**Research Data Sharing Incentivisation Toolkit**

Taras Zhyhulin1, Karol Zając1, Marek Kasztelnik2, Maciej Malawski1,2, Jan Meizner1,2,   
Piotr Nowakowski1,2, Piotr Połeć2

1Sano Centre for Computational Medicine, Czarnowiejska 36, 30-054 Kraków, Poland  
2ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Kraków, Poland  
{*t.zhyhulin,k.zajac.,m.malawski,j.meizner,p.nowakowski*}@*sanoscience.org*, {*m.kasztelnik,p.polec*}@*cyfronet.pl***Keywords**: data sharing, FAIR principles, Dataverse, RODBUK, HPC

1. Introduction

Data sharing plays a crucial role in the modern scientific community, as it can significantly reduce research costs and improve the quality of future related studies. To encourage scientists to share their data, the process must be simplified and made more convenient through the development and implementation of user-friendly tools. Additionally, providing incentives can further motivate researchers to engage in data sharing.

1. Description of the problem

Medical simulations require large volumes of input and output data, which must be securely stored and easily accessed by research teams. Without standardised, user-friendly infrastructure, researchers risk data loss, duplication, and security breaches. These simulations also demand a lot of computational power, requiring secure data transfer to external environments such as HPC.

Although data-sharing tools exist, researchers often lack integrated workflows that connect storage, collaboration, and execution. This fragmentation hinders reproducibility and discourages data sharing. Incentives for collaboration and reuse remain limited, further contributing to data loss.

Additionally, to increase citation and reuse potential, data must be properly disseminated. Therefore, a seamless, secure, and incentivised data sharing framework tailored to computational medicine is urgently needed..

1. Related work

This work is motivated by the 2009 findings of Paul Glasziou and Iain Chalmers, who reported that around 85% of reusable research data is lost due to being unpublished or poorly documented [2]. This highlights the need for structured data sharing and preservation, especially in data-intensive fields like computational medicine.

The Galaxy Project [3] offers an integrated platform for data storage, computation, and analysis across disciplines. However, its general-purpose design makes it less suited to the specific demands of medical simulations, where data sensitivity and workflow complexity require a more focused, domain-specific approach.

1. Solution to the problem

We deployed an institutional Dataverse instance, named Sano Dataverse [4], providing a secure and user-friendly platform for internal data sharing and convenient publication.

To improve visibility, credibility, and security, Sano joined RODBUK – a Polish federation of Dataverse instances. It enforces advanced security policies, aggregates datasets in a central portal for better discoverability, and holds CoreTrustSeal certification, ensuring FAIR compliance and long-term preservation.

To streamline data transfer between storage and HPC environments, we integrated the Model Execution Environment platform [5] with Dataverse, Zenodo, and InvenioRDM. This enables an end-to-end workflow: retrieving data, processing it in HPC, and uploading results back.

We also developed a rule-based data sharing strategy, illustrated by the DPValid [6] dataset on Sano Dataverse. It incentivises users to contribute related data post-publication, supporting ongoing dataset growth. The strategy uses built-in repository features and aligns with the integration framework.



**Fig.1.** Toolkit for enhanced data sharing and collaboration.

1. Conclusions and future work

The developed solutions – Sano Dataverse as part of the RODBUK federation, the integration between the Model Execution Environment (MEE) and data repositories such as Dataverse, Zenodo, and InvenioRDM, as well as the implementation of a rule-based data sharing strategy – collectively form a toolkit that simplifies the data sharing process and encourages researchers to make their data available. Furthermore, it is in line with Open Science and FAIR principles, supporting the broader shift toward collaboration and data reuse in the scientific community.

Acknowledgements. This publication is partly supported by the EU H2020 grants Sano (857533), ISW (101016503) and by the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" number MEiN/2023/DIR/3796. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017022.

References

1. „RODBUK”. URL: https://home.rodbuk.pl/en/ (visited on 2025-07-20).
2. Chalmers, Iain and Glasziou, Paul. *Avoidable waste in the production and reporting of*

*research evidence.* J. The Lancet 374.9683 (2009), pp. 86–89. ISSN: 0140-6736. DOI:

10.1016/S0140-6736(09)60329-9.

1. „The Galaxy Community”. *The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update*. Nucleic Acids Research (2024). URL: <https://doi.org/10.1093/nar/gkae410>
2. „Sano Dataverse”. URL: <https://sano.rodbuk.pl/> (visited on 2025-07-20).
3. „Model Execution Environment”. URL: <https://mee.cyfronet.pl/> (visited on 2025-07-20).
4. Davico, G., Bottin, F., Labanca, L., Baruffaldi, F., & Viceconti, M. (2024). *DPValid collection – TKA (2.0)*. DOI: 10.71580/sano/56RDVO.