D4.3
Data Management Plan

Project Title	dealii-X: an Exascale Framework for Digital Twins of the Human Body
Project Number	101172493
Funding Program	European High-Performance Computing Joint Undertaking
Project start date	1 October 2024
Duration	27 months

dealii-X has received funding from the European High-Performance Computing Joint Undertaking Programme under grant agreement Nº 101172493

Deliverable title	Data Management Plan
Deliverable number	D4.3
Deliverable version	4.0
Date of delivery	March 31, 2025
Actual date of delivery	March 28, 2025
Nature of deliverable	Report
Dissemination level	Public
Work Package	WP4
Partner responsible	RUB

Abstract	The report for D4.3 aims to provide a data management plan for the dealii-X project according to the European Commission guidelines. It presents the use of software and data in the project, the way results are structured to ensure its accessibility during the runtime of the project and beyond, and the resources devoted to achieve these goals.
Keywords	data management; FAIR data; data accessibility; data security; resource allocation

Document Control Information

Version	Date	Author	Changes Made
0.1	20.03.2025	M. Kronbichler	DPMOnline template
0.2	21.03.2025	M. Kronbichler, I. Prusak, R. Schussnig	Initial draft
0.3	26.03.2025	M. Kronbichler, I. Prusak, R. Schussnig	Partners' feedback collected
1.0	28.03.2025	M. Kronbichler, I. Prusak, R. Schussnig	Final version

Approval Details

Approved by: M. Kronbichler
Approval Date: 28.03.2025

Distribution List

- Project Coordinators (PCs)
- Work Package Leaders (WPLs)
- Steering Committee (SC)
- European Commission (EC)

Disclaimer: This project has received funding from the European Union. The views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European High-Performance Computing Joint Undertaking (the "granting authority"). Neither the European Union nor the granting authority can be held responsible for them.

Contents

1 Plan Overview

Project Title	dealii-X: an Exascale Framework for Digital Twins of the Human Body
Plan Creator	Martin Kronbichler
Principal Investigators	L. Geris, G. Brandino, F. De Giorgi, A. Caiazzo, P. Steinmann, S. Budday, L. Dede', W. A. Wall, A. Buttari, P. D'Ambra, M. Mazza, V. Cardellini, S. Filippone, G. Rozza, A. Cangiani, G. Mathias, A. Carini, A. Ghidoni, A. Salvadori, B. Meini, C. Pagliantini, S. Massei, F. Durastante, L. Heltai, M. Kronbichler
Data Managers	G. Stanic, R. Schussnig, I. Prusak, M. Kronbichler, A. Caiazzo
Project Administrators	G. Stanic, R. Schussnig, I. Prusak
Project Funder	European Commision
Project ID	173914
Project Start Date	01-10-2024
Project End Date	31-12-2026
Project Grant Number	101172493
Project Abstract	"dealii-X: an Exascale Framework for Digital Twins of the Human Body" is a pioneering project aimed at developing a scalable, high-performance computational platform using the deal.II library to create accurate digital twins of human organs, with emphasis on the human brain. This framework will leverage exascale computing capabilities and existing lighthouse applications to simulate complex biological processes in real-time, aiding in personalized medicine and advancing the diagnosis and treatment strategies of neurological disorders. The computational complexity to solve the underlying mathematical models has previously prevented the simulation knowledge from being translated into clinical practice. By integrating cutting-edge HPC technologies with multiphysics and multidisciplinary approaches, dealii-X will deliver unprecedented computational efficiency and fidelity in biological modeling. The project represents a significant leap toward the future of medical diagnostics and treatment planning, offering a robust tool for researchers and clinicians alike.

2 Detailed Data Management Plan

2.1 Data summary

State the purpose of the data collection/generation

- **Source code:**
  The source code and scientific software developed within the dealii-X project, in particular the deal.II, PSCToolkit, MUMPS codes well as the application codes, support open science. With open-source codes, advancements of the community can be accelerated, and individual results are reusable by all members of the community. All aspects of open science immediately apply here.
- **Simulation input and simulation results:**
  Simulation results of large-scale simulations of various complex physical phenomena mainly serve three purposes: i) verification of newly developed software, ensuring its correctness in well-established benchmark problems, ii) comparison of performance in various aspects of a numerical method, and iii) enabling reuse by the community not primarily focused on the derivation of simulations tools themselves, but in further processing the data to perform data analysis such as uncertainty quantification or sensitivity analysis, model order reduction by traditional surrogate modeling techniques, polynomial chaos expansion or methods based on machine learning.

Explain the relation to the objectives of the project

The data collections and open-source software development are central parts of the project's identity. The overarching goal of the dealii-X project is to make digital twin simulators for exascale supercomputers directly accessible to the scientific community and to provide examples for successful transfer from pre-exascale application codes by means of leveraging the exascale-ready building blocks and capabilities of the deal.II library. Hence, keeping the underlying repositories and source codes open-source is pivotal for the success of the project, which is in some sense measured by the reception of the approaches established within the dealii-X project. The data collections comprising problem descriptions, required data and related simulation results of various physics problems and digital twin applications add another layer of reusability and interoperability to the obtained results.

Specify the types and formats of data generated/collected

- **Source code:**
  The source code developed will be primarily in the programming languages C++ and C.
- **Simulation input and simulation results:**
  Solution quantities and derived measures to ease comparison, reduce storage and improve reusability will be collected. The data will primarily be generated in plain text, binary output of VTU type, HDF5, and csv data for reduced output. The data format will be adapted to the individual problems and assessed continuously during the project's run time. Decisions will be made by the individual partner in charge of the respective work package.

Specify if existing data is being reused (if any)

- **Source code:**
  The deal.II, PSCToolkit and MUMPS libraries are part of a larger open-source software network that have been developed over many years. These software dependencies are to be considered, but the source code is not duplicated. The interfaces will be adopted to use the provided functionality. The specific resources reused here are to be distinguished per code project and are summarized in the related software dependencies overviews.
- **Simulation input and simulation results:**
  Depending on the specific digital twin application, different data sources are being used as the problem formulations of interest and the derived simulation results are based on openly available data. This data is either i) already openly available to the general public, or ii) will be made public through a data repository in accordance to European and national laws. Decisions in this regard will be made by the individual partner in charge of the respective work package.

Specify the origin of the data

The data will be generated by the individual partners themselves, meaning, as the product of software development, or by executing and testing the developed application software. Since many of the sought-after improvements and optimizations

can only be verified and measured in real applications, the data generation is part of the natural software development and testing cycle. The extended datasets produced for further reuse by the community for the purpose of data analysis is then a minor incremental effort once the application runs operates on the exascale and delivers credible results.

State the expected size of the data (if known)

- **Source code:**
  10-50MB of source code is expected.
- **Simulation input and simulation results:**
  Several tens of TB will be generated, which depends on the individual digital twin application. Reduced storage will be used where viable, including output of average/rms quantities, sparse output, and similar. Decisions in this regard will be made by the individual partner in charge of the respective work package.

Outline the data utility: to whom will it be useful

- **Source code:**
  The source code provided will be useful to the entire open-source scientific software community and beyond. Exchanging and reusing software components and profiting from each others' improvements and sourcing inspiration from each others' individual tailored solutions is an integral part of the open-source culture.
- **Simulation input and simulation results:**
  Scientists in the field aiming to verify their simulators will profit from having access to reference results and data scientists will profit from the datasets made available, based on which alternative methods might be derived to obtain cheaper approximations, or through which a detailed statistical analysis is made possible.

2.2 FAIR data

2.2.1 Making data findable, including provisions for metadata

Outline the discoverability of data (metadata provision)

Concerning the discoverability of both data and software contributions, all datasets and code will be accompanied by detailed and standardized metadata to ensure their discoverability in open repositories such as Zenodo and Git. The metadata will be structured to support efficient search, retrieval, and indexing across repositories and search engines, facilitating access to both the datasets and the software. This ensures that both data and code are discoverable, supporting the transparency and reproducibility of the entire research process.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

All publicly available datasets and software will be assigned a Digital Object Identifier (DOI) via Zenodo or other recognized repositories. The use of DOIs guarantees the long-term accessibility, traceability, and proper citation of both datasets and software. By assigning DOIs to both data and code, we ensure that these contributions are reliably identifiable and can be properly referenced in scientific publications, maintaining their persistence and accessibility over time.

Outline naming conventions used

A structured naming convention will be implemented for both datasets and software contributions to maintain clarity, consistency, and ease of identification. The naming scheme will include:

- Data type (e.g., input, output, mesh, solution)
- Project component (e.g., work package number, task identifier)
- Version number (for traceability and reproducibility)
- File format extension (e.g., .vtk, .csv, .hdf5)
- Software-related tags (e.g., simulation model version, solver used)

Outline the approach towards search keyword

Each dataset and software package will be tagged with relevant keywords that reflect their content, context, and specific computational methods. These keywords will be selected in alignment with standardized scientific terminology, ensuring that they can be easily indexed, searched for, and retrieved across repositories and search engines. Special attention will be given to including keywords related to both the data and the software, making it easier to discover the full research framework.

Outline the approach for clear versioning

Version control will be a central aspect of the project. All code will be managed through Git repositories, providing a detailed history of versions, contributors, and associated metadata. For datasets, Zenodo or similar repositories will be used, where version numbers will be assigned to ensure traceability and reproducibility. Every dataset and software version will be tagged with a structured numbering system (e.g., v1.0, v1.1) to ensure that both data and software are correctly versioned and easily accessible.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

The metadata associated with both datasets and software will follow widely recognized standards. Each dataset and software package will include key metadata elements, such as:

- Title and description (providing clear identification of the dataset and the software used)
- Authors and affiliations (acknowledging the contributors to both the data and the software)
- Keywords (to enhance searchability and discoverability, with a focus on software-related terms)
- Creation date and versioning (to ensure traceability and proper version control for both data and software)
- Licensing information (outlining the usage rights for both datasets and software)

- Associated publication references (if applicable, linking datasets and software to relevant publications)

2.2.2 Making data openly accessible

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

As much of the software and computational tools will be made openly available as possible, following the best-practices of open-source software the partners have established in the past. Simulation data used in scientific publications will also be made publicly accessible, ensuring transparency and reproducibility. In this context, it can be expected to cover the vast majority of all research results, covering all core algorithms. Any data that may be subject to restrictions, if applicable, will be handled in accordance with legal or ethical guidelines. However, the project does not expect to work with highly sensitive data.

Specify how the data will be made available

The scientific software will be made available through the main repositories of the deal.II, PSCToolkit, and MUMPS projects, as well as through individual application software projects like lifex, 4C, ExaDG, polyDEAL, LiverX, and ExaBrain. These repositories will be organized on GitHub/GitLab instances and will include full documentation for the codes and contributions, supporting transparency and reproducibility.

Data will be shared through Zenodo, GitHub, and institutional repositories. Publications will include direct links to the relevant datasets, supporting verification and reproducibility, ensuring that both data and software are easily accessible. The datasets will be linked to the corresponding software versions used to generate them, promoting complete reproducibility.

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

To access the simulation data, scientific computing software, such as deal.II, PSCToolkit, and MUMPS, will be required. Full documentation for data formats and

processing steps will be provided to facilitate reuse. The necessary software tools, along with instructions for running simulations and processing the data, will be included as part of the metadata and documentation. Where possible, the relevant software will be made available as open-source code, allowing full access to both the data and the computational tools used to generate it. Additionally, the dealii-X project collaborates with the CASTIEL 2 project at the EuroHPC level, ensuring adherence to common standards and conventions within the scientific computing community.

Specify where the data and associated metadata, documentation and code are deposited

Data, associated metadata, and documentation will be deposited in:

- Zenodo (for curated datasets and publication-related data)
- GitHub/GitLab (for software and version-controlled simulation data)
- Institutional repositories (for long-term archiving)

These repositories will be structured to ensure proper versioning, clear attribution, and seamless access to both the data and the software needed to reproduce the results.

Specify how access will be provided in case there are any restrictions

For any datasets with restrictions, access will be granted upon request with justification to ensure compliance with EU ethical and legal requirements. The procedure for requesting restricted data will be clearly outlined in the repository documentation, ensuring transparency and proper handling of any restricted data.

2.2.3 Making data interoperable

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

Data formats will be aligned with widely recognized scientific standards, such as HDF5, VTK, and CSV, to ensure interoperability across different platforms and

research domains. Project-specific standards will be defined for metadata to ensure seamless integration with the data and software.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow interdisciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

The standard vocabulary will be used for all data types present in our datasets, ensuring inter-disciplinary interoperability. In cases where in-house formats are necessary, conversion tools will be developed as part of the individual software projects to translate data into widely accepted formats. These tools will facilitate data exchange across domains while maintaining compatibility with other research areas.

2.2.4 Increase data reuse (through clarifying licenses)

Specify how the data will be licensed to permit the widest reuse possible

As much of the datasets and software produced in the project as possible will be made publicly available under open-source licenses, following the partners' established best practices in open-source development. Licensing will align with the policies of the relevant software repositories, ensuring the widest possible reuse while maintaining compatibility with the broader scientific computing community.

Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed

Data will be made available upon publication. No embargo will be applied, except for proprietary data, and if required, any embargo will not exceed 12 months. This ensures that data will be publicly accessible in a timely manner, promoting transparency and reproducibility.

Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the reuse of some data is restricted, explain why

Datasets and software will be designed for reuse by third parties, even after the project's completion. Any reuse restrictions will be based solely on legal or ethical considerations, but highly sensitive data is not anticipated in the project.

Describe data quality assurance processes

Data quality will be ensured through:

- Version control and automated testing for software
- Peer review of datasets before publication
- Consistency checks with validation scripts to ensure the integrity and accuracy of the data

Specify the length of time for which the data will remain reusable

Data will remain accessible for at least five years beyond the project's completion, with institutional repositories ensuring long-term archiving.

2.3 Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

The open source development of code via platforms like github or gitlab as employed within the present project is low. The main costs stem from professional workflows offered by these platforms to automatically and routinely check code/hardware compatibility, as well as performance and integration testing. The related costs are already covered by external sources unrelated to the present project, hence zero costs in relation to this project are generated. The costs for computing large sets of simulation data are to be covered by the individual partners. These resources are either freely available to some partners, or need to be acquired by taking part in respective calls by the remaining partners. Access to supercomputing facilities, e.g. the SuperMUC-NG, JUPITER, Fritz and Alex systems at FAU Erlangen Nürnberg in Germany, or the Leonardo system for Italian partners, are set up in this phase of the project, and will be provided through the respective funding of this infrastructure.

Clearly identify responsibilities for data management in your project

The PIs and principal developers of the respective repositories are responsible for guaranteeing the full adherence to the FAIR principles. Consequently, the PI of

the respective partner ensures that the data provided by the principal developers and all involved persons adheres to the FAIR principle at the time of generating a dataset or source code for publication. The steering committee oversees the adherence to the FAIR principle within the dealii-X project.

Describe costs and potential value of long term preservation

Long-term storage of generated data including backup are to be organized by the respective partner. In case resources are required for long-term storage, the resources reserved for publication purposes are recruited for this purpose.

The value of long-term storage of the simulation results lies in the grade of reusability of the generated results. Hence, the FAIR principle increases the chances of the value of the preserved data being valuable in the future. This value is only in scarce cases to be understood as a monetary value, but in terms of scientific merit gained from making the data accessible publicly. Therefore, the FAIR publishing of datasets will always be connected to related scientific output such as journal articles, conference proceedings or contributions to any other kind of peer-reviewed written scientific output with a digital object identifier.

2.4 Data security

Address data recovery as well as secure storage and transfer of sensitive data

The institution selected for long-term storage by the partners of dealii-X has to provide regular back ups of the stored data and version control if applicable. The datasets do not contain sensitive data, as only anonymized data previously cleared for public release by agencies unrelated to the present project will be used within dealii-X. The transfer of sensitive data is hence not an issue.

2.5 Ethical aspects

The ethical aspects are covered in the context of the ethics review in the ethics section of DoA and ethics deliverables. No further comments on ethical aspects are added at this point.

2.6 Other

No further data management regulations are to be followed, as the FAIR principle covers the requirements formulated by the involved funding organizations.