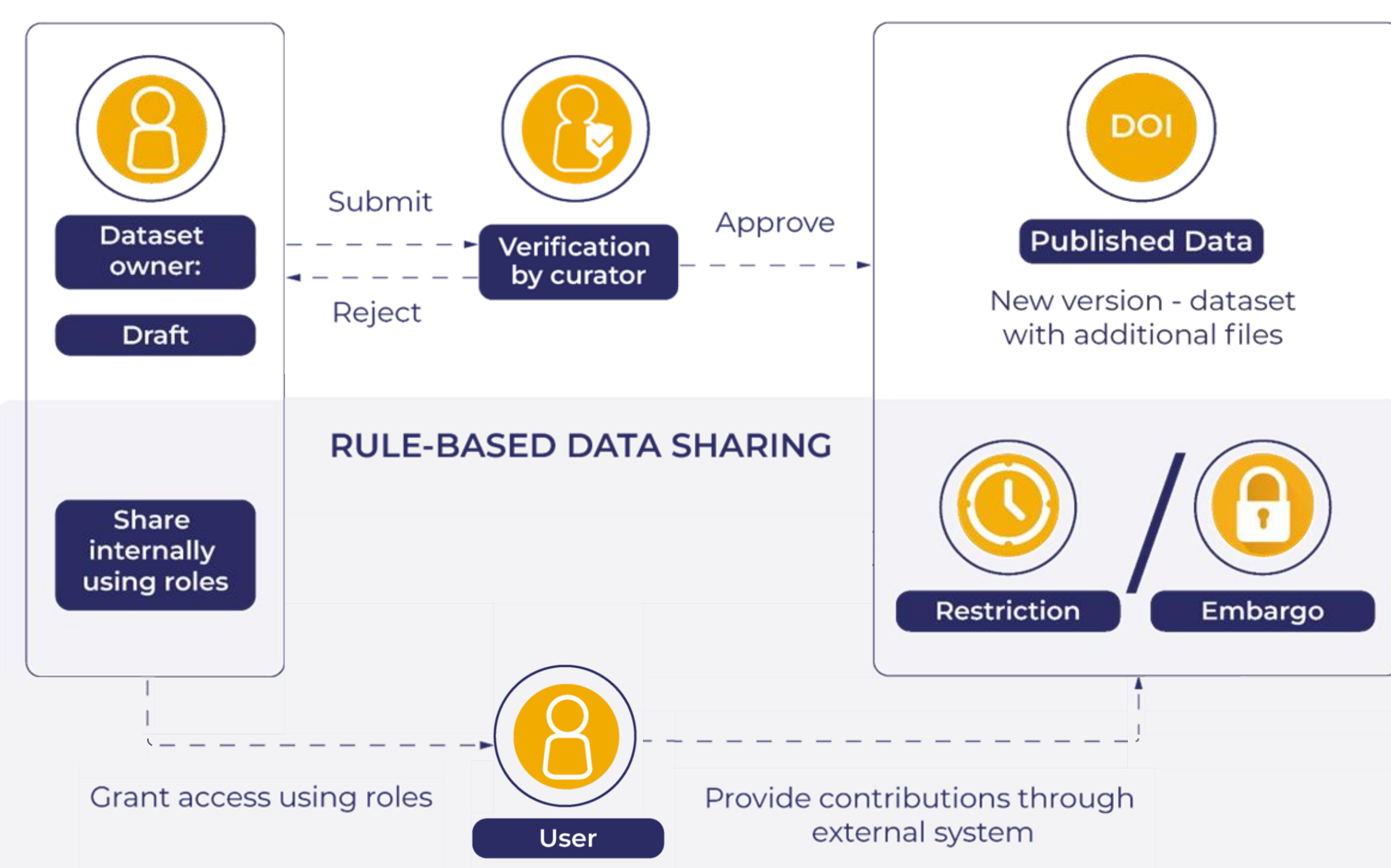
Conclusions

These solutions create a comprehensive research data management framework for in-silico medical simulations, empowering the scientific community to manage and share research data while advancing open science and collaboration. It supports scientists at various research stages, including data processing on HPC, internal data sharing, data publication, and post-publication dataset development. Additionally, the framework encourages users to uphold FAIR and Open Science principles.