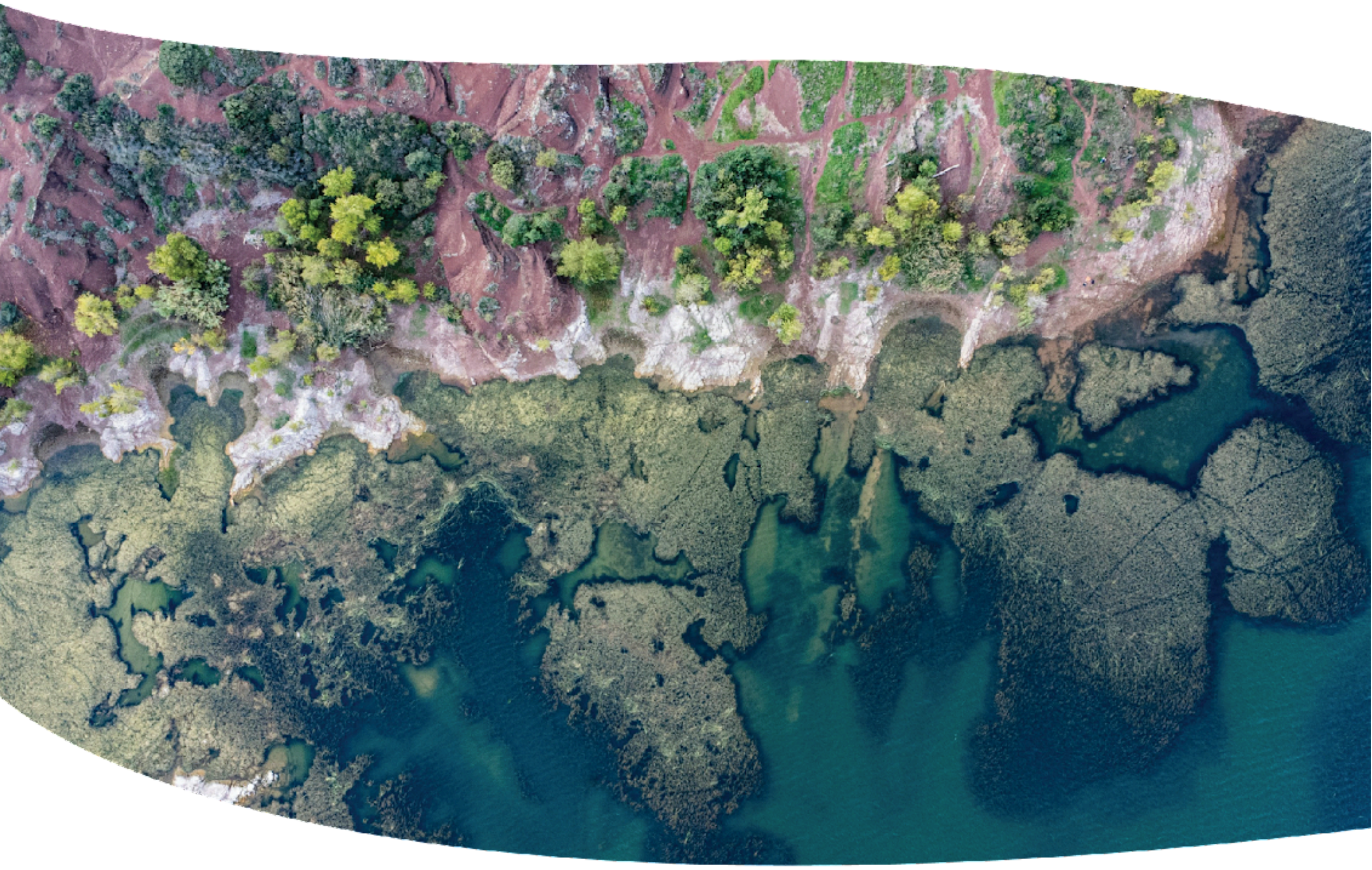
MICROBES 4CLIMATE



Deliverable D9.1

Data Management Plan and Processing of Personal Data

Date – 31-05-2024

Funded by the European Union

DOCUMENT CONTROL SHEET

PROJECT INFORMATION

<table>
<tr><td>Project Number</td><td colspan="3">10113181</td></tr>
<tr><td>Project Acronym</td><td colspan="3">MICROBES-4-CLIMATE</td></tr>
<tr><td>Project Full title</td><td colspan="3">Microbial services addressing climate change risks for biodiversity and for agricultural and forestry ecosystems: enabling curiosity-driven research and advancing frontier knowledge</td></tr>
<tr><td>Project Start Date</td><td colspan="3">1 February 2024</td></tr>
<tr><td>Project Duration</td><td colspan="3">60 months</td></tr>
<tr><td>Funding Instrument</td><td>Horizon Europe</td><td>Funding Scheme</td><td>HORIZON-RIA - HORIZON Research and Innovation Actions</td></tr>
<tr><td>Topic</td><td colspan="3">HORIZON-INFRA-2023-SERV-01-02 - Research infrastructure services advancing frontier knowledge</td></tr>
<tr><td>Coordinator</td><td colspan="3">MIRRI ERIC</td></tr>
</table>

DELIVERABLE INFORMATION

Deliverable No	D9.1
Deliverable Title	Data Management Plan and Processing of Personal Data
Work-Package No	9
Work-Package Title	Project Management and Ethics
WP-Leader (Name and Short Org. Name)	Microbial Resource Research Infrastructure – European Research Infrastructure Consortium (MIRRI-ERIC)
Task No	9.3
Task Title	Quality, ethics, data & risk management
Task Leader (Name and Short Org. Name)	Microbial Resource Research Infrastructure – European Research Infrastructure Consortium (MIRRI-ERIC)
Main Author (Name and Short Org. Name)	Ana Mellado (LifeWatch ERIC)
Other Authors (Name and Short Org. Name)	Ana-Claudia Sima (SIB), Marco Beccuti (UNITO), Matthew Ryan (CABI), Roland Pieruschka (FZJ), Tarcisio Mendes (SIB)
Reviewers (Name and Short Org. Name)	Ana Portugal Melo (MIRRI-ERIC), Aurora Zuzuarregui (UVEG), Cristina Varese (UNITO), Dimitris Hatzinikolaou (NKUA), Paula Marques (MIRRI-ERIC), Rosa Aznar (UVEG)
Status	Draft ☐ Final ☒
Deliverable Type	Report ☒ Data ☐ Demonstration ☐ Other ☐

Dissemination Level	Public (PU) ☒ Sensitive (SEN) ☐ Classified ☐ PU: Public, fully open SEN: Sensitive, limited under the conditions of the Grant Agreement Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No 2015/444 Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No 2015/444 Classified S-UE/EU-S – EU SECRET under the Commission Decision No 2015/444
Date Approved by Coordinator	31-05-2024

DOCUMENT VERSION HISTORY

Version	Date	Author	Description of Change
0.0	26.04.2024	Ana Mellado	Writing of the first draft
1.0	31.05.2024	Ana Mellado	Writing of the final version 1.0 of D9.1

DOCUMENT REVIEW

Reviewer	Date	Reviewer Name (Short Organisation Name)
Ana Portugal Melo	27.05.2024	MIRRI-ERIC
Paula Marques	27.05.2024	MIRRI-ERIC
Roland Pieruschka	27.05.2024	FZJ
Aurora Zuzuarregui	27.05.2024	UVEG
Cristina Varese	27.05.2024	UNITO
Rosa Aznar	27.05.2024	UVEG
Dimitris Hatzinikolaou	27.05.2024	NKUA
Matthew Ryan	28.05.2024	CABI

Legal disclaimer

The MICROBES-4-CLIMATE project is funded by the European Union under Grant Agreement no. 101131818. This document reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains.

Table of contents

Index of tables

About the project

Global changes are imposing substantial pressures on terrestrial biodiversity and ecosystems, presenting significant threats to agricultural and forestry systems. Among these changes, Climate Change stands out as a major concern, closely linked with biodiversity loss and ecosystem service degradation. Despite being a key life support system of the biosphere, microbes are often overlooked in the context of Climate Change, and our understanding of their responses to these changes remains limited. The impact of Climate Change on the assembly and functions of microbiomes, as well as their interactions with plants and soil, and their effects on plant performance and productivity, are still poorly understood.

MICROBES-4-CLIMATE (M4C) is a five-year Horizon Europe project that aims to fill these knowledge gaps by granting researchers across various fields efficient access to a diverse array of world-class Research Infrastructures across Europe and beyond. These infrastructures offer integrated, advanced services, along with training and scientific/technical support. At the heart of the M4C project lies an excellence-driven Transnational Access (TNA) Programme, empowering users to engage in curiosity-driven research to explore the complex interactions among microbiomes, plants, soil, and the environment in the context of Climate Change. The ultimate goal of the TNA programme is to encourage collaboration, drive innovation, and expedite scientific progress by facilitating researchers' access to and utilization of state-of-the-art resources.

With a consortium of 31 committed partners from 13 countries, M4C seeks to unravel the complex relationships among microorganisms, plants, soil and the environment, driving forward our understanding of ecosystem functioning and resilience. The project is set to generate an extensive array of data across its work packages and TNA grants, which will be relevant and beneficial to a diverse set of stakeholders involved in or affected by microorganisms and related disciplines. It is crucial that this data is managed effectively, adhering to best practices and in accordance with the requirements of Horizon Europe and European policies.

Executive Summary

The Data Management Plan (DMP) for the MICROBES-4-CLIMATE project outlines strategies and procedures for managing all types of data assets generated and reused throughout the project's lifecycle, ensuring adherence to Horizon Europe guidelines. It emphasizes data quality, accessibility, and long-term preservation, covering practices for data collection, storage, sharing, and preservation, while addressing legal, ethical, and security considerations. This initial DMP aims to provide comprehensive guidelines for preserving data and other research outputs collected or generated during the project's implementation by consortium members and outcomes produced by TNA grants. Effective data management practices, including accurate data recording, metadata documentation, and secure storage and dissemination of results, will be performed to ensure safe and reliable data preservation.

Given the diverse range of data expected over the project's five-year duration, the DMP incorporates various strategies for responsible data management and is regularly revised to accommodate newly generated data throughout the project's lifecycle.

Definitions

Term	Definition
Data asset	Refers to any identifiable piece of data or information, including various databases, documents, spreadsheets, multimedia files, etc.
Data Management Plan	The document outlines from the start of the project the main aspects of the lifecycle of research outputs, notably including data. This includes their provenance, organization and curation, as well as adequate provisions for their access, preservation, sharing, and eventual deletion, both during and after a project.
Digital data repository	A digital data repository is an established and reliable online platform or infrastructure that stores, manages, and provides access to research data in a secure, long term and sustainable manner. Ideally, such repositories adhere to recognized standards and best practices for data management, ensuring the integrity, authenticity, and long-term preservation of the stored data. Trusted repositories typically provide features such as persistent identifiers, version control, metadata standards, access control, and data curation services. They play a crucial role in facilitating data sharing, reuse, and reproducibility, thereby advancing research, innovation, and collaboration within the scientific community.
Metadata	Descriptive information accompanying the data assets, including details on data collection methods, equipment used, sampling protocols, and any associated contextual information.
Open Science	Open science is an approach based on open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the research process.
Research Data Management	The process within the research lifecycle that includes the organization, storage, preservation, security, quality assurance, allocation of persistent identifiers (PIDs) and rules and procedures for sharing of data including licensing

Research Data (H2020)	Research data is information (particularly facts or numbers) collected to be examined and considered, and to serve as a basis for reasoning, discussion or calculation.
Research outputs	Results to which access can be given in the form of scientific publications, data or other engineered results and processes such as software, algorithms, protocols, models, and workflows.

Abbreviations

Abbreviation	Definition
CBD	Convention on Biological Diversity
DMP	Data Management Plan
DOI	Digital Object Identifier. This is a unique identifier managed by the International DOI Foundation (IDF), a non-profit membership organization responsible for governing and managing the federation of Registration Agencies that offer DOI services and registration. The IDF serves as the registration authority for the ISO standard (ISO 26324) governing the DOI system.
DSI	Digital Sequence Information
FAIR	Findable, Accessible, Interoperable, Reusable (FAIR). This acronym delineates a collection of characteristics that scientific data assets should possess to facilitate reproducible research and maintain relevance within the research community.
GDPR	General Data Protection Regulation. This is the EU law that establishes the legal framework for personal data management
IPR	Intellectual Property Right
M4C	Abbreviation of the present project’s title “Microbes-4-Climate”.
OR	Open Repository
RDM	Research Data Management
RDF	Resource Description Framework
TNA	Trans-National Access involves granting researchers from foreign institutions access to an organization's research facilities and services.
SME	Small and medium-sized enterprises
WP	Work Package

1. Purpose of the data management plan

The purpose of the M4C Data Management Plan is to establish best practices for managing research outputs throughout the project lifecycle. The DMP covers various aspects such as data provenance, organization, curation, access, preservation, sharing, and eventual deletion. It is integral to the research methodology and contributes to efficiency, time savings, information security, and data impact. Research data management involves multiple stakeholders, including data creators, database designers, data managers, project managers, and IT staff. This interdisciplinary field ensures effective handling of research data, maximizing its value and impact.

The DMP is an adaptable document that evolves alongside the project. This initial version of the DMP will be continually refined and updated throughout the project's duration; it is foreseen to be updated by Month 24, 36 and 48 to ensure its relevance over time. Revisions will be driven by project milestones, feedback, consortium reviews, external insights, and advancements in technology and best practices. To facilitate these updates, we have implemented a comprehensive change management framework outlined in the next section of the present document.

The data assets generated in the context of the M4C project will be published in domain-specific or domain-generic trusted digital data repositories, employing standardized data formats and vocabularies, depending on data type. The metadata will be centralized within a dedicated M4C Information Platform that is presently under development. The Platform will be implemented using the open-source Harvard Dataverse. It will utilize a standardized metadata schema, guarantee persistent identifiers, offer open access to metadata (either through public or restricted access policies) and enable machine-readable metadata

While the M4C consortium will not directly manage all data assets, this document sets forth guidelines and best practices on how individual project partners and TNAs applicants should generate, acquire, handle, share and curate data during all activities carried out within and beyond the M4C project.

The Data Management Plan will be publicly accessible to ensure that all consortium members are informed about the practices to be adhered to during these activities. TNA applicants together with the service provider are required to review this document and adhere to the policies outlined herein when planning their data management activities.

1.1 Supporting policy

The present deliverable, D9.1 – Data Management Plan and Processing of Personal Data, serves as a comprehensive guideline adhering to both European and national regulations on the acquisition, handling, processing, and archiving of data generated throughout the M4C project and beyond its completion. This chapter reviews relevant European policy concerning open access and data management. This helps to better appreciate the context regarding the development of the M4C DMP.

1.1.1 European Commission - Open Science Policy

The 2018 Recommendation of the European Commission on access to and preservation of scientific information (2018/790/EU) [1] which builds on and replaces the 2012 Recommendation (2012/417/EU) [2], outlines a number of policy recommendations to support open access to research results. Open access policies aim to provide researchers and the public at large with access to peer-reviewed scientific publications, research data and other research outputs free of charge in an open and non-discriminatory manner as early as possible in the dissemination process, and enable the use and re-use of scientific research results. Open access helps enhance quality, reduce the need for unnecessary duplication of research, speed up scientific progress, help to combat scientific fraud, and can contribute to economic growth and innovation. Any licensing decisions should be aimed at facilitating the dissemination and re-use of scientific publications. Helping researchers to adequately manage research outputs is becoming a standard scientific practice [1]. The 2018 recommendations are broadly broken down into a number of measures as outlined in Table 1 [1]. These measures are useful to give M4C context and guidance when developing and implementing the project's DMP.

Table 1. European Commission's open science policy measures

Measure	Summary
Open access to scientific publications	Policies for the dissemination of and open access to scientific publications resulting from publicly funded research.
Management of research data, including open-access	Policies for the management of research data resulting from publicly funded research, including open access.
Preservation and re-use of scientific information	Policies for reinforcing the preservation and re-use of scientific information (publications, datasets and other research outputs).
Infrastructures for open science	Policies for further developing infrastructures underpinning the system for access to, preservation, sharing and re-use of scientific information and for promoting their federation within the European Open Science Cloud (EOSC).
Skills and competences	Policies for the necessary skills and competencies of researchers and personnel of academic institutions regarding scientific information.
Incentives and rewards	Policies for adjusting, with regards to scientific information, the recruitment and career evaluation system for researchers, the evaluation system for awarding research grants to researchers, and the evaluation systems for research performing institutions.
Multi-stakeholder dialogue on open science at national, European and international level	Multi-stakeholder dialogues on the transition towards open science at national, European and international level.
Structured coordination	National point of reference expert group to support coordination among EU members.

1.1.2 Horizon Europe - Open Science

The open science policy of the European Commission is integrated into the Horizon Europe 2021–2027 funding program for research and innovation. As described in the Horizon Europe Programme Guide, open science is characterized as collaborative and systematic sharing of knowledge and tools at the earliest stages and to the widest audience possible. This approach has the potential to enhance research quality and efficiency, speeding up the progress of knowledge and innovation through result sharing, enhancing reusability, and improving reproducibility. It requires the participation of all relevant stakeholders. Horizon Europe includes both mandatory and recommended open science practices, which may vary depending on specific work programs or call conditions. The typical mandatory open science practices are detailed in the following table:

Table 2 . Mandatory open science practices (Horizon Europe)

	Open science practices
1	Open access to scientific publications under the conditions required by the grant agreement.
2	Responsible management of research data in line with the FAIR principles of “Findability”, “Accessibility”, “Interoperability” and “Reusability”, notably through the generalized use of data management plans, and open access to research data under the principle “as open as possible, as closed as necessary”, under the conditions required by the grant agreement.
3	Information about the research outputs/tools/instruments needed to validate the conclusions of scientific publications or to validate/re-use research data.
4	Digital or physical access to the results is needed to validate the conclusions of scientific publications, unless exceptions apply.
5	In cases of public emergency, if requested by the granting authority, immediate open access to all research outputs under open licenses or, if exceptions apply, access under fair and reasonable conditions to legal entities that need the research outputs to address the public emergency.

Recommended open science practices extend beyond the mandatory ones. Examples include engaging all relevant knowledge actors, including citizens, early and open sharing of research, managing outputs beyond research data, implementing open peer review, etc.

The Horizon Europe Programme Guide [3] also defines research data management (RDM) as the series of activities within the research lifecycle, including data collection or acquisition, organization, curation, storage, (long-term) preservation, security, quality assurance, assignment of persistent identifiers (PIDs), provision of metadata to meet disciplinary standards, licensing, and rules and procedures for data sharing. RDM is essential for any project that generates, collects, or reuses data, requiring provisions to ensure responsible management (e.g., selecting the appropriate repository, ensuring adequate access provisions, complying with legal regulations such as the General Data Protection Regulation (GDPR), etc.). Additionally, data management must align with the FAIR principles, enabling researchers to easily Find, Access, Interoperate and Reuse each other’s data, thereby improving research effectiveness and reproducibility.

Data management plans are pivotal for responsibly managing research outputs. They play a key role in helping researchers to adequately manage a wide range of research outputs in line with the FAIR principles. Beyond publications and data, the range of research outputs may also be physical or digital, and include original software created during the project, workflows, protocols, and new materials such as samples, and cell lines, among many others. A DMP should be a living document, which is updated as the project evolves. A DMP should include:

Table 3. Data management plan aspects (Horizon Europe)

DMP aspect	Summary
Dataset description	Detailed description of the data generated or re-used.
Standards and metadata	The protocols and standards used to structure the data.
Name and persistent identifier for the datasets	A unique and persistent identification (an identifier) of the datasets and a stable resolvable link to where the datasets can be directly accessed.
Infrastructures for open science	Policies for further developing infrastructures underpinning the system for access to, preservation, sharing and re-use of scientific information and for promoting their federation within the European Open Science Cloud (EOSC).
Curation and preservation methodology	Information on the standards that will be used to ensure the integrity of the datasets and the period during which they will be maintained, as well as how they will be preserved and kept accessible in the longer term.
Data sharing methodology	Information on how the datasets can be accessed, including the terms-of-use or the license under which they can be accessed and re-used, and information on any restrictions that may apply or relevant security and privacy considerations.

These Horizon Europe open access and data management policies will be implemented in the M4C project.

1.1.3 GDPR

The General Data Protection Regulation (GDPR) is a European Union law requiring organizations to safeguard personal data and uphold the privacy rights of anyone in the EU. Personal data is defined as any information that relates to an individual who can be directly or indirectly identified (e.g. names, email addresses, location, ethnicity, gender, etc.). The regulation includes seven principles of protection that must be implemented and a series of privacy rights that must be facilitated. The seven data protection principles are [4]:

Table 4. GDPR - data protection principles

Data protection principle	Summary
Lawfulness, fairness and transparency	Processing must be lawful, fair, and transparent to the data subject.
Purpose limitation	You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.
Data minimization	You should collect and process only as much data as absolutely necessary for the purposes specified.
Accuracy	You must keep personal data accurate and up to date
Storage limitation	You may only store personally identifying data for as long as necessary for the specified purpose.
Integrity and confidentiality	Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. by using encryption).
Accountability	The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.

GDPR also recognizes a series of privacy rights, including:

- The right to be informed
- The right to access
- The right to erasure
- The right to restrict processing
- The right to data portability
- The right to object
- Rights in relation to automated decision making and profiling

The GDPR regulation must apply to any personal data managed by M4C.

1.1.4 Open data directive

The European Union Directive on open data and the re-use of public sector information provides common rules for government-held data at national, regional and local levels. It aims to overcome barriers to fully exploit the potential of public sector information for the European economy and society. The Open Data Directive implementation is ongoing and requires the adoption by the European Commission, via a future implementing act, with a list of high-value datasets taking into account guidelines on recommended standard licenses, datasets and charging for the reuse of information resources. To achieve maximum impact and to facilitate re-use, the high-value datasets should be made available for reuse with minimal legal restrictions, free of charge, and published via Application Programming interfaces (APIs). However, the Directive does not prevent public sector bodies from charging no more than marginal costs (e.g. reproduction costs) for the reuse of their data. However, particular charging exceptions are allowed for public sector bodies that are required to generate revenue to cover a substantial part of their costs; libraries, including university libraries, museums and archives; and public undertakings. Focus areas include [5]:

Table 5. Focus areas of the Open Data Directive

Open data focus areas
Provision of real-time access to dynamic data
Increasing the supply of public data for re-use, including public undertakings, research performing organizations and research funding organizations
Tackling the emergence of new forms of exclusive arrangements and the use of exceptions to the principle of charging at marginal cost
Outline the relationship between this Directive and other Directives such as: ● GDPR: align with GDPR personal data protection measures ● Database: public bodies should not be exercising data ownership protection measures to prevent data re-use

In line with the European Commission's Open Science Policy, the Directive describes obligations regarding re-use of publicly funded research data:

"*The Commission Recommendation of 25 April 2018 on access to and preservation of scientific information describes, among other things, relevant elements of open access policies. Additionally, the conditions, under which certain research data can be reused, should be improved. For that reason, certain obligations stemming from this Directive should be extended to research data resulting from scientific research activities subsidised by public funding or co-funded by public and private-sector entities. Under the national open access policies, publicly funded research data should be made open as the default option. However, in this context, concerns in relation to privacy, protection of personal data, confidentiality, national security, legitimate commercial interests, such as trade secrets, and to intellectual property rights of third parties should be duly taken into account, according to the principle 'as open as possible, as closed as necessary'.*"

1.2 DMP change management

M4C recognizes the potential need to adjust its data management policies and IT requirements over the project's duration. This is due to the dynamic nature of regulations and best practices, as well as the rapidly evolving IT service environments. To facilitate effective change management, a committee composed of representatives from WP4 (Transnational Virtual Access Programme), WP5 (Data Access and Management), WP6 (Legal, regulatory, ethical, and intellectual property affairs), and WP9 (Project Management and Ethics) will oversee three revisions of the DMP over the course of the project.

The process for DMP revisions involves several steps:

- Identification of change requests: The DMP committee will identify new data management requirements imposed externally or emerging internally, evaluating their urgency and feasibility for implementation within the scheduled DMP update.
- DMP change specification: For each change request, the committee will design appropriate procedures and draft a candidate revision to the DMP.
- Pilot phase: The revised DMP procedures will be tested on a voluntary basis with a data item impacted by the change. WP9 will oversee and coordinate this pilot phase.
- Change implementation: Upon successful completion of the pilot, the DMP will be updated, and the new management procedures will be implemented.

External requirement providers such as the EU, the ESFRI community, the EOSC association, and projects funded under the INFRA-EOSC initiative may influence new data management requirements. The three foreseen DMP revisions are anticipated to take effect by month 24, 36 and 48, respectively.

Given the relevance of Digital Sequence Information in the M4C project, The DMP will be updated to reflect any changes proposed by the Conference of the Parties to the CBD with respect to the use of DSI and monetary benefit sharing mechanisms thereof.

In the rare event of disruptive external change requests requiring an immediate response, the committee reserves the right to conduct extraordinary DMP revisions, which may bypass the pilot phase and become immediately effective. In such cases, the consortium will be promptly informed of the ongoing revision activities.

2. Data summary

The objective of this chapter is to provide a summary overview of the different types of data and information being used in M4C, including existing data sources and new data being generated by the project. During its operational phase, M4C will deal with a variety of data types. Some of these data will be directly generated and managed by M4C WPs, while others will be independently managed by TNA projects. In either scenario, for the sake of uniform data management, each distinct data asset should be categorized according to the descriptions outlined in this chapter and will be carefully managed to ensure its quality, integrity, accessibility, and (long-term) preservation. This chapter will be updated and enriched as the implementation of the project progresses.

2.1 Data types and formats

The M4C project will gather data on a wide variety of parameters, including genetic information, functional traits, phylogenetics, growth patterns, and more, related to plants, fungi, viruses, yeasts, and other (micro)organisms, as well as a range of environmental data. M4C will also generate other types of outputs, such as models, algorithms, workflows, training materials and guidelines, among others. Table 6 outlines various types of resources anticipated to be generated as part of the M4C project.

Regarding the selection of file formats, it is crucial to publish data in open and interoperable formats whenever feasible. This promotes enhanced data reusability and eliminates dependency on proprietary software licenses for data reuse. Typically, the original 'raw' file format or a lossless format derived directly from it is most suitable for long-term storage. Table 6 presents a variety of recommended file formats. Other formats can be used while processing/analyzing the data, but it is encouraged to convert to a recommended file format before archiving/publishing the data. Ideally, file formats should be non-proprietary (open), unencrypted, uncompressed, and commonly used by the research community. If it's not possible to use one of these file formats, the software name, version, license, and company should be documented in the metadata. M4C anticipates generating or reusing data ranging from gigabytes to terabytes in size. M4C will revise and update the data types and recommended formats listed in Table 6 throughout the project duration.

Table 6. Data types and formats

Data Type	Data format
Documents	● PDF (.pdf) ● OpenDocument Format for text documents (.odt and .fodt) ● Text file (.txt) ● Microsoft Word (.doc)
Presentations	● OpenDocument Format for presentations (.odp and .fops) ● Microsoft PowerPoint (.pptx)
Tabular	● Comma-separated values (CSV) (.csv) ● OpenDocument Format for spreadsheets (.ods and .fods) ● Microsoft Excel (i.e., .xls, .xlsx) ● Relational data (e.g., sql)
Images	● PNG (.png) ● JPEG (.jpg) ● GIF (.gif) ● TIFF (.tiff) ● Hyperspectral images (e.g., BIP, BSQ, BIL) ● X-ray Computed Tomography (slice images or 3D volumes; tif, .vol, .raw)
Videos	● MPEG-4 (.mp4)
Audio	● MP3 (.mp3) ● FLAC (.flac) ● WAV (.wav)
Spatial	● Shapefiles for vector files (.shp, etc.) ● GeoTIFF for raster files (.tif, etc.)

Sequences	● FASTQ, FASTA / FAS,
Code	● Plain text (usually with an extension that represents the source language)
Other	● Compressed files (e.g., .zip, .gz) ● Hierarchical data such as those based on XML (e.g., RSML to store 2D or 3D image metadata) ● Resource Description Framework (RDF) for metadata

2.2 TNA Grant Applications

TNA grant applications are essential to the M4C project, serving as a necessary preliminary step toward establishing a grant agreement between the service provider and the applicant (service user). Each grant application provides a detailed description of a transnational research activity (TNA) that may occur during the project's duration, along with its anticipated outcomes. The contents of these applications are considered confidential between the applicant and the service provider and should not be disclosed under any circumstances, except with explicit agreement between the applicant and M4C. Grant applications are managed within the IT Platform that M4C is currently developing, and adhere to the ethical principles and policies outlined in this document.

2.3 Data overview

All data assets used in the M4C project, whether pre-existing or generated during the project, will be thoroughly documented using a metadata file template (Annex 1) that adheres to the meta-information structure outlined in Table 7. Each data item considered within the project will be accompanied by its respective completed metadata file, which will be appended to the DMP deliverable (D9.1). Updated versions of these files will be provided throughout the project's duration. The metadata file comprehensively outlines descriptive information about the corresponding data asset, ensuring a detailed data report aligned with the guidelines outlined in the Horizon Europe DMP template (HE, 2022) [6].

2.4 Data reuse

Extensive review of archive data and data available in trusted repositories will be conducted periodically. M4C expects to reuse data from a variety of sources, including long-term monitoring, biodiversity databases, DNA databases, etc. A list of potential sources of public data which will be used in the M4C project will be provided in future versions of the DMP.

2.5 Data utility

Data used and produced in the M4C project will serve to gain a deeper understanding of the impact of climate change on the assembly and functions of microbiomes, as well as their interactions with plants and soil, and their effects on plant performance and productivity. M4C's outcomes will cater to individual researchers and research organizations across microbiology, plant soil and agricultural sciences, forestry, ecology, biodiversity, climate change, and related disciplines. These will also be useful for companies active in the biotechnology, agricultural/farming, forestry and environmental services sectors, among others. The project results will enhance awareness and promote evidence-based decisions for policy makers in areas linked with climate change risks for biodiversity and agricultural and forestry ecosystems, and will be used to actively engage, educate, and inform the broader public on climate resilience and adaptation, biodiversity conservation, and the promotion of sustainable agriculture among citizens.

3. FAIR data

The FAIR principles of “Findability”, “Accessibility”, “Interoperability” and “Reusability” are at the core of the M4C data management plan, and are discussed in this chapter.

3.1 Making data findable

All data assets generated within M4C must be associated upon publication with a persistent identifier (DOI or another globally unique and persistent identifier) assigned by a trusted digital data repository (e.g GenBank, Zenodo, GBIF, etc.) for effective and persistent citation. This persistent identifier can be used in any relevant publication to direct readers to the underlying data.

In order to make data findable, we need metadata (data about data) to help discover that data. Metadata contains information that describes the basic characteristics of each data asset, helping users to evaluate it and ultimately use it for their needs. Metadata must adhere to standardized formats consistent with community standards. It must be machine-readable, which enables computer processing techniques such as metadata harvesting, indexing, and the ability to cross-link between different research outputs. The structure of the minimum meta-information that will be recorded for each data asset generated within M4C is shown in Table 7. This aligns with the guidelines outlined in the Horizon Europe DMP template (HE, 2022) [6], and encompasses the minimum set of metadata elements as per the DataCite Metadata Schema standards [7]. It includes search keywords aimed at maximizing opportunities for reuse. Other standard metadata formats and vocabularies will also be used (e.g., MIAPPE) and will be updated in future versions of the DMP. Moreover, metadata about the different data assets in the project will be described and structured using vocabularies based on the Resource Description Framework (RDF) such as schema.org and Data Catalog Vocabulary (DCAT).

For all data assets a metadata record will be created in the M4C Dataverse repository, linking to the other repositories where the data can be found. The research outputs generated within M4C are owned by the partner/TNA grant beneficiary that produces it and should be acknowledged when others use the data, along with acknowledging M4C. Both, M4C partners and TNA applicants, must be aware that their published data assets may be included in a number of catalogs, such as the EOSC marketplace, ENVRI-Hub or Research Infrastructures data portals, among others.

Table 7. Structure of the minimum meta-information record for each data asset generated within M4C

DATA DENOMINATION	
Data reference name *	The reference name will start with the prefix “DA_M4C” indicating it’s a data asset
Data title *	The title of the data asset which should be easily searchable and findable
Description	A brief description of the data asset
Keywords	The keywords associated with the data asset
Version Number *	To keep track of changes to the data assets
DATA ORIGIN	
Name(s) of data creator(s) *	Authors responsible for the creation of the data asset, and their affiliation
Data Source	How/why was the data asset generated/re-used
Creation Date *	Date of generation of the data asset
Quality Assurance	A brief description of the quality assurance processes the data has gone through
DATA SPECIFICATIONS	
Type *	Type of data contained in the data set
Format *	This could be CSV, PCAP, HTML, etc.
Expected Size *	The approximate size of the data set
DATA ACCESSIBILITY	
Data Location *	The repository where the data are stored: partner owned, M4C Information Platform, or other trusted data repository
Date of Repository Submission *	The date of submission to the repository can be added once it has been submitted
Persistent Digital Identifier *	The PDI can be entered once the data set has been deposited in the trusted repository
Access Status *	Whether data is “open”, “consortium” or “restricted”
Embargo Period *	Duration of embargo foreseen for a data set
Funding statement	A disclaimer indicating that the specific data set was generated within the M4C project that received funding from the EU
DATA UTILITY	
Significance to the project *	The associated work package or TNA grant this data originates from
Significance outside the project	To whom the data might be useful outside the project
DATA PUBLICATIONS	
Related Publications	Bibliographical details of publications based on the data set
Data Citation	A ‘ready-to-use’ citation reference for the data asset

3.2 Making data accessible

The research data created in M4C is owned by the partner / TNA grant beneficiary that generates it, who is responsible for data dissemination unless there is a legitimate interest to protect it (e.g SMEs). The data will be made accessible following the accessibility criteria:

- Restricted data: accessible by the owner only, deposited in the repository of the owner partner or TNA grant beneficiary. Data must be kept for at least 10 years after the project ends. This applies only to SMEs. All other applicants should be eligible to publish the data; otherwise, they cannot apply for TNA.

- Open data: accessible by the public, deposited in trusted data repositories. Data is kept for the lifetime of the selected trusted data repository.

In cases where it is considered necessary, data owners may request embargo periods of up to two years for specific data assets. This allows time for publication or seeking intellectual property protection. During this period, the data will be stored in the trusted repository with an embargo date specified, after which the data will be made accessible. For data assets that cannot be shared, or need to be shared under restricted access conditions, data owners have to explain why, clearly separating legal and contractual reasons from intentional restrictions. Information about the restricted data will be collected in the metadata file (Annex 1). Where a restriction on open access to data is necessary, attempts will be made to make data available under controlled conditions to other individual researchers.

In general, M4C will consider data assets accessible if they are available through one or more trusted repositories. TNA beneficiaries who choose alternative approaches are required to outline in their TNA data management plan how they will ensure accessibility for their assets, how they intend to secure adequate storage and backup facilities to maintain all produced data assets after the project ends, and take full responsibility for long-term data preservation.

3.3 Making data interoperable

Data generated in the M4C project will adhere to standard formats upon publication and be as much as possible compliant with available (open) software applications, facilitating interoperability. All data providers are responsible for making sure their data adhere to international standards before publication to ensure interoperability. Which standards apply, depends on the data type.

While M4C does not mandate specific formats and vocabularies, data assets must be hosted in a trusted repository and appropriately annotated and formatted according to the repository's best practices. If any partner or TNA applicant opts for alternative hosting solutions, they must outline the metadata annotations and formats they intend to provide for each of their published data assets in their data management plan.

Acceptable domain-specific metadata formats and vocabularies are listed in FAIRsharing, and their use is further described in RDMkit and the FAIR Cookbook. The FAIR Cookbook also provides extensive documentation on domain-specific formats for omics and other Life Sciences fields.

3.4 Making data reusable

The data generated in the M4C project will be made available as Open Data to permit the widest possible reuse. Only SMEs can request exceptions from this rule for commercial, ethical, legal or security reasons. Project partners and TNA beneficiaries may request data publishing embargoes of up to two years, e.g., to maximize the scientific impacts of the project or prepare exploitation measures. After this period, all data will be made open. For data that can be publicly released open licenses such as CC BY or CCO will be used.

The open data will be made available in trusted data repositories, as described earlier. When necessary, to facilitate data reuse and data analysis validation, documentation about data sources, used methodologies, codebooks, data cleaning, analyses, variable definitions, units of measurements, etc., will be provided for each data asset.

If data assets are updated, the owner of the data has the responsibility to manage the different versions and to make sure that the latest version is available in the case of publicly available data. The quality control of the data is the responsibility of the owner of the data.

3.5 Publications

M4C conforms with Horizon Europe requirements to make all peer-reviewed scientific publications generated by M4C Open Access immediately after publication. Publications will either (1) be published in an Open Access journal, or (2) be deposited in a repository for scientific publication (e.g. Zenodo) and be made open access.

3.6 Allocation of resources

The recommended trusted data repositories are free of user-charge. Project partners have a specific budget for Open Access publication.

4. Harvard Dataverse

In the MICROBES-4-CLIMATE project, following the Open Data, Software, and Code Guidelines suggested by the European Commission, we intend to implement a central repository for storing, sharing, and preserving project metadata using Harvard Dataverse. Indeed, Harvard Dataverse is widely recognized as an open-source web application designed to share, preserve, cite, explore, and analyze research data according to the FAIR (findability, accessibility, interoperability, and reusability) principles. Therefore, the primary objective of utilizing Harvard Dataverse in this project is to provide a standard framework for the accessibility, transparency, and long-term preservation of all generated metadata. By depositing such metadata into Harvard Dataverse, we aim to facilitate data sharing within the research community, promote reproducibility, and comply with funding agency requirements regarding data management and sharing policies.

We now discuss some aspects of data management through such an implemented repository.

Data Deposit:

Upon the completion of each phase of data collection and analysis, all relevant research metadata, including raw datasets, processed data, code, and associated documentation, will be deposited into our Harvard Dataverse repository, following standard metadata schemas (defined by experts in the field) to enhance discoverability and facilitate future reuse.

Access Control:

Access to the deposited datasets will be managed under our data-sharing policy. While some datasets may be made openly accessible to the public, others may require restricted access due to privacy, confidentiality, or proprietary concerns. Access permissions will be configured accordingly, ensuring compliance with applicable legal and ethical standards.

Data Sharing and Collaboration:

The use of Harvard Dataverse facilitates seamless data sharing and collaboration among researchers by providing tools for versioning, licensing, and citation of datasets. We will actively encourage other researchers to access and utilize our datasets for secondary analyses, replication studies, and interdisciplinary collaborations. Proper citation of the original data source will be enforced to acknowledge the contributions of the data producers.

Data Citation:

To promote proper attribution and recognition of our research outputs, we will assign persistent identifiers (DOIs) to each deposited dataset. These identifiers will facilitate accurate citation of the data in scholarly publications, enabling others to trace the provenance of the data and provide appropriate credit to the data creators.

Data Retrieval:

The Harvard Dataverse API, which is REST-based, offers a range of endpoints designed for accessing metadata and data from Harvard Dataverse repositories. This functionality empowers stakeholders to develop customized applications, scripts, and workflows that seamlessly interact with Dataverse repositories, automating the retrieval of datasets and their associated metadata.

Integration with External Systems:

By utilizing the Harvard Dataverse API, we facilitate integration with external systems and platforms, including research data portals, content management systems, and data analysis tools. This interoperability fosters seamless data exchange and collaboration between our repositories and other research infrastructure, achieved through the exportation of data in XML and JSON formats.

5. Data security

Data will be stored in trusted repositories, which have provisions in place for data security.

6. Data ethics

As stated in the M4C Consortium Agreement, results, thereby including all types of data assets, are owned by the Party that generates them.

Where it is not possible to establish the respective contribution of each participant to the result or to separate the result, then for the purpose of applying for, obtaining or maintaining their protection, the provisions for Joint Ownership (detailed in Grant Agreement Article 8.2 and Article 16.4 and its Annex 5) apply.

MICROBES-4-CLIMATE is therefore committed to a fair and transparent management of intellectual property, and condemns all predatory behaviour aimed at superseding, circumventing, or working around the above described data ownership principles.

Furthermore, the M4C consortium acknowledges that FAIR data and Open data practices should be an instrument to promote reproducible research and knowledge transfer and not means to transfer intellectual property and/or to prevent data owners from exercising their rights.

7. Personal data and privacy

Personal data is information relating to an identified or identifiable natural person. An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Article 2(a) EU General Data Protection Regulation 2016/679 (GDPR)). Individuals are not considered ‘identifiable’ if identifying them requires excessive effort.

M4C uses personal data only for internal purposes, i.e., for TransNational (TNA) applications, organizing meetings and workshops, sending email etc. and does not handle sensitive data. Personal data is not an object of research in M4C. The data can be categorized as:

1. Personal data: e.g. contact lists of consortium members, mailing lists, etc. that are used by the consortium is stored in the M4C Synology Platform managed by MO, in compliance with GDPR policy (Data Privacy and Accessibility)

2. Personal data collected from third parties, includes but is not limited to:

a. M4C TNA applications that will require registration from TNA applicants

b. Registration forms for representatives external organizations, participating in a M4C organized workshops

c. Seminars organized by M4C, that require participant registration

d. Surveys (either technological or user surveys) which is not anonymous for any reason

For each activity, M4C will follow EU GDPR regulations in the EU countries and in the UK.

The M4C project has a Synology platform that is shared with all project participants and is used to store and share common datasets (experimental data), project results, contact lists, templates and other data that is required for the implementation and progress of the project. The data collected will be stored in a secure IT environment. Access to the data requires a member of quest status in the MICROBES-4-CLIMATE Synology Platform, which is given to project members by the Project Coordinator. Privacy of data subjects will be secured by fully complying with the General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council). The project consortium has appropriate technical and organizational measures in place to carry out data protection during the project. Each Consortium partner has an appointed Data Protection Officer (DPO), who is responsible for advising the project with regard to compliance with the GDPR.

8. Conclusions

The present document describes the DMP of the M4C project and emphasizes the importance of effective data management in achieving project objectives. By following this DMP, the M4C project can ensure the efficient and responsible management of data generated by project partners and TNA grants, maximizing the impact and value of research outcomes and promoting collaboration and knowledge exchange within the research community.

9. References

[1] European Commission, “Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information,” Official Journal of the European Union, vol. 61, no. L 134, pp. 12-18, 31 May 2018.

[2] European Commission, “Commission Recommendation of 17 July 2012 on access to and preservation of scientific information (2012/417/EU),” Official Journal of the European Union, vol. 55, no. L 194, p. 39–43, 21 July 2012.

[3] European Commission, “Horizon Europe Programme Guide (V2.0),” 11 April 2022. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf. [Accessed January 2023].

[4] Proton AG, “What is GDPR, the EU’s new data protection law?,” 2023. [Online]. Available: https://gdpr.eu/what-is-gdpr.

[5] European Commission, “Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast),” Official Journal of the European Union, vol. 62, no. L 172, pp. 56-83, 26 June 2019.

[6] EU Grants: Data management plan (HE):V1.1 – 01.04.2022

[7] DataCite Metadata Working Group. (2024). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.5. DataCite e.V. https://doi.org/10.14454/g8e5-6293

10. Annexes

10.1 Annex 1 – Metadata file template

DATA DENOMINATION
Data set reference name The reference name will start with the prefix "DA_M4C" indicating it is a data asset *Ex: M4C_stakeholders_survey*
Data set title The title of the data asset which should be easily searchable and findable *Ex: Stakeholders_Survey_Q2_20232*
Description A brief description of the data asset *Ex: Survey on technological tools used to quantify microbial diversity.*
Keywords The keywords associated with the data asset *Ex: survey, tools, microorganisms*
Version Number To keep track of changes to the data assets *Ex: V 1.0*
DATA ORIGIN
Name(s) of data set creator(s) Lead partners responsible for the creation of the data asset, and their affiliation *Ex: Name Surname, Affiliation, Country*
Data Source How/why was the data asset generated/re-used *Ex: generated data set*
Creation Date Date of generation of the data asset

Ex: dd.mm.yyyy

Quality Assurance

Brief description of the quality assurance processes the data has gone through

Ex: response rate monitoring, data profiling,

DATA SPECIFICATIONS

Type

Type of data contained in the data asset

Ex: survey data

Format

This could be CSV, PCAP, HTML, etc.

Ex: .xlsx

Expected Size

The approximate size of the data asset

Ex: 1 MB

DATA ACCESSIBILITY

Data Location

The repository where the data are stored: partner owned, M4C Information Platform, other trusted data repository (e.g., Zenodo, GBIF, etc.)

Ex: GBIF

Date of Repository Submission

The date of submission to the repository can be added once it has been submitted

Ex: yyyy.mm.dd

Persistent Digital Identifier

The PDI can be entered once the data asset has been deposited in the repository

Ex: https://doi.org/10.xxxx/dome.xxxxx

Access Status

Whether a data asset is “open”, “consortium” or “restricted”

Ex: consortium

Embargo Period Duration of embargo foreseen for a data asset *Ex: no embargo*
Funding statement A disclaimer indicating that the specific data asset was generated within the M4C project that received funding from the EU
DATA UTILITY
Significance to the project The associated work package or TNA grant this data asset originates from *Ex: WPx – Task y.z*
Significance outside the project To whom the data might be useful outside the project *Ex: policy makers*
DATA PUBLICATIONS
Related Publications Bibliographical details of publications based on the data asset *Ex: deliverable x, scientific article y, …*
Data set Citation A ‘ready-to-use’ citation reference for the data asset *Ex: Author1 Surname, Name, Author2 Surname, Name. (yyyy). Publication title. Publisher.*