Clara

Center for Artificial Intelligence and Quantum Computing in System Brain Research

Deliverable 2.4

Title: Data Management Plan (DMP)

Funded by the European Union

Project Number: 101136607

Project Acronym: CLARA

Call: HORIZON-WIDERA-2023-ACCESS-01

Project title	**Center for Artificial Intelligence and Quantum Computing in System Brain Research (CLARA)**
Grant Agreement Nr.	101136607
Funding scheme	HORIZON Coordination and Support Actions (CSA) Call: HORIZON-WIDERA-2023-ACCESS-01 Topic: HORIZON-WIDERA-2023-ACCESS-01-01-two-stage - Teaming for Excellence
Project duration	1 November 2024 - 31 October 2030 (72 months)
Project Coordinator	INDRC – International Neurodegenerative Disorders Research Center
Deliverable number	**D2.4**
Title of the deliverable	**Data Management Plan (DMP)**
WP contributing to the deliverable	**WP2**
Deliverable type	**DMP** R: document, report; DEM: Demonstrator, pilot, prototype, plan designs; DEC: Websites, patents filings, press & media actions, videos, etc.; OTHER: Software, technical diagram, etc.
Dissemination level	**PU** PU = Public, fully open, e.g. web; CONFIDENTIAL-C; RESTRICTED-R; SEN – Sensitive
Due submission date	**30 April 2025 (M6)**
Actual submission date	30 April 2025 (M30)
Author(s)	Vit Vondrák, Kateřina Slaninová (VSB)
	Smaran Chaudhary, Stephan Hachinger (BADW-LRZ), RP leads
Internal reviewers	Cross-reviewed by VSB/BADW-LRZ co-authors, RP-specific parts by RP leads.
Final approval	Václav Snášel

Document Status – History of Changes

Version	Date	Contributor(s)	Description of Changes
v0.1	28/02/2025	Smaran Chaudhary, Jan Martinovic, Katerina Slaninova	Concept and First Draft
v0.2	13/03/2025	Stephan Hachinger	Reviewed and completed for being worked on by RPs
v0.3	11/04/2025	RP leaders, LRZ	With RP inputs and revision/harmonisation
v0.4	14/04/2025	Smaran Chaudhary	LRZ cross-review, contents beta-final, typos corrected
v0.5	15/04/2025	Stephan Hachinger, Katerina Slaninova	Style corrections, RP3 additions, IT4I input, last corrections by RP3 and partners
v0.6	18/04/2025	Jiri Damborsky	Additions to RP1 sections
v1.0	23/04/2025	Jean-Marie Boutellier, Ara Khachaturian, Stephan Hachinger	Completed RP3 Sections, other minor corrections.
v2.0	27/04/2025	Vít Dočkal	Minor formal revisions, spellcheck
V3.0	27/04/2025	Vít Dočkal	Revisions accepted, pre-final version
FINAL	29/04/2025	Václav Snášel	Final version

Document Keywords

Data Management Plan, CLARA, Research Data Management

Confidentiality

Does this report contain confidential information?	Yes ☐ No ☑
Is the report restricted to a specific group?	Yes ☐ No ☑ *If yes, please precise the list of authorised recipients*:

Table of Contents

Executive Summary

This Data Management Plan (DMP) specifies the main aspects of the life cycle of the project data (organisation and their long-term storage, access, preservation, and sharing). This document also includes a preliminary specification of outputs (what data will be generated during the project). This document will be continuously updated during the project (updates foreseen in M30, M60). The DMP contains a discussion of the most relevant aspects of the CLARA research and infrastructure efforts and of their role in relation to the FAIR (Findable, Accessible, Interoperable, Re-usable) principles of research data management. The main part of the deliverable contains a summary of datasets that will be used and produced in the Research Programmes (RP1-RP3) and the Center of Excellence (CoE) efforts in general (see Section 2). The project's approach to complying with the FAIR data principles is described, together with a recap of the principles themselves, in Section 3. Sections 4 briefly touches upon ethical and security issues. Our plans for managing the project's outputs in the FAIR spirit is further detailed upon in Section 5. The outputs are again divided into sections related to the RPs and to general CoE activities. Section 6 concludes the document.

1. Introduction

CLARA (Center for Artificial Intelligence and Quantum Computing in System Brain Research) represents the very first interdisciplinary Center of Excellence (CoE) in the Central and Eastern Europe (CEE) focused on utilising the next generation of artificial intelligence/machine learning (AI/ML) applications and quantum-accelerated supercomputing tools to solve the etiology of neurodegenerative diseases (NDs).
The setting of the CoE, where top-grade research with massive amounts of (partially sensitive) data is accompanied by significant infrastructure support sets the scene for our data-management efforts. In strong previous collaboration of the computing-centre partners (VSB – IT4Innovations National Supercomputing Centre, BAdW-LRZ – Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities), data-management concepts for computational workflows have been developed which can be applied to CLARA. These, in particular, pertain to the Research Data Management (RDM) concepts in the LEXIS platform[1] [1,2,3], which has emerged from the homonymous Horizon 2020 project[2] (GA #825532). These methods are currently evolved by the partners within the EXA4MIND Horizon Europe project[3] (GA #101092944), and strongly take into account the FAIR principles [4] of RDM according to the state of the art. The resulting RDM planning for CLARA is laid out in the present **Data Management Plan**, which will be evolved with further releases at Months 30 and 60.
Below, we give some more details on project hardware – in particular storage – used for data management (Section 1.1), and on which external systems and repositories we plan to involve in our data-management efforts (Section 1.2), as well as the personnel helping with these efforts (Section 1.3). Section 2 of the DMP introduces the datasets used and produced in the Research Programmes (RP1-RP3) and the CoE efforts in general (see Section 2). The project's approach to complying with the FAIR data principles is described in detail, together with a recap of the principles themselves, in Section 3. Section 4 briefly touches upon ethical and security issues. Our plans for managing the project's outputs in the FAIR spirit is further detailed upon in Section 5. The outputs are again divided into sections related to the RPs and to general CoE activities. Section 6 concludes the document.

1.1. Project Hardware and Storage

Within the project consortium, we have two HPC centres (European and national level): BAdW-LRZ and IT4I at the VSB; both have significant experience in the operation of multi-petascale computational resources at Tier-0 and Tier-1 levels. The centres support diverse hardware architectures from a range of vendors and will provide a range of environments for the execution of research applications. In particular, both IT4I and BAdW-LRZ feature quantum-computing and GPU-computing systems essential for CLARA's ambitious research concept. The HPC centres also host significant data storage with fine-grained access-rights management, various security levels and remote-access possibilities – an essential basis for the management of primary output data in particular. HPC resources are usually provided through local/regional, national or EuroHPC JU access calls. A summary of the key HPC resources and data storage, as well as access modalities and conditions is available at the institutional documentation websites[4]. Furthermore, the CLARA project can build upon distributed data and computing infrastructure from the LEXIS project - the LEXIS Platform[5].
Existing systems at LRZ and IT4I (which can be integrated into the CLARA Testbed as appropriate) will be augmented by two dedicated main Testbed parts: The GPU based compute part in form of a cluster (further addressed as HPCQC-AI cluster) and an intermediate storage solution (part of so called Project storage).
The HPCQC-AI cluster consists of three units (GPU computing part, scratch storage and high speed interconnect network):
a) The GPU computing part of the system will consist of approx. 78 so-called razor blades, each of which will contain a so-called node consisting of 1x CPU and 2x next-generation GPU units. Internally, there will be a high-speed

[1] https://www.lexis.tech; https://docs.lexis.tech
[2] https://lexis-project.eu
[3] https://exa4mind.eu
[4] https://doku.lrz.de; https://docs.it4i.cz
[5] https://www.lexis.tech; https://docs.lexis.tech

connection between the CPU and GPU, which will ensure coherence up to the cache memory level. This technical solution will enable very efficient training of large-scale AI models and simulation of quantum systems with high numbers of q-bits. The expected memory throughput to the memory from the local GPU will be at least 5.5 TB/s and from the CPU at least 900 GB/s. The memory throughput to the next GPU will be approx. 1800 GB/s. Each node will be connected to both other nodes and data storage using a high-speed network (see below). One computing node will have an expected maximum (theoretical – Rpeak in FP64 precision) performance of approx. 180 TFlop/s with a total GPU computing performance of approx. 14 PFlop/s. In calculations for AI operations (FP32 precision for dense matrix operations), one node will have an expected maximum theoretical performance of approx. 5 PFlop/s with a total performance of all GPUs of approx. 390 PFlop/s. The expected consumption is approx. 237 kW using direct liquid cooling (DLC) and a water inlet temperature of at least 36 °C.

b) SCRATCH storage will be completely based on flash technology with an emphasis on performance and a large volume of data read and write (so-called DWPD). The software layer will provide shared, high-performance, parallel storage for all computing nodes with the possibility of direct access to the file system from GPU accelerators. The storage will ensure both fast data loading into GPU accelerators and the possibility of fast saving of intermediate state of simulations (or neural network training) for the needs of a possible restart (so-called "checkpoint and restart"). The storage will be connected to the computing nodes via a high-speed network (see below). The expected total storage capacity will be approx. 2665 TB of gross capacity and throughput of approx. 1000 GB/s and intensity of IO operations up to 6145 kIOPS. The estimated storage consumption will be approx. 34 kW.

c) A specialized high-speed network will be built to connect the above-mentioned units. This network will be built on technology enabling transmission speeds at the physical layer of at least 400 Gb/s. The software part of this network will enable direct access to the memory of the GPU accelerators of the computing part using RDMA technology. The network configuration will include 4x 400Gb/s ports for each GPU node. The network will be sufficiently dimensioned to connect data storages, in such a way as to enable the GPU part to achieve high aggregated read and write throughputs (especially for the powerful SCRATCH storage). The topology of this high-speed network will be a fat-tree with a blocking factor of 1:2 or similar in performance. The total consumption will be approx. 8 kW.

The intermediate storage solution will be built to provide medium-term storage capacity for both the active parts of the project life cycle and for subsequent parts where data needs to be available for later (re-)validation, or for projects that are, by their nature, a continuation of the original project.

As part of the project, a data storage will be purchased that will provide data capacity for the needs of the project. The repository will be integrated with the HPCQC - AI computing cluster and will also enable the so-called fairification of research data as an intermediate link between computing capacities and data repositories. The storage will provide at least the NFSv4 protocol, snapshot functionality, extended ACLs and quotas at the software layer. The estimated storage capacity is approx. 5.3 PB of net, user-available capacity. The estimated storage consumption is approx. 9.5 kW.

1.2. Systems for Holding FAIR Data - Including Domain Science, Institutional and General (European and International) FAIR Data Ecosystems

Going beyond the management of input data and freshly generated data on appropriate file systems at the computing centres, FAIR management of “lukewarm” and “cold” data (i.e. data which is unchanged in the mid- to long term) becomes a priority. FAIR RDM – a de-facto standard approach in management of scientific data today – focuses on Findability, Accessibility, Interoperability and Re-usability of data, for which the assignment of appropriate metadata and unique, persistent identifiers (PIDs, e.g. Digital Object Identifiers – DOIs – according to ISO 26324:2025) to each dataset produced is paramount [4]. While we describe our measures for following the FAIR principles in more detail in Section 3, the systems for holding FAIR data are briefly described here.

FAIR management of warm data is a key feature of the LEXIS platform (see above) and its distributed data infrastructure [3]. The LEXIS distributed data infrastructure uses the iRODS data middleware [5] for data and

metadata storage, with EUDAT's B2SAFE[6] module guaranteeing the assignment of EUDAT-B2HANDLE[7] PIDs to data where appropriate.
For publication of cold data, Zenodo[8] (where a CLARA "community" is being created to bundle CLARA contributions) is certainly one of the most-used general-purpose FAIR research data repositories in Europe. It allows for (free) uploads of 50GB per data product by default and is operated with support of CERN and the EU. Where more domain specific research data repositories[9] exist, they will be preferentially targeted. As a fallback possibility and for bundling data products by institution, the academic partners in CLARA run research data repositories[10] as well. They have also been exploring mechanisms for publishing large datasets directly from persistent HPC storage (or HPC-centre archives) without the need to involve a repository; a service prototype which can already be used on matching occasions is available e.g. at LRZ[11].
For source codes and computational model data, specialised platforms excel in facilitating adapted data management and sharing. The computing centres in CLARA run institutional GitLab-based platforms[12] for (open and non-public) source-code management (SCM), which can be combined with PID/metadata management and code publication via the institutional RDM solutions mentioned above. The probably most used international public SCM platform is GitHub[13], which has been integrated with Zenodo for FAIR SCM. Models, mostly AI/ML-based, are globally shared on platforms such as Hugging Face[14] and Weights & Biases[15], and on application-and-domain-specific platforms such as ModelDB[16]. The relevance of these platforms makes it crucial for CLARA to consider them for primary or secondary publication of source codes and models.

1.3. Allocation of Personnel

The appropriate effort for the creation of the data management plan is allocated at IT4I and at BAdW-LRZ, where the latter leads the Task 2.3 "Data management" (M1 - M72). These institutions play also key roles within WP5, where the testbed is created and managed, practically implementing the activities envisaged within the DMP. Data management related activities – mixed with some workflow-/testbed-management activities – are thus backed in total by about one dedicated FTE at BAdW-LRZ. All project partners will contribute to the DMP and its continuous updates during the project, and will work with data, metadata, and other outputs according to the principles defined in this DMP.

2. Summary of relevant data

This section describes the input and output data that will be involved in CLARA research. The project encompasses a diverse range of research activities, necessitating a data-type agnostic RDM accommodating any data type.

2.1. RP1 - Quantum-accelerated Supercomputing and Machine Learning (ML) to Address Protein Dynamics, Aggregation, and Modulation by Small Molecules

RP1 will develop and train new ML foundation models for understanding the dynamics and aggregation of proteins involved in NDs and their interactions with other biomolecules, focusing on crucial proteins like ABeta, APOE, and Tau. The initial focus will be on studying three isoforms of apolipoprotein E (APOE) due to its strong genetic association with AD. This will address the incomplete comprehension of APOE activity, and the arduous task of

[6] https://www.eudat.eu/services/b2safe
[7] https://www.eudat.eu/services/b2handle
[8] https://www.zenodo.org
[9] e.g. https://openneuro.org, https://www.ebrains.eu, https://mdrepo.org, https://mddbr.eu, https://www.dsimb.inserm.fr/ATLAS
[10] e.g., https://dspace.vsb.cz
[11] https://rdm.lab.lrz.de (prototype)
[12] https://gitlab.lrz.de, https://opencode.it4i.cz (and other instances depending on purpose)
[13] https://github.com
[14] https://huggingface.co
[15] https://wandb.ai
[16] https://modeldb.science

studying interactions between APOE and lipids. The research aims to develop a rapid and precise deep learning model for the blind docking of lipids and small molecules to the APOE protein. This will be achieved by employing accurate simulations and experimental data to train and evaluate the model. The anticipated outcomes include a high-throughput structure-based virtual screening pipeline and a deep learning model capable of accurately predicting protein-lipid interactions.

Purpose of the data generation or re-use

The proposed investigation will utilize comprehensive training datasets for machine learning models, focusing on molecular docking simulations. These datasets will encompass interactions between proteins and various molecules, including lipids, biomolecules, and small molecules such as potential drug candidates. The data will be sourced from established drug databases, protein structure repositories, and molecular dynamics (MD) simulation results. These diverse datasets will cover individual proteins, their dimeric and multimeric forms, and their interactions with other molecules. By incorporating this wide range of high-quality, relevant data, the aim is to develop robust and accurate machine learning models that can effectively predict and analyse protein-ligand interactions, potentially accelerating drug discovery processes and enhancing our understanding of molecular biology.

Type and Size of Data

The study will generate and utilize a range of input and output data formats for molecular docking and dynamics simulations, with a moderate overall dataset size to ensure efficient storage and analysis.

- Input Data:
  - SDF and MOL2 (compressed), SMILES, and InChIKey (text-based) for molecular structures.
  - PDB files for protein structures.
  - Libraries of ligand parameter files for molecular mechanics calculations.
- Output Data:
  - Compressed tar packages containing:
    - PDB files for protein-ligand complex geometries.
    - MOL2/SDF files for compound structures.
  - CSV files summarizing results, including calculated molecular properties, energies, predicted affinities, and timing metrics.
  - Trained machine-learning models

Protein-structure (input) data for a single binding site can range from 100 GB to 1 PB.
MD simulation output data will typically be organized in folders containing a mix of small files (<100 MB) and larger files (<5 GB), with an estimated total of 1000–2000 files. Thus, one simulation corresponds to about 25-150 GB in data. The total amount of produced MD data is estimated to be ~15 TB.

Origin of data and potential users

Protein-binding input is retrieved from open databases and APIs, for example:

- https://www.lipidmaps.org/: Open-access database of a large number of lipid structures.
- https://zinc20.docking.org/: Commercially available database for virtual screening.
- https://pubchem.ncbi.nlm.nih.gov/: Open chemistry database of molecules with a wide variety of properties.
- https://www.ebi.ac.uk/: EMBL-EBI open database of bioinformatics sources.

In addition, the molecular dynamics simulations and data from molecular docking will be generated as part of ongoing research. We will also assemble *in house* experimental data in our databases FireProt DB and SoluProtMut DB. These databases are currently under active development, with planed release in summer 2025. The manuscript describing the database will be submitted to the NAR Database Issue.
Potential users include protein engineers, synthetic biologists, ML practitioners and developers, researchers studying the molecular basis of NDs, biotechnology and biopharmaceutical companies.

2.2. RP2 - Expanding Systems Biology with Clinical Phenomenology of Alzheimer's Disease to Understand Time and Scale Coupling Using Generative AI and HPCQC

RP2 aims to develop advanced AI-driven models to address the complexity of neurodegenerative diseases (NDs) by integrating molecular, biochemical, and patient-level data over decades. Using multimodal datasets from over 40,000 neuroimaging and 400,000 genetic participants, the project will design generative AI tools capable of modelling relationships across diverse data, filling gaps in missing modalities, and predicting disease progression. These open-source tools, distributed via platforms like Clinica[17] are expected to significantly advance patient care by uncovering novel mechanisms for diagnostics and therapeutic interventions, while also delivering precision medicine tools to personalize treatment strategies and optimize the timing of interventions for each patient.

Purpose of the data generation or re-use

The input and generated datasets will facilitate the development of sophisticated generative AI models for understanding and predicting the progression of neurodegenerative diseases (NDs) across decades. By leveraging multimodal datasets from leading consortia, the project aims to integrate diverse data types such as neuroimaging, genetics, and clinical observations. These datasets will enable the creation of tools capable of modelling complex relationships across molecular, biochemical, and patient-level data, addressing gaps in missing modalities, and generating predictions for disease progression. The ultimate goal is to use these datasets to identify novel mechanisms for diagnostics and therapeutic interventions while developing precision medicine tools that can personalize treatment strategies and optimize intervention timing for individual patients. This approach will contribute significantly to advancing patient care and improving outcomes for those affected by neurodegenerative diseases.
To complement the multimodal data cited above, the studies will incorporate both publicly available and proprietary Electroencephalography (EEG) datasets. This EEG data will provide valuable insights into neurological activity and brain function, enhancing the overall understanding of neurodegenerative processes.

Type and Size of Data

The study will utilize diverse data types, including:

- Medical images as NIFTI files
- Socio-demographic information as TSV files
- Clinical data (diagnosis, scores) as TSV files
- Genotyping data as PLINK files
- Associated metadata as JSON files
- EEG data as HDF5, BIDS, MEF3, EDF and EDF plus formats.
- EEG annotations and labels might extend beyond HDF5 to CSV, TXT, and XML formats.

Furthermore, the total volume data that would be stored and processed will be about 1000 TB (including both raw and processed data). Volume-wise the majority of the data is from the UKBiobank. The data considered in this RP involves larger parts of non-public patient-related data, for which the suitability of target storage systems must be evaluated; in particular such data can normally not be stored on standard HPC-related filesystems.
The output of the RP consists in trained models, whose total size is not expected to exceed that of the input data.

Origin of data and potential users

The input data are from individuals who volunteered to participate in the above mentioned studies. All individuals provided informed consent and all studies have been approved by adequate ethics committees. There is no need for special certification but – except for open data portions – there is a need to have a way to handle access rights so that only the researchers who are authorised to use the data can access it.
Neuroimaging data with associated data (patient-characterising, sociological etc.) are envisaged to be primarily sourced from ADNI, AIBL, ARWIBO, NIFD, OASIS, and UKBiobank.

[17] https://www.clinica.run

Open EEG data come from Donders Repository, OpenNeuro, DataDryad, Temple University Hospital EEG Data Corpus, PhysioNet, iEEG.org DataBase and EEG Database for BCI Applications. Proprietary EEG data are sourced from Mayo Clinic, University of Melbourne and the Motol Hospital.
Machine Learning and AI model artifacts, including code, trained weights, and configuration files, will be primarily version-controlled and shared via dedicated GitHub and/or on-premises GitLab repositories, ensuring collaborative development and reproducible results. For AI models, particularly large language models or complex neural networks, we will leverage Hugging Face's model hub for streamlined storage and distribution of pre-trained models and fine-tuned checkpoints, facilitating wider accessibility and fostering community contributions. Additionally, we will utilize Weights & Biases (W&B) to track and visualize experiment metrics, model performance, and system metadata. W&B will serve as a centralized platform for monitoring training runs, comparing different model versions, and analysing results, providing crucial insights for model improvement and reproducibility. This integrated approach, combining Git, Hugging Face, and W&B, optimizes development, deployment, and analysis workflows.

2.3. RP3 - Development of Multiscale/Cross-Modal Patient/Deep-Learning Models of AD in the Hybrid Quantum Classical Computing Environments

RP3 intends to develop an integrated multi-scale knowledge platform to better understand AD. It seeks to bridge the gap between molecular-level interactions and emergent brain behaviours by creating a platform for exploring and visualizing data across different levels of brain organization. Utilizing a hybrid quantum-classical computational framework and cross-modal deep learning models, it will simulate brain system functions and their temporal sequences. By initially focusing on a hippocampal CA1 pyramidal cell model and establishing neuronal performance metrics influenced by APOE and clinical action potential data, the programme aims to conduct multi-scale simulations, quantify neuron performance, and identify potential markers and therapeutic targets for AD. The ultimate goal is to provide a knowledge platform, models, and tools for ethical, open-source distribution, enhancing our understanding of AD development and progression.

Purpose of the data generation or re-use

Firstly, data generation and reuse aim to support the development, validation, and refinement of mechanistic models that simulate the complex interactions underlying Alzheimer's disease (AD). The data will primarily consist of inputs and outputs associated with these models, enabling the verification of their validity and the calibration of simplified, data-driven models. Additionally, datasets will be generated to characterize the mechanistic models, ensuring they accurately represent the intricate dynamics of neuronal behaviour across electrical, chemical, and metabolic domains.
A second key aspect of this effort includes addressing the requirements for 3D modelling and visualization. These needs will depend on the granularity required for the proof-of-concept use case, which focuses on a single neuron and its internal processes. The generated data will facilitate detailed 3D representations, aiding in the exploration of spatial and temporal interactions within neurons. By reusing these datasets across various simulations and analyses, RP3 aims to enhance our understanding of AD pathogenesis while providing a robust foundation for future research and model development.

Type and Size of Data

Input data for this research project will consist of established elementary models that are to be integrated in the platform, as well as their associated datasets (for unitary model validation). It will also comprise a database of publications and datasets that will be used to train/estimate models‘ dynamics.
Output data of the project will comprise model parameters, large datasets (also for visualisation) from mechanistic models, and trained machine-learning models, as well as source codes / scripts / tools – and data – related to (and partially accessible by) the knowledge platform (e.g. visualisation).
Data for this programme are expected to reach a size of 500 TB in relatively small files.

Origin of data and potential users
The input data will be sourced from existing established models, as well as published data and datasets from established data portals and repositories. We anticipate potential users to be scientists coming from fields such as computational neuroscience, system biology, quantitative systems pharmacology, etc. Notably, all (mechanistic and data-driven) models, and their associated configuration files will be version-controlled and shared via dedicated GitHub and/or on-premises GitLab repositories, thereby ensuring collaborative development and reproducible results.

2.4. Data Related to Other CLARA Activities (Testbed, Education, etc.)

The comprehensive approach taken at RDM within CLARA includes the FAIR treatment of data related to infrastructure (e.g. infrastructure-as-code source codes for deployment and maintenance of testbed components), to the management of computations (e.g. workflow scripts and descriptions) and to training activities (e.g. educational slide decks). This in particular pertains to data significant and worthwhile to be published.

Purpose of the data generation or re-use
WP5 sets up a testbed for HPCQC computation and data management, leveraging experience from earlier projects such as LEXIS. Such efforts usually generate reproducible and useful installation recipes (e.g. in a infrastructure-as-code approach) and coded computational or management workflows, which can be made FAIR on SCM platforms and be re-used by other infrastructure providers or peer researchers using similar workflows. Apart from this, CLARA as an CoE can be expected to produce educational material (e.g. slide decks) and other research-management related materials useful for internal purposes and a more general public.

Type and Size of Data
Testbed and education-related data usually come as source codes (text files) or as office-document files (e.g. docx, pptx) with a low volume, typically in the sub-TB range (in total) even after a project of several years.

Origin of data and potential users
Most of these data will probably be related to the testbed activities of WP5; training and educational or methodical outputs may come from WP3 (capacity building) and WP4 (HR development), or also from WP6 (dissemination and communication).
Materials from WP5 will be interesting for academic or general IT-infrastructure providers and for scientific-computing users and consultants. WP3/4/6 material is potentially very useful for academia or also, e.g., for industrial/SME educational activities.

3. Approach to FAIR Data

3.1. Repositories and Publication Mechanisms

For safe long-term archival and backup, data can be transferred to archives of the participating computing centres, but also be published, e.g., on the repositories of the academic partners in CLARA, on domain-specific research data repositories, or on generalised platforms such as Zenodo. For source codes and model data, specific approaches are supported by platforms such as GitLab, GitHub, Hugging Face, Weights & Biases and ModelDB. More details and references for all these facilities have already been given in Section 1.2.

3.2. General Concept to Fulfil FAIR Criteria

We fulfil the domain-agnostic DataCite [6] metadata standard with our outputs (and intermediate data, where appropriate) as a minimum. This standard is the prerequisite for obtaining Digital Object Identifiers (DOIs), as technically enabled by RDM platforms, and allows for references to further metadata and basic provenance tracking by relating data products. The consortium is aware of consistent annotation needs. The assignment of basic metadata and unique, persistent identifiers (PIDs) is key to the fulfilment of the FAIR criteria [4].

Once basic findability or accessibility, metadata and PID requirements are fulfilled, interoperability and re-usability (the I and R in FAIR) depend crucially on the use of standard (meta-)data formats, which CLARA scientists are aware about – in particular also through the internal activities of Task 2.3. The fulfilment of findability and accessibility requirements in the RPs and in general – usually coming with a steep learning curve, is laid out in Section 3.3.

3.3. Findability and Accessibility via Metadata, PIDs and Publication Mechanisms

Published or shared data will be equipped with basic metadata generally fulfilling at least the DataCite least-common-denominator standard to facilitate the discovery of datasets in the RPs and in general. Recommendations on an optimum metadata-scheme usage [7] – for example on the usage of keywords – will be ensured to optimise the possibility for discovery and then potential re-use, as well as harvesting and indexing. Datasets will be unambiguously identified by a (PID), for example DOI, or B2HANDLE PID.

As DOIs in European research are usually registered at DataCite, findability is greatly enhanced for datasets with a DOI – without any further effort – through the DataCite Commons Portal[18]. Data within the LEXIS platform will in practice be findable via a FAIRification approach currently developed within the EXA4MIND project (possibly involving landing webpages for datasets, OAI-PMH interface and/or harvesting via B2FIND), or via a DOI where registered. For long-term access, data will be published in reliable repositories, in particular in Zenodo. Where appropriate, special domain repositories will be investigated. For such datasets, DOI minting will be supported by the repositories where the data is deposited. Datasets that are not covered through these possibilities can be equipped with a DOI by DOI-registration access of the participating computing centres.

We will provide data access through the centre-specific data-access mechanisms IT4I and LRZ offer[19], but also through the access mechanisms on the LEXIS Distributed Data Infrastructure [3], which can accommodate CLARA data. Data and code deposited on public general-purpose or domain-specific RDM or SCM platforms is usually directly accessible, indexed and shared via web access.

The table below details somewhat more on the plans on assigning metadata and PIDs (including DOIs) and providing access for output data from the specific RPs and the other CLARA activities. Open outputs are listed as well as embargoed/proprietary ones.

Data source	Nature of data and access modalities	Metadata and PID assignment strategy
RP1	MD-simulation output; predicted docking scores and metrics; input data collections (as far as interesting for re-usage); trained ML models: Important data can mostly be made open and web-downloadable (cf. right column), after possible justified embargo periods. Internal databases will will be used for mutagenesis data.	Metadata are stored as JSON, CSV, TOML, or XLSX files, mostly with the datasets and outputs. Significant MD simulations will be deposited to public databases, such as MD repo, The European Repository for Biosimulation Data, or The ATLAS database, provided their accessibility. ML input and output data will be deposited on Zenodo and/or provided as supplementary files to the corresponding publications, generating PIDs. Trained ML models will be published via appropriate channels (GitLab or GitHub repositories, Hugging Face, etc.), assigning PIDs or DOIs where appropriate.

[18] https://commons.datacite.org
[19] cf. https://docs.it4i.cz/storage/proj4-storage; https://doku.lrz.de/dss-how-globus-data-transfer-and-globus-sharing-for-dss-works-11484489.html

RP2	Trained AI models; training data collections (as far as interesting for re-usage): Access will be restricted by access-control (login) mechanisms for most data due to their sensitivity.	Metadata are stored as JSON, CSV or XLSX files, mostly with the datasets and outputs. Datacite-compliant subsets of these metadata can thus be extracted to obtain DOIs for significant datasets.
RP3	i) data related to trained AI models / developed mechanistic/simulation models (including generative, topological, and simulation frameworks) ii) multi-modal clinical biological data, such as nanopore-based and optical biosignal readouts; other -Omics level data, and biomarker data iii) derived visualization and simulation outputs (temporal pathway maps, resilience trajectories, AI-predicted molecular cascades) iv) visualisation and knowledge platform codes Implementation-critical open code parts will be put on established publicly-accessible SCM platforms (GitHub, on-premises GitLab, etc.). Notable open datasets will be made available e.g. on ModelDB or Zenodo, after a possible justified embargo period. Data under special protection (IPR, GDPR aspects) will be provided with appropriate access control (e.g. on-premises gitlab with appropriate authentication).	DataCite-compliant metadata are stored as JSON, CSV or XLSX files with datasets and outputs, or within the databases of publication portals. DOIs will be obtained for significant published outputs; PIDs are considered for other data (where commit hashes on SCM platforms can substitute PIDs to some degree). This can be realised through publication platforms (left column) or also through assignment of data to publications as supplementary files.
Other activities (testbed, educa-tion, ...)	Source codes for testbed component instantiation and maintenance: can be made public via SCM platforms as long as no security-relevant contents is present. Source codes for facilitating computational workflows and data processing: can be made public via SCM platform after potential embargo periods. Open training and education material: can be made publicly accessible via repositories.	Metadata/PIDs are assigned by a coupled SCM-RDM approach with metadata directly ingested in a RDM framework (GitHub-Zenodo, or combination of GitLab with institutional repository framework). Note that commit hashes on SCM platforms can substitute PIDs to some degree. Metadata/PIDs are assigned by a coupled SCM-RDM approach, with metadata directly ingested in a RDM framework (GitHub-Zenodo, or combination of GitLab with institutional repository framework). Note that commit hashes on SCM platforms can substitute PIDs to some degree. Metadata and PIDs are ingested/assigned when uploading the data into a repository (institutional repositories or Zenodo).

Table 1: Details on nature, accessibility and findability of output datasets from CLARA (with focus on metadata and PID assignment)

4. Security and Ethics Compliance

The computing centres participating in CLARA provide an optimum basis for secure data treatment depending on the application case. At IT4I, internal processes and regulations for safe work with information that prevent misuse of information, unauthorized changes and data loss will be followed for data management and their safe transfer within individual locations. IT4I@VSB holds the ISO 27001 certificate (ISO/IEC 27001:2013, ČSN ISO/IEC 27001:2014). The certificate was awarded for providing services to the national supercomputer infrastructure, solving computationally demanding problems, advanced data analysis and simulation, and processing big data. LRZ has been certified according to ISO 20000 and 27001 since 2019, ensuring consistent service management and quality as well as proper information security management. The core of this management is realised via an integrated "I/SMS" (Information Security Management System (ISMS) + Service Management System (SMS)). In 2022, BADW-LRZ's research unit – though usually not directly responsible for the operation of services – has been fully added to the scope of IT security management according to ISO 27001. At the Paris Brain Institute, rigorous internal processes and policies, as well as a state-of-the-art security management ensure secure data processing and transfer. These measures prevent data misuse, unauthorized modifications and data loss. Via this common background in the consortium, a choice of storage systems guaranteeing appropriate data safety and security will be warranted as a key to the sustainable success of the CLARA endeavour.
The CLARA partners participating in ethically critical studies, in particular those involving data from patients, have a long-standing experience in assessing related ethics requirements and will ensure an appropriate assessment of CLARA studies.

5. Details on Output Datasets

We have identified specific outputs that we can provide during the project for each use case. This is a preliminary plan that will be reviewed and updated two times as CLARA runs - as specified in the DoA.

5.1. RP1

From RP1s investigations on protein interaction, MD simulation data will be yielded as well as machine-learning models trained on the MD data, drug databases and protein structure repositories. The data, as listed below, are planned to be made public via the mechanisms discussed in Section 3 gradually with the release of publications.

Item	Description
Dataset ID	CLRP1-01
Name and Reference	CLARA-RP1: MD Simulation Data
Description	Molecular dynamics data for proteins associated with ND
Standards, Format, and Metadata	.csv, .xtc, .pdb, .txt, .nc; HDF5; metadata in DataCite-json/yaml or TOML format
Data Sharing (including License)	Public (CC-BY 4.0) after paper publication; accessible via mechanisms discussed in Sec. 3 and below.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived, as appropriate, at tape archives (e.g. LRZ, IT4I) and deposited at domain-specific repositories (MD repo, MDDB, ATLAS) according to their availability.

Table 2: RP1 - Output Dataset 1 - MD Data

Item	Description
Dataset ID	CLRP1-02
Name and Reference	CLARA-RP1: Docking Results
Description	Molecular docking results for APOE, produced by the trained ML models and in house docking pipelines
Standards, Format, and Metadata	.csv, .mol2, .sdf, .pdb, .mmcif; HDF5; metadata in DataCite-json/yaml format
Data Sharing (including License)	Public (CC-BY 4.0) after paper publication; accessible via mechanisms discussed in Sec. 3 and below.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived, as appropriate, at tape archives (e.g. LRZ, IT4I), published on Zenodo and/or as a supplement to corresponding papers (e.g., docking scores and ranking).

Table 3: RP1 - Output Dataset 2 – Docking Results

Item	Description
Dataset ID	CLRP1-03
Name and Reference	CLARA-RP1: ML Models
Description	Trained ML models for understanding the dynamics and aggregation of proteins involved in NDs and their interactions.
Standards, Format, and Metadata	pb, .pkl, .pmml, HDF5; metadata in DataCite-json/yaml format
Data Sharing (including License)	Public (CC-BY 4.0) after paper publication; accessible via mechanisms discussed in Sec. 3 and below.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived, as appropriate, at tape archives (e.g. LRZ, IT4I) and published on GitHub and/or Hugging Face.

Table 4: RP1 - Output Dataset 3 - ML Models

5.2. RP2

RP2 aims to design, train and validate large generative AI models for understanding the correlations between molecular, biochemical and patient-level (in particular neuroimaging and EEG) data on ND progression, with the goal of delivering precision medicine tools to personalize treatment strategies. Moreover, these models will also be able to replicate past and predict future observations for patients. Ethics and security aspects are relevant for input and output data. The output model data to be generated are described in the following table.

Item	Description
Dataset ID	CLRP2-01
Name and Reference	CLARA-RP2: ML Models
Description	Trained ML models for understanding dynamics of ND and their progression, depending on observables.
Standards, Format, and Metadata	.bin, .pt, .pb, .onnx, .safetensors, .pkl, .pmml, HDF5; metadata in DataCite-json/yaml format
Data Sharing (including License)	Depending on privacy concerns and exploitation strategy, only parts are excepted to be made public; those will be accessible

	via mechanisms discussed in Sec. 3 and below.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived at tape archives (e.g. LRZ, IT4I) where feasible and appropriate and published on Hugging Face.

Table 5: RP2 - Output Dataset 1 - ML Models

5.3. RP3

The multi-scale knowledge platform for a better understanding of AD, developed by RP3, will comprise models (mechanistic simulations, AI-based models) as well as visualisations (with the respective data) and source codes related to the platform.

Item	Description
Dataset ID	CLRP3-01
Name and Reference	CLARA-RP3: Model data
Description	Data related to trained AI and classical (simulation/mechanistic) models representing brain functions and their temporal sequence as part of the knowledge platform developed by RP3. Where appropriate, source codes are delivered with the data.
Standards, Format, and Metadata	.pb, .pkl, .pmml, HDF5, UTF-8 source code; metadata in DataCite-json/yaml format
Data Sharing (including License)	Public (CC-BY 4.0 for data; GPL, Apache 2.0 or similar for code) after possible justified embargo period via mechanisms discussed in Sec. 3; source codes via GitLab/GitHub SCM platforms.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived at tape archives (e.g. LRZ, IT4I) as feasible (regarding also the sensitivity of data); codes to be archived on SCM platforms and Zenodo.

Table 6: RP3 - Output Dataset 1 - Model data

Item	Description
Dataset ID	CLRP3-02
Name and Reference	CLARA-RP3: Multi-Modal Biological Data
Description	Multi-modal clinical biological data (e.g. cfDNA (5hmC), ATP flux, Ca^{2+} dynamics, transcriptomics), nanopore-based and optical biosignal readouts, other -Omics levell and biomarker data
Standards, Format, and Metadata	Various (de-facto-)standard file formats; metadata in DataCite-json/yaml format
Data Sharing (including License)	Mix of restricted (IPR, GDPR) and public data. Public (CC-BY 4.0) portions accessible via mechanisms discussed in Sec. 3.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived at tape archives (e.g. LRZ, IT4I) as feasible (regarding also the sensitivity of data).

Table 7: RP3 - Output Dataset 2 – Multi-Modal Biological Data

Item	Description
Dataset ID	CLRP3-03
Name and Reference	CLARA-RP3: Derived Visualisation Data and Simulation Outputs
Description	Visualisation data displayed within the knowledge platform (as far as static extraction makes sense), and derived simulation outputs (as far as not covered by CLRP3-01) – for example temporal pathway maps, resilience trajectories, AI-predicted molecular cascades.
Standards, Format, and Metadata	.mp4, .swc, .obj, .fbx; metadata in DataCite-json/yaml format
Data Sharing (including License)	Public (CC-BY 4.0) after possible justified embargo period; accessible via mechanisms discussed in Sec. 3.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived at tape archives (e.g. LRZ, IT4I).

Table 8: RP3 - Output Dataset 3 – Visualisation Data

Item	Description
Dataset ID	CLRP3-04
Name and Reference	CLARA-RP3: Visualisation and Platform Codes
Description	Codes from the visualisation efforts and presentation layer related to the knowledge platform.
Standards, Format, and Metadata	UTF-8 source code; metadata in DataCite-json/yaml format
Data Sharing (including License)	Public (GPL, Apache 2.0 or similar) after possible justified embargo period via mechanisms discussed in Sec. 3; source codes via GitLab/GitHub SCM platforms.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	Codes to be archived on SCM platforms and Zenodo.

Table 9: RP3 - Output Dataset 4 – Visualisation and Platform Codes

5.4. Other Outputs

WP5 will generate source codes (including infrastructure-as-code descriptions) and WPs 3, 4, and 6 will generate text and presentation material which can be useful, e.g., for education.

Item	Description
Dataset ID	CLOO-01
Name and Reference	CLARA-OtherOutputs: Testbed Codes
Description	Infrastructure-as-code codes (or also container creation recipes) relevant to the CLARA testbed creation and reproducibility.
Standards, Format, and Metadata	UTF-8 source code; DataCite metadata and readme files managed via SCM-RDM platform combination.
Data Sharing (including License)	Public (GPL, Apache 2.0 or similar for code) as far as useful and publishable (after security and IP assessment),

	managed/shared via GitLab/GitHub SCM platforms in combination with a RDM platform (e.g. Zenodo or institutional).
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived at tape archives (e.g. LRZ, IT4I) and SCM platforms.

Table 10: OtherOutputs - Output Dataset 1 - Testbed Codes

Item	Description
Dataset ID	CLOO-02
Name and Reference	CLARA-OtherOutputs: Workflow Descriptions
Description	Workflow-management and –description code from CLARA use cases.
Standards, Format, and Metadata	UTF-8 source code; DataCite metadata and readme files managed via SCM-RDM platform combination.
Data Sharing (including License)	Public (GPL, Apache 2.0 or similar for code) as far as useful and publishable (after security assessment and embargo/IP considerations), managed/shared via GitLab/GitHub SCM platforms in combination with a RDM platform (e.g. Zenodo or institutional).
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	To be archived at tape archives (e.g. LRZ, IT4I) and SCM platforms.

Table 11: OtherOutputs - Output Dataset 2 – Workflow Descriptions

Item	Description
Dataset ID	CLOO-03
Name and Reference	CLARA-Other Outputs: Material with Educational or Training Value
Description	Documents from WPs 3/4/6 useful for general public (e.g. due to educational value), for example training material.
Standards, Format, and Metadata	Typical Office-Software or graphics formats (e.g. .pptx, .docx, .odt, .png, .svg), annotated with DataCite metadata on submission to repository.
Data Sharing (including License)	CC-type license chosen as appropriate for the purpose where the data can be made open.
Are Datasets Accessible?	Not yet.
Is Dataset Reusable?	Not yet.
Archiving and Preservation (Storage and Backup)	Backup at tape archives (LRZ, IT4I) and preserved in repositories where published.

Table 12: OtherOutputs - Output Dataset 3 – Material with Educational or Training Value

Conclusion

In this Data Management Plan, we have summarised the awareness and planning of the consortium towards a FAIR data approach in the work with our research datasets and other data products such as infrastructure-related source

codes. A consistent work with basic metadata and persistent identifiers, as well as institutional, domain-based and public data and code publication frameworks/repositories is envisaged as described. The type and size of data to be used in all four application cases were discussed and a preliminary specification of outputs was delivered to lay out ambition and requirements.
This document will be continuously updated during the project as described in the DoA.

References

[1] Golasowski, Martin et al. (2022). “The LEXIS Platform for Distributed Workflow Execution and Data Management.” In: HPC, Big Data, and AI Convergence Towards Exascale. Boca Raton (FL): CRC Press, pp. 17–35. ISBN: 978-1-003-17666-4. DOI:10.1201/9781003176664-2.

[2] Hachinger, Stephan et al. (2022). “Leveraging High-Performance Computing and Cloud Computing with Unified Big-Data Workflows: The LEXIS Project.” In: Technologies and Applications for Big Data Value. Ed. by Edward Curry et al. Cham: Springer International Publishing, pp. 159–180. ISBN: 978-3-030-78307-5. DOI:10.1007/978-3-030-78307-5_8.

[3] Munke, Johannes et al. (2022). “Data System and Data Management in a Federation of HPC/Cloud Centers.” In: HPC, Big Data, and AI Convergence Towards Exascale. Boca Raton (FL): CRC Press, pp. 59–77. ISBN: 978-1-003-17666-4. DOI:10.1201/9781003176664-4.

[4] Wilkinson, Mark D. et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship.” Sci Data 3, 160018. DOI:10.1038/sdata.2016.18

[5] Xu, Hao et al. (2017). iRODS primer 2: Integrated Rule-Oriented Data System. Williston, VT: Morgan & Claypool Publishers. DOI:10.2200/S00760ED1V01Y201702ICR057

[6] Starr, Joan; Gastl, Angelika (2011). isCitedBy: A Metadata Scheme for DataCite. D-Lib Magazine, 17(1-2). DOI:10.1045/january2011-starr

[7] Bayer, Christiane et al. (2024). DataCite Best Practice Guide (Version 3.0). Zenodo. DOI:10.5281/zenodo.7099881