Project data management plan

MultiXscale Deliverable 8.2
Deliverable Type: Report
Delivered in June, 2023

MultiXscale
EuroHPC Centre of Excellence for
Multiscale Modelling

Acknowledgement
Funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) under grant agreement No 101093169.

Disclaimer
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European High Performance Computing Joint Undertaking (JU). Neither the European Union nor the granting authority can be held responsible for them.

Project and Deliverable Information

Project Title	MultiXscale: EuroHPC Centre of Excellence for Multiscale Modelling
Project Ref.	Grant Agreement 101093169
Project Website	https://www.multixscale.eu
EuroHPC Project Officer	Dr. Linda Gesenhues
Deliverable ID	D8.2
Deliverable Nature	Report
Dissemination Level	Public
Contractual Date of Delivery	Project Month 6(30th June, 2023)
Actual Date of Delivery	30th June, 2023
Description of Deliverable	Data management Plan (DMP) to outline how research data will be collected, generated and handled during and after the project.

Document Control Information

Document	Title:	Project data management plan
	ID:	D8.2
	Version:	As of June, 2023
	Status:	Accepted by Steering Committee
	Available at:	https://www.multixscale.eu/deliverables
	Document history:	Internal Project Management Link
Review	Review Status:	Reviewed
Authorship	Written by:	Neja Šamec(NIC)
	Contributors:	Alan Ó Cais (UB)
	Reviewed by:	Alan Ó Cais (UB)
	Approved by:	Neja Šamec (NIC)

Document Keywords

Keywords:	MultiXscale, Data management, licensing

30th June, 2023

Disclaimer:*This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements.*

[1] neja.samec@cmm.ki.si

Contents

List of Tables

Executive Summary

This is short summary of the contents of the deliverable with section referencing where appropriate for further details.

- *Data* (Section 2)
  The MultiXscale **project generates and manages various types of data** throughout its life cycle. These include mailing lists, internal meeting notes, images, website content, presentations, training materials, participant information, input data for workshops, video recordings of lessons, target audience information for course advertisement, and survey data. The project team follows best practices for data management, using appropriate formats and standards. Data is securely stored in backed-up locations and made available to researchers in compliance with relevant policies and regulations. Proper documentation, annotation, and labeling are implemented to facilitate data reuse and interpretation.

  To promote data accessibility and interoperability, the project prioritizes the use of open and non-proprietary formats.

  Various **data storage platforms** are utilized to meet diverse requirements and enable data sharing. Microsoft OneDrive is used for synchronous online editing of working documents, while services such as GitHub/GitLab serves as a repository for backup, versioning, and project monitoring. GitLab.com or GitHub.com are employed for collaborative design and development of source code and training materials, with each course having its own repository. Dissemination and communication channels are also utilized to share project information.

  The **Data Structure section** outlines the arrangement and management of data in the MultiXscale project. Guidelines for file naming and organization are established to ensure efficient data organization across platforms like GitHub and Microsoft OneDrive. Git version control is utilized for tracking changes and secure backup. The lesson folder structure in the GitHub repository is defined, including folders for training materials, survey data, survey code, survey results, and administration. The structure may evolve as needed. The MultiXscale website serves as an information hub, providing access to training materials, news updates, social media links, and partner information. State-of-the-art web technologies ensure compatibility across devices and browsers. Section **Metadata and documentation** section emphasizes the importance of facilitating data discovery, interpretation, and re-use. GitLab/GitHub repositories are used for version control and documentation, with each lesson having its own repository and a README file containing essential details. Metadata adhering to community standards is provided, including descriptive information, organization details, dates, description, format, and licensing information. Documentation efforts cover data collection methodologies, processing and analysis steps, and a comprehensive history of changes. Clear guidelines and project structure documentation are provided within the repositories for easy navigation.

- *Software* (Section 3)
  This section provides an overview of software management in the MultiXscale project. It is not foreseen that MultiXscale creates large software projects de novo, but rather contributes to existing software projects. MultiXscale is therefore bound by the licencing and contribution agreements of those software projects.

  The section outlines the **management of the code** generated and utilized within the project, including availability and considerations for long-term support. Additionally, it addresses the adherence to the **FAIR principles** for data publication.

  Regarding storage and backup, the project employs multiple platforms to enhance accessibility, collaboration, and data integrity. Metadata for materials is generated following platform guidelines, enhancing discoverability and accessibility for the scientific community.

  To ensure long-term accessibility, persistent identifiers such as DOIs are applied to entire lesson materials folders, including code, for each official version of the lessons. Platforms like **Zenodo** are utilized to assign DOIs, guaranteeing the longevity and accessibility of project resources.

- *Access and security* (Section 4)
  The section on access protocols and security measures emphasizes the **protection of sensitive data** in the MultiXscale project. Access to administrative and management data will be granted based on involvement and data requirements, utilizing platforms like GitHub and Microsoft OneDrive with appropriate permissions and access controls. Data security practices, including secure passwords, encryption, and secure communication channels, will be followed by all project partners and collaborators. Regular data backups will be performed, and any security incidents or breaches will be promptly reported. The training materials will be released under the Creative Commons Attribution license, promoting sharing and re-use while respecting intellectual property rights. These measures ensure the security, integrity, and responsible sharing of project data.

- *Legal aspects* (Section 5)
  The legal aspects section of the document covers two key areas: **handling of personal/sensitive data** and **data ownership/intellectual property**. For personal data collected from participants, the project ensures data protection by obtaining user consent, providing mechanisms for data subject requests, establishing retention periods, and anonymizing feedback data. Such personal data will be managed through the members.cecam.org platform with adherence to its terms and conditions. Regarding data ownership and intellectual property, training materials are shared among project partners. The training materials will be licensed under Creative Commons Attribution (CC-BY), allowing sharing and adaptation while ensuring proper attribution. The project's measures align with data protection regulations, respect intellectual property rights, and foster collaboration and knowledge sharing among project partners.

1 Introduction

1.1 Scope of the deliverable

This deliverable relates to the management of data objects gathered and handled during the project's implementation and can be categorized into four primary groups:

- Data - addresses data which will be used in the projects including how it will be managed and stored. It explains how the data lifecycle will be made compliant with the FAIR principles.
- Software - addresses software created by the project, and software tools used for generating, processing, and analyzing data in the MultiXscale project.
- Access and security - addresses access protocols and security measures in place to safeguard the project's data.
- Legal aspects - addresses the legal aspects and data management practices related to sensitive data usage and the measures taken to protect participants' data and ensure compliance with data protection regulations.

1.2 Target audience

The recommendations of this document are intended for members of the MultiXscale project who need to follow the requirements of the collected data.

2 Data

2.1 Type and format of data

Throughout the project life cycle, various data types and formats will be generated, carefully managed, and preserved. The following data will be produced:

- Mailing lists: Generated to facilitate communication with participants throughout the project.
- Internal meeting notes: Records of discussions held during consortia/event preparation meetings.
- Images and illustrations: Visuals utilized during workshops, training materials, course registration platform, website, and other dissemination and communication activities.
- Website content: Descriptions of the training program, participant testimonials, and other relevant information.
- Presentations delivered at conferences: Presentations showcasing the training program at conferences and events.
- Training materials: Presentations, documents and other materials used for delivering training and educational content to participants.
- Course participant information: Data including participant names, contact details, and demographic information used for course registration, participant selection, and impact measurement in MultiXscale reports.
- Input data used during workshops: Datasets employed in workshop practicals to exercise and test data analysis and processing skills.
- Survey data: Data collected from surveys designed to evaluate training program effectiveness and gather participant feedback.
- Software: software lines of code (LoC) generated by those working within the context of the project (further discussion in Section 3).

To ensure proper management and preservation, the project team will adhere to best practices for data management, utilizing appropriate data formats and standards. All data will be securely stored in backed-up locations and made available to researchers in accordance with relevant policies and regulations. The team will also ensure proper documentation, annotation, and labeling of the data to facilitate its reuse and interpretation by others. MultiXscale will prioritize the use of open and non-proprietary formats to promote data accessibility and interoperability. Please refer to the following table1 for a tabular description of the formats employed in the project.

Data Type	**Format(s) Used**
Course participant information	Spreadsheet (CSV), structured text (JSON, XML)
Mailing lists	Spreadsheet (CSV), email list (Mailing list software such as Mailchimp)
Information about entities for advertisement	Spreadsheet (CSV), structured text (JSON, XML)
Training materials	PDF, HTML, Markdown
Images and illustrations	SVG, PNG
Input data used during workshops	Spreadsheet (CSV), structured text (JSON, XML), other open formats specific to tools and software used (for biological sequences handling: FASTA, FASTQ, GFF/GTF, BED, BAM/SAM)
Internal meeting notes	Text files (OneDrive)
Survey data	Spreadsheet (CSV), online survey tools (SurveyMonkey, Google Forms)
Website content	HTML, CSS, JavaScript
Presentations delivered in conferences and meetings	PDF, HTML, Markdown, Google Slides
Software LoC	source code (C, C++, Python, bash,...)

Table 1: Type of data

The description of the Data Type “Input data used during workshop” will be expanded in following versions of this document. Compressed folders (e.g. ZIP or TAR.GZ) might be used, mainly for sharing purposes during the training, but an uncompressed version of the data will always be archived for backup.

2.2 Data storage platforms

To meet the diverse requirements throughout the MultiXscale project's lifespan and facilitate data sharing with different audiences, data storage platforms (Table 2) will be employed:

Microsoft OneDrive: This platform will store and enable synchronous online editing of working documents, that don't contain personal data of course participants. All project partners will have access to this shared drive, which is owned by the organization (XYZ), allowing the project management team to control and monitor the data as well as delete it if necessary. More details about this platform's usage can be found in the Document storage practices section.

GitHub.com: After live editing sessions, documents will be migrated to a Markdown-based format and transferred to a repository in GitHub.com for backup, versioning, and project monitoring purposes. Only project partners will have access to this repository. It also serves as a tool for project management, monitoring project progression through issues, tasks, boards, and milestones. Extensive information about this project management tool will be provided in the MultiXscale Dissemination and Communication Plan.

GitLab.com or GitHub.com: These platforms will be used for collaborative design and development of training materials or software. After a training course delivery, lesson materials will be deposited in a repository that will assign a digital object identifier (DOI) to them (e.g., Zenodo), enabling systematic and combined search based on deposited metadata.

members.cecam.org: This platform, managed by EPFL, will handle course registration, participant selection, and result notification. It will host all participant data, as well as most of the data related to each workshop instance (e.g., dates). Further information about this platform can be found in the Legal aspects section.

MultiXscale website: The project's website will contain relevant information and data. A comprehensive description can be found in the Website section, with additional details included in the Dissemination and Communication Plan.

It's important to note that information about the project will also be disseminated through various MultiXscale communication channels. Further information regarding the usage and practices of these channels will be provided in the MultiXscale Dissemination and Communication Plan.

Data Type	Platform(s) Used
Course participant information	members.cecam.org
Mailing lists	Mailing list software
Information about entities for advertisement	Mailing list software, Microsoft OneDrive(continuous update)
Training materials	GitLab/GitHub
Images and illustrations	members.cecam.org, Microsoft OneDrive, GitLab/GitHub, Website
Input data used during workshops	GitLab/GitHub
Internal meeting notes	Microsoft OneDrive (while editing), GitHub (for storage)
Survey data	GitHub
Website content	Website server
Presentations delivered in conferences and meetings	Microsoft OneDrive (while editing), GitHub (for storage)
Software LoC	GitLab/GitHub (or similar)

Table 2: Platforms used for storage

2.3 Reuse of existing data

This section lists existing data sets that will be used and specify the terms of use (e.g. licence, collaboration with the data producing group). When re-using public data, it provides links to the source.

2.3.1 Training materials

All the training materials in these projects will be openly available under permissive licences, such as Creative Commons, and can be accessed via their respective websites. Links to these materials will be provided in the MultiXscale's website and in any other relevant documentation. The version information and dates of the materials that are used will be tracked to ensure reproducibility and transparency of the work, as the materials are continuously evolving.

Additionally, all MultiXscale training materials will be versioned using GitLab/GitHub to keep track of changes and facilitate collaboration within the project team and with the wider community. The materials will be released under a Creative Commons Attribution (CC-BY) licence, which allows others to use, remix, and build upon MultiXscale's work, provided they give appropriate credit to the original authors.

2.4 Data Structure

This section outlines the arrangement and administration of data throughout the project. For managing data using a file system, it provides details on the directory structure, naming conventions, and quality control procedures for data content, structures, and conventions.

2.4.1 Practices for Storing Documents

To ensure efficient organization and management of data during the project, a set of guidelines for file naming has been established. These conventions are applicable to all platforms where data and documents will be stored, such as GitHub and Microsoft Onedrive. Each file will have a concise and meaningful short filename describing its content. Spaces in filenames will be replaced with underscores to maintain consistency. To ensure unique identification of each version, files will start with the creation date following the ISO 8601 standard (e.g., 2023-04-03), and a version number identifier may be added if necessary.

Utilizing GitLab/GitHub and Git version control will enable tracking of all data changes and secure backup of each version. When required, project documentation will include links to relevant documents, providing further information about data organization, including the directory structure and naming conventions.

2.4.2 Lesson Folder Structure

Building upon the previous section, a folder structure with clear names and descriptions will be implemented to organize training materials and project-generated data. Maintaining consistent content within these folders is crucial as they may be publicly shared. The following structure, including the specified sub-folders, will be used for all project lessons:

- **Training materials**: Contains all training program-related documents, including manuals, presentations, handouts, and scripts used by participants and instructors throughout the course. Project members and collaborators will have access to this folder. The e-learning versions of the training materials will be made public under the CC-BY license and accessible via the project website.
- **Survey data**: Stores anonymized data generated during the training program, such as attendance records, surveys, and evaluations.
- **Survey results**: Holds all results and outputs generated from the analysis of survey data. Access to this folder will be given to project members responsible for data analysis and project managers.
- **Admin**: Contains administrative and management documents specific to the course, such as notes from preparation meetings among MultiXscale partners.

This structure is subject to change throughout the project, and additional folders and subfolders may be introduced as necessary to maintain an organized and clear file structure.

2.4.3 Website

MultiXscale website will serve as an information hub for training materials and related events. The website is developed using state-of-the-art web technologies to ensure compatibility across various devices and browsers.

The website will feature a "download" (or similar) section where users can access software repositories, training materials and related resources. Additionally, a news section will keep visitors up to date with the latest developments and events. We will also include links to our project's social media accounts, providing regular updates and information about the training program. Furthermore, the MultiXscale website will provide insights into the project partners, showcasing their expertise and roles within the project.

2.5 Metadata and Documentation

This section provides an overview of the methods used to document and track data, as well as the process for accessing and linking relevant documentation to the corresponding data. Our documentation efforts will adhere to community metadata standards and encompass all necessary information to facilitate the discovery, interpretation, and re-use of both data and training content.

To effectively track and document our data, we will utilize GitLab/GitHub repositories for version control and documentation purposes. Within each repository, a README file will contain essential details about the software or lesson. Additionally, a contribution guide will be provided to enhance accessibility and ensure the quality of contributions.

2.5.1 Metadata

To ensure that the data are discoverable and interpretable, metadata that describes the contents of the software repositories and the training materials will be provided. The metadata will adhere to community standards such as Dublin Core or DataCite and include the following information:

- Descriptive title
- Name of the responsible organization
- Relevant dates
- Brief description encompassing the purpose, scope, and methods employed
- Format
- Licensing information

We will incorporate this metadata into the README files within each repository on GitHub. By doing so, users will have a clearer understanding of the training materials and can more easily discover, comprehend, and reuse them.

2.5.2 Documentation

In order to facilitate interpretation and re-usability, we will diligently document all relevant information pertaining to the software repositories and training materials. This documentation will be regularly updated whenever a new version of an official document is created, or a new software release is made. General guidelines on navigating the project folders, along with documentation of the project structure, will also be provided in the README files within each repository on GitHub and Microsoft Onedrive.

3 Software

One of the primary outputs of MultiXscale is software LoC. However, it is not foreseen that MultiXscale will create any software project de novo, but rather will make contributions to existing open source software projects. As such, MultiXscale is bound by the (existing or future) licensing and contribution agreements of the relevant software project.

In the MultiXscale project, Git, a version control system, is utilized to manage all code created by the project. This ensures efficient collaboration and version tracking among project partners. Each partner involved in software (or training) development stores their code in a designated code repository. These repositories are typically hosted via services such as GitHub/GitLab, fostering accessibility and collaboration. In addition, these services facilitate the implementation of code development best practices such as code review and Continuous Integration (CI) workflows.

To guarantee compliance with the requirements of FAIR software, the project management team supervises the code repositories to ensure adherence to the FAIR principles. Following best practices in the software development community, including regular updates, bug fixes, and comprehensive documentation, guarantees compatibility with long-term support.

The project team will also ensure that workflow created by the project comply with the FAIR principles and are assigned persistent identifiers, facilitating accessibility and long-term reference.

3.1 Storage and Back-up

This section outlines the storage locations for data and metadata, as well as the backup strategy implemented to ensure data integrity and accessibility.

To ensure accessibility and facilitate collaborative editing, the software, training materials, data, and metadata will be stored on multiple platforms throughout different phases of usage. The following storage platforms will be utilized:

- **Microsoft Onedrive**: The Onedrive platform will serve as a cloud storage and collaborative editing tool for documents such as meeting agendas, presentations, and other files requiring easy access and live editing. This platform ensures seamless collaboration and version control.
- **Git Repository**: The Git repository (hosted under services such as GitHub/GitLab) will store source code, training materials, data, metadata, and documents during the backup phase. It provides a secure and centralized location for storing these resources. Additionally, code and scripts will be maintained in GitHub, promoting accessibility and encouraging external contributions. This fosters collaboration and enhances the quality of the development cycle and the training program.
- **Internally Managed List Server**: Mailing lists will be managed through an internally managed list server, ensuring data security and providing project partners with access to all communications. This centralized approach facilitates effective communication and information sharing.

To ensure long-term accessibility, persistent identifiers such as DOIs will be applied source code and lesson material releases. Platforms like Zenodo will be utilized to assign DOIs, guaranteeing the longevity and accessibility of the resources. By implementing this storage strategy and utilizing various platforms, the MultiXscale project ensures data integrity, accessibility, and collaboration throughout its duration.

4 Access and Security

This section outlines the access protocols and security measures in place to safeguard the project's data. As the project involves handling a limited amount of sensitive information, including personal data of participants, the section highlights the necessary documents, agreements, and risk mitigation steps related to data access and sharing:

- **Access to administrative and management data** will be granted to project partners and external collaborators based on their involvement and data requirements. Data access will primarily be facilitated through the GitHub platform and the Microsoft Onedrive shared storage. To ensure data security and integrity, appropriate permissions and access controls will be implemented.
- All project partners and external collaborators will be required to adhere to appropriate **data security practices**, such as using secure passwords, employing encryption, and utilizing secure communication channels. **Regular data backups** will be performed to ensure data preservation and facilitate recovery in the event of data loss or corruption. Any security incidents or breaches will be promptly reported to the project officer and relevant authorities, as necessary.
- In line with the **Open Science model** embraced by the project, the training materials will be released under the Creative Commons Attribution (CC-BY) license. This license promotes sharing, re-use, and adaptation of the content while respecting intellectual property rights.

5 Legal ascpects

5.1 Handling of personal/sensitive data

For the MultiXscale project, participants will be required to provide personal data, specifically demographic information, during the course application process. However, sensitive data will not be necessary for participant selection. The management of personal data, excluding sensitive data, is described in this paragraph.

Personal and demographic data will be collected and stored through the members.cecam.org platform, provided by CECAM. All data will be hosted on their internal server. The platform is specifically designed to handle course registration, participant selection, and communication, offering the required functionalities for the MultiXscale project. Further details regarding the platform and it's usage are described in the CECAM Member platfom terms and conditions.

To ensure the protection of participants' data, the project will implement the following measures:

- Obtain user consent for processing personal data.
- Inform users about the existing mechanism to address data subject requests, such as requests for access, correction, or deletion of personal data.
- Establish a retention period for storing personal data, after which all personal data will be deleted.
- Anonymize course feedback data and demographic information utilized for evaluating the project's success in reports, safeguarding participants' privacy.

This document serves as an internal agreement among all project members who have access to personal data, and its terms are mutually agreed upon.

5.2 Data Ownership and Intellectual Property

This section clarifies the ownership of software LoC, training materials, addresses intellectual property matters, and explains the impact on intellectual property rights and their management.

- The project team will ensure that **all materials are appropriately licensed**. Software will be contributed under open source licences and according to the contribution agreement of the corresponding software project. Ownership of the training materials will be shared among the project partners. The training materials will be published under the **Creative Commons Attribution** (CC-BY) license, enabling sharing and adaptation of the materials while requiring proper attribution to the original creators.
- Any **intellectual property rights** associated with the data generated through the project will be shared among the project partners, without granting exclusive ownership to any single partner.

By implementing these measures, the MultiXscale project ensures compliance with data protection regulations, respects intellectual property rights, and promotes collaboration and knowledge sharing among project partners.

References

Acronyms Used

LoC	lines of code
CI	Continuous Integration

URLs referenced

Page ii

https://www.multixscale.eu . . . https://www.multixscale.eu
https://www.multixscale.eu/deliverables . . . https://www.multixscale.eu/deliverables
Internal Project Management Link . . . https://github.com/orgs/multixscale/projects/1
neja.samec@cmm.ki.si . . . mailto:neja.samec@cmm.ki.si
http://creativecommons.org/licenses/by/4.0 . . . http://creativecommons.org/licenses/by/4.0

Page 2

members.cecam.org . . . https://members.cecam.org

Page 10

members.cecam.org . . . https://members.cecam.org
CECAM . . . https://www.cecam.org/
CECAM Member platfom terms and conditions . . . https://members.cecam.org/terms-and-conditions