MIBIREM – Toolbox for Microbiome based Remediation

Deliverable 1.2

Data management plan version 1

Due Date:	31.03.2023
Submission Date:	31.03.2023
Dissemination Level:	PUBLIC
Lead Beneficiary:	Utrecht University - UU
Lead Author:	Alraune Zech, a.zech@uu.nl
Project Acronym: MIBIREM	**Project Number**: 101059260
Start Date of Project: 01/10/2022	**End Date of Project**: 31/03/2027

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Document Control Information	
Title	*Data management plan version 1*
Editor	*Alraune Zech, UU*
Contributors	*Aristotelis Kandylas, UU & Sona Aseyednezhad, UU*
Reviewer/s	*Thomas Reichenauer, AIT & David Donnerer, RTDS*
Dissemination Level (please choose just one according to the DoA)	☐ **SEN** - Sensitive (please provide one-page Publishable Summary) ☒ **PU** - Public
Approved by (please tick the boxes after all the partners have provided their written approval of the deliverable)	☒ RTDS (COO) ☒ AIT (SCO) ☒ UHAS ☒ CNRS ☒ UGENT ☒ SENSA ☒ ALTAR ☒ DND ☒ UU ☒ UNIPI ☒ Tauw
Underlying IPRs	*Are there intellectual property rights included in this deliverable?* **No**
Underlying Datasets	*Are there relevant datasets included in this deliverable?* **No**

Version/Date	***Change/Comment***
Version 1, 10/03/2023	Alraune Zech
Version 2, 17/03/2023	Alraune Zech
Version 3, 27/03/2023	Alraune Zech
Final version, 30/03/2023	Last edits by RTDS

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Table of Contents

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

1 Publishable summary

The purpose of task T1.4 is data management, particularly setting up and updating a data management plan. This document assists consortium members on high-quality data management in compliance with funder requirement:

- strategy of data management
- general measure of data management
- summary of data expected and re-used in the project
- application of FAIR principles to data produced during the project
- handling other research outputs (digital like software or protocols and physical output)
- strategies for data storage, allocation of resources and data security
- activity plan for data management

The project data management plan will be updated at later stages.

Disclaimer: The document does not in any way overrule the Grant Agreement (including its associated Annexes) and the Consortium Agreement.

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

2 Introduction

Bioremediation is a cost-effective and eco-friendly way to remediate sites polluted with organic contaminants. A successful application of bioremediation requires the understanding and control of the microbial networks that lead to degradation of contaminants. The MIBIREM project will adapt and streamline microbiome science to the needs of applications, to exploit microbiomes for bioremediation, creating and applying a TOOLBOX to identify, analyse, cultivate and up-scale microbiomes, while ensuring safety and policy alignment.

In the MIBIREM project, 11 international partners will perform research together in 8 work packages over a period of 4.5 years. This does not only require good project management but also a well organised data management. This includes foremost the data management plan (DMP), but also other work performed by the data management group as outlined in the next chapter. The Data Management Plan will particularly focus on developing strategies of data handling in line with the FAIR principle.

The data management (DM) is part of work package WP1 as task T1.4. It will continue over the entire duration of the project with 3 main deliverables: the DMP at month 6, month 26 and month 54. All partners are involved in the data management. For T1.4, a DM group is formed with one representative from each project partner at least. The DM group will discuss open points on data management and will decide on common DM strategies. This will be taken in agreement with legal data privacy regulations. We will here consult GDPR (General Data Protection Regulation of EU) , the Grant Agreement (GA) and the Consortium Agreement (CA)[1].

All partners will benefit from good data management as it provides them guidance on data storage, data publication, meta-data standards, etc. During the course of the project, the DM group will also define standard file formats for sample data, including metadata aligned with T1.1 to guarantee data usability for all partners and avoid data loss due to a lack of documentation.

We expect the production of new data from field investigations, laboratory investigations, data processing and visualisation, simulations and interpretation method development. We will also make use of existing data which will be clearly specified and distinguished from newly produced data.

In terms of data types and formats, we expect extensive datasets from DNA and RNA next-generation sequencing (fastq.gz), including taxonomic marker gene datasets and DNA shotgun libraries from soil, groundwater, bac-traps and stable isotope probing-SIP. Besides, we collect physical and chemical soil and groundwater data, environmental conditions, and metabolite datasets. Metadata will potentially be stored in excel, csv, txt, and/or word files, adapted to database and repository requirements. We will write R and python scripts for data processing and analysis and a detailed Python-software suite will be developed. Alongside, simulation data on bioremediation modelling (vtk files) will be generated.

For data produced during the project, we follow the “as open as possible but as closed as necessary” principle for data management. The DM particularly focuses on strategies of data handling in line with the FAIR principle. Raw data storage (field investigations, laboratory investigations) will be managed by each group producing it using cloud or server solutions with back-ups and possibilities to share data between groups.

At mature status, results will be published and data will be made findable and accessible by publication in trusted repositories including unique identifiers (DOIs). Processed observation data, software development and modelling results will be published either in the supplementary material along with a publication (having a DOI) or separately in a

[1] An overview table of used acronyms is provided in the Annex at the end of the document.

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

trusted data and software repositories, or both. However, licences for data sharing and re-use (e.g. Creative Commons, Open Data Commons) remain subject to the individual groups.

We will perform Intellectual property (IP) management for generated datasets specifically in task 7.3 (WP7). The IP management will be led by RTDS to avoid conflicts of interest, as RTDS will not own any results /IP generated in MIBIREM.

IP management started during proposal preparation with a confidentiality agreement signed by all partners. Project results and their readiness for exploitation and partner contributions will be assessed to clarify ownership. At the end of MIBIREM a results ownership list (ROL) including taken or recommended IP protection measures will be included in D7.4. For all results the possibility to protect IP and commercially exploit them will be analysed before they are disseminated. IP management will not only look at key exploitable results (KERs), but at all results generated in MIBIREM. Data and software are covered by copyright; most will be freely available under specific licences (e.g. Creative Commons CC-BY).

Prediction tool software will be freely available but licensed for use to end-users. Several results might be suitable for protection via patents: soil bac-traps to capture pollutant-degrading bacteria (CNRS); modified microbiomes for bioremediation (several partners); direct evolution equipment (ALTAR); single bacterial strains could be patented but are maybe not worth the effort and cost. Post-project IP agreements: The IP of microbiomes undergoing evolution in ALTARs equipment will be co-owned by the provider of the sample and ALTAR, so agreements will be needed between partners. The specific are outlined in the Consortium Agreement.

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

3 General measures for data management

3.1 Data management plan online tool

The data management plan will be in place over the entire period of the project. We make use of the online-tool DMP-online (https://dmponline.dcc.ac.uk). The DMP in this online tool will be regularly updated and saved in the project administration repositories with related data. We will create 3 main versions:

- 1st version at M6, being part of this deliverable
- 2nd version at M26, being deliverable D1.3
- Final version at M54 at the end of the project.

We use the Horizon Europe Template, Version 1.0 (05 May 2021) provided in the DMP-online tool. It provides a list of questions on the main topics:

1. Data summary
2. FAIR data
3. Other research outputs
4. Allocation of resources
5. Data security
6. Ethics
7. Other issues

We address all these questions to outline the details of the DMP. This will be the content of the following sections which are structured following the main topics and providing the questions as well as our answers to them.

3.2 Additional measures of data management

We furthermore take a list of general measure to ensure high-quality data management:

- The key measure is the formation of a data management group where each project partner is represented. We meet regularly (and will continue to do so) at intervals of 2 weeks to a maximum of 2 months, depending on the status of the DMP. During meetings, we assess the status of the DMP, assess the dissemination of data and their potential publication. We will assist researchers in the project to prepare the data according to project needs as well as following the FAIR principles.
  Each meeting will be documented including minutes being checked and approved by all participants. Meeting material and minutes will be stored in the MIBIREM intranet (GoogleDrive) and will be available to all project partners throughout the entire project duration.

- To assist project partners in preparing their data according to FAIR principles, we will make use of FAIRness assessment tools such as FAIRdat or SATIFYD (or other tools as recently evaluated by Krans et al, (2022)).

- The DM group will collaboratively setup data file templates (next step). This will be done aligned with metadata standards used in the project (see section 5). Using metadata standards of the individual disciplines guarantee that variables are exactly defined. An example is the MIxS checklist (https://gensc.org/mixs/) for genomics. This will maximise interoperability and reusability of the material which are crucial for the data use in the tools to be developed during the project: the prediction tool (WP4) and the MIBIREM TOOLBOX post-project.

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- We consult with Research Data Management Support Staff of UU providing expert knowledge and guidance on the DMP as well as direct data management, e.g. handling of data storage solutions, data publication, metadata standards, licences, and infrastructural support for database setup.

3.3 Dissemination and exploitation of data

Owners of resulting research datasets are required to develop and implement a dissemination and exploitation strategy for such data. As a basic principle, it is vital to understand that the obligation to exploit results does not replace the obligation to disseminate results, and vice versa. Hence, there could be something to disseminate about the exploitation strategies for generated research datasets, provided that such data is not marked as confidential information according to the conditions established in Section 10 of the MIBIREM Consortium Agreement (CA) (i.e. non-disclosure of information).

Art. 16 of the GA stipulates that exploitation strategies may refer to post-project Research & Development (R&D) activities, the use of research data for the development and commercialization of products and services, and/or in standardisation activities. It is required that MIBIREM partners balance open access & open data requirements with potential exploitation opportunities, in particular commercial ones. Thus, all datasets will be first checked for their exploitation potential before dissemination. Data underlying scientific publications will be made available without restrictions. For all other datasets, it will be decided on a case-by-case basis whether access can be granted without giving away sensitive information. For example, data whose release would limit the patentability of a result will not be made accessible to the public. In general, access will be as open as possible, but as restricted as necessary.

The commercial potential of expected MIBIREM results, especially of results with potential to be KERs, has been already addressed in the proposal which turned into the DoA. Consequently, it is crucial that post-project use of generated research datasets contribute to the further development and delivery of partners' exploitation plans.

Along these lines, project partners may also request access rights to the owner of resulting research datasets under fair and reasonable conditions, whenever it is needed for the development and implementation of their exploitation plans. This is stipulated in Art. 9.4.1 of the MIBIREM CA:

> *Access Rights to Results if Needed for Exploitation of a Party's own Results shall be granted on Fair and Reasonable conditions. Access Rights to Results for internal research and for teaching activities shall be granted on a royalty-free basis.*

According to Art. 17 of the GA, the research data generated in MIBIREM must be disseminated as soon as possible and by appropriate means (other than those resulting from protecting or exploiting the results), including in scientific publications (in any medium). Such means and conditions of the dissemination activities fall under the responsibility of the owner of the resulting research dataset, with support and guidelines established by the project's DMP subgroup.

The GA has also established the obligation to share research data (and metadata) underlying (peer-reviewed) scientific publications, in Art. 17.4 on open access publications:

> *The beneficiaries must ensure open access to peer-reviewed scientific publications relating to their results.*

Such research data can also refer in this context to relevant data as background, whenever it is needed to validate the analytical activities described in a (peer-reviewed) scientific publication generated in MIBIREM.

Moreover, MIBIREM partners have agreed during the proposal development to adhere to Open Science principles with regard to research data generated (as result) in the project. This applies not only to the data and metadata needed to validate results in scientific publications, but also to other curated and/or raw data and metadata that may be required

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

for validation purposes or with re-use value. All legal requirements are described in Art. 17 of the GA (Open science: research data management):

> *The beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:*
>
> *• establish a data management plan ('DMP') (and regularly update it)*
>
> *• as soon as possible and within the deadlines set out in the DMP, deposit the data in a trusted repository; if required in the call conditions, this repository must be federated in the EOSC in compliance with EOSC requirements*
>
> *• as soon as possible and within the deadlines set out in the DMP, ensure open access — via the repository — to the deposited data, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, following the principle 'as open as possible as closed as necessary', unless providing open access would in particular:*
>
> > *• be against the beneficiary's legitimate interests, including regarding commercial exploitation, or*
> >
> > *• be contrary to any other constraints, in particular the EU competitive interests or the beneficiary's obligations under this Agreement; if open access is not provided (to some or all data), this must be justified in the DMP*
>
> *• provide information via the repository about any research output or any other tools and instruments needed to re-use or validate the data.*
>
> *Metadata of deposited data must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following: datasets (description, date of deposit, author(s), venue and embargo); Horizon Europe or Euratom funding; grant project name, acronym and number; licensing terms; persistent identifiers for the dataset, the authors involved in the action, and, if possible, for their organisations and the grant. Where applicable, the metadata must include persistent identifiers for related publications and other research outputs.*

3.4 Intellectual Properties and Rights

The balance between the project's openness and confidentiality requirements for research data generated within the project will be discussed on a case-by-case basis by project partners. It is, however, evident that whenever a generated research dataset has potential commercial or industrial value, the owner of such a dataset must take all pertinent measures to protect it, in line with the obligations of the Grant Agreement (GA). Specifically, Art. 16 of the GA details the obligation to protect results:

"Beneficiaries which have received funding under the grant must adequately protect their results — for an appropriate period and with appropriate territorial coverage — if protection is possible and justified, taking into account all relevant considerations, including the prospects for commercial exploitation, the legitimate interests of the other beneficiaries and any other legitimate interests. "

A research dataset is any organised collection of data defined by a theme or category that reflects what is being measured, observed, or monitored. In this regard, an original research dataset may be protected in Europe by Copyright

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Law. This means that the developer of such a dataset may claim an exclusive right for 70 years (this duration can differ in non-EU countries) to exclude others from using, copying, and exploiting this dataset without prior given consent.

Also, partners may claim the sui generis right for the protection of original databases for a period of 15 years. The sui generis database right protects a database, which is defined as follows in Article 1 (2) of the EU Database Directive (96/9/EC):

"[…] a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means." (DIRECTIVE 96/9/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (March 1996) on the legal protection of databases. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009)

Furthermore, research datasets generated by partners within the project can also be marked as "Confidential Information" according to Section 10 of the CA, with the purpose to limit access for disclosure outside of the consortium. Such restriction applies not only for the duration of project implementation, but also for a period of 5 years after the project has ended.

Taking all this into consideration, all partners are highly encouraged to mark their data as confidential whenever it might be needed for the implementation of their exploitation strategies.

(Joint) ownership claims and further suitable protection measures (e.g. IPR) are analysed and presented in WP7 (DEC).

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

4 Data Summary

Will you re-use any existing data and what will you re-use it for?

Data will be reused in the project. The two main sources of existing data reused in MIBIREM is data from preliminary investigations at the project field sites as well as databases on microbiological and genetic information. Details on the reused data sets are selected in a table "DMP_Data_Reused_VX.xlsx" being accessible by all Consortium members. The current version is added to the Annex. The table outlines type and description of the data, as well as owner, source, formats, software required, maximum size, number of data files, purpose of use in the MIBIREM project, metadata standards, storage location, linked publications, accessibility and licence.

What types and formats of data will the project generate or re-use?

Data types and formats of reused data can be found in the table "DMP_Data_Reused_V1.xlsx" (in the Annex).

A similar table "DMP_Data_Produced_V1.xlsx" is set up containing all data-management relevant aspects. As soon as data is produced during the project, the information on the data sets will be specified in the table. At this early state of the project no new data sets have yet been produced.

We foresee research data to be newly produced in the project of the following origins with types and formats:

- Laboratory experimental (raw) data on
  - analytical data
  - molecular biomass
  - sequencing data
  - genomic data
  - bacterial diversity data
  - gene quantification
  - list of bacteria and cultures
  - list of functional genes
  - list of qPCR primers
  - (additional) numerical lab data
  - lab journal and protocols
    Formats: csv/.biom/.fastq/.fasta/.xlsx/.txt/.doc/.docx/hand written notes
- Field - experimental data:
  - site descriptions, like sedimentary descriptions, hydrological parameters, biogeochemical conditions,
  - coordinates of sampling points
    Formats: csv/.xlsx/.txt/.doc/.docx/hand written notes
- Numerical data/Processed data:
  - statistical output data & results visualisation
    Format: .csv/.tsv/.tiff/.jpeg/.png/.pdf/.svg/.rmd/.RData/.qmd
  - Software implementation data
    Format: .r/.py/.sh
  - Simulation output:
    Format: .vtk, .csv, .pdf

Non-scientific data produced in the project include documents on:

- Activity plans, project descriptions, contracts, templates and guidelines, meeting protocols, presentations
- DMP
- website

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The purpose of each data set is specified in the tables "DMP_Data_Reused.xlsx" and "DMP_Data_Produced.xlsx" (in the Annex).

The general purpose of generating sequencing data is to gain novel information on bacterial consortia. The data will be interpreted in combination with the metadata to understand the functioning of a bacterial consortium. This also includes statistical analysis (processed data) and modelling (numerical data). All these data will provide new insights into the behaviour of microbiological processes and dependencies in the subsurface. They will be used to find solutions for enhancing bioremediation which is the purpose of the project.

What is the expected size of the data that you intend to generate or re-use?

The number of files and size of each data set is specified in the tables "DMP_Data_Reused.xlsx" and "DMP_Data_Produced.xlsx" (in the Annex).

What is the origin/provenance of the data, either generated or re-used?

The origin of each data set is specified in the tables "DMP_Data_Reused.xlsx" and "DMP_Data_Produced.xlsx" (in the Annex).

The research data (newly) generated during the project will origin from research performed by the project partners through:

- experiments (raw data)
- interpretations/visualisations of experimental data (processed data)
- implementation of algorithms (data interpretation methods)
- modelling output (predictions based on observation data)
- documentation data produced by researcher (manuscripts)

To whom might your data be useful ('data utility'), outside your project?

- other researchers
- stakeholders
- public companies

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

5 FAIR data

5.1 Making data findable

Making data findable, including provisions for metadata: Will data be identified by a persistent identifier?

All research data sets produced during the project will be findable through a unique identifier (UID). Data will either be published directly with a persistent identifier (DOI), in repositories/databases providing UIDs) or along a scientific publication (with DOI), e.g. in the supporting material or with link to an open archive.

We foresee for the different types of research data during the project:

- Experimental (raw) data: direct publication of the data set (with DOI) or deposition into a public repository providing unique identifier (e.g. sequence accession numbers)
- Processed experimental data: published in scientific articles.
- Numerical data: direct publication of the data set (with DOI)

Making data findable, including provisions for metadata: Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Newly created data sets will be provided with rich metadata.

As a consortium project with research groups from various disciplines, we expect data of various kinds. As for all the project related disciplines metadata standards exist, we will make use of them.

For sequencing data, we will follow the metadata standards of the NCBI repositories NCBI Genbank and NCBI SRA as these are the standard discipline archives/repositories/databases (see also questions at 5.2) for depositing nucleotide sequences and raw sequencing data.

For one, NCBI uses the metadata standard "BioSample Data Model" to store metadata associated with biological samples. The BioSample Data Model is a standardised format for describing biological samples and their associated metadata, including sample identifiers, organism names, sample attributes, and other relevant information. BioSample Data Model is compatible with other metadata standards, such as MIAME (Minimum Information About a Microarray Experiment) and MIGS/MIMS (Minimum Information about a Genome Sequence/Minimum Information about a Metagenome Sequence).

The SRA metadata describes the technical aspects of sequencing experiments: the sequencing libraries, preparation techniques and data files. Most of descriptive information is captured at the level of the SRA EXPERIMENT and will be displayed in the public record.

Moreover, SRA provides specialised bioinformatic tools, (SRA Toolkit 3.0.3[2]). The tools facilitate deposit, conservation, identification and management of raw sequencing data, related metadata and also elaboration: The SRA Toolkit and SDK from NCBI is a collection of tools and libraries for using data in the International Nucleotide Sequence Database Collaboration (INSDC[3]). The INSDC is a long-standing foundational initiative that operates between DDBJ (DNA DataBank of Japan), ENA (European Nucleotide Archive) and NCBI. INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

[2] https://github.com/ncbi/sra-tools#the-sra-toolkit

[3] https://www.insdc.org/

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Otherwise, we will make use of the common metadata standard "DataCite". This is also the one Zenodo's metadata is compliant with (again see also questions at 2.2 on repositories).

The outlined metadata standards will guide the design of the standard file formats for sample data which will be defined in the second phase (M6-M12). The Handbook for sampling and sample treatment (T1.1) will also be aligned with the metadata standards to guarantee data usability for all partners and avoid data loss due to a lack of documentation.

Details on the metadata standards will then be specified for each data set after publication in Table "DMP_Data_Produced.xlsx", Annex.

Making data findable, including provisions for metadata: Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?

Keywords will be provided in the metadata according to the scientific domain-discipline, each dataset is related with. In this way we can ensure that interested parties can find our datasets.

Making data findable, including provisions for metadata: Will metadata be offered in such a way that it can be harvested and indexed?

The metadata for our datasets will comply with the "*Datacite v4*" and "*Dublin Core*" metadata standards and will be offered in file formats (.xml, .json) that can be harvested and indexed.

In addition, Zenodo, which will be used as a main repository of our project, facilitates any metadata harvesting and indexing with the use of the OAI-PMH protocol (Open Archives Initiative's Protocol for Metadata Harvesting) and its own REST API.

5.2 Making data accessible

Making data accessible – Repository: Will the data be deposited in a trusted repository?

The research data produced in the project will be accessible in trusted repositories. Following the terms of the Grant Agreement all data will be open access, unless it is for commercial exploitation or contrary to any other constraints by data (co-)owner. In particular, site owners engaged in MIBIREM can request the site data collected by the project to be confidential for safety and business reasons.

We will use trusted repositories:

- Zenodo being a general purpose repository. Here a project specific community for MIBIREM has been set up.
- ENA (European Nucleotide Archive) and NCBI Genbank and NCBI SRA (Sequence Read Archive) being domain repositories, i.e. a trusted repository established for the specific research domain of microbiology and bioinformatics.

The consortium agreed on using the listed repositories with Zenodo as a major repository to bundle all project data. However, the use of ENA and NCBI for genomic and sequencing data is agreed within the group given several reasons:

- Using NCPI/ENA increase re-usability of the produced research data:
  - NCBI and ENA are the most widely used and recognized repositories for sequencing data with a high level of recognition and acceptance (which cannot be provided by the multi-purpose repository Zenodo).
  - open-source software (e.g. source code deposited on GitHub) to access information is optimised on their application to NCPI/ENA
  - NCBI offers specialised data management tools (SRA ToolKit v.3.0.3) designated for retrieving and reusing data
- Some discipline specific research journals require authors to deposit their sequence data in one of the discipline specific repositories (ENA, NCBI, DDBJ).

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Details on the used repository will then be specified for each data set after publication in Table "DMP_Data_Produced.xlsx", Annex.

Making data accessible - Repository: Have you explored appropriate arrangements with the identified repository where your data will be deposited?

We explored the arrangements of the repositories. As stated, a Zenodo community for MIBIREM has been setup:

https://zenodo.org/communities/mibirem/?page=1&size=20

It allows all partners to deposit data at unlimited size.

Making data accessible - Repository: Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, the mainly used repository (Zenodo) for our project provides by default an unique identifier (DOI) to every paper/data publication, ensuring the accessibility of our datasets on this platform.

Data published in ENA and/or NCBI SRA also is provided a unique identifier, like DOI or BioProject and sequence accession numbers.

Making data accessible – Data: Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

We follow the principle 'as open as possible but as closed as necessary'. Most research data produced in the project will be openly accessible. Following the terms of the Grant Agreement (Annex 5) all data will be open access unless it is for commercial exploitation.

We currently expect that data produced solely by research groups of the consortium will be shared openly. Industry partners in the consortium indicated that opening specific data goes against their legitimate interests. Research includes field sampling at contaminated sites owned by third parties. Some of the gathered data will be subject to restrictions due to legal and contractual reasons. This includes:

- Data on exact composition of the solution(s) used for remediation at sites as industry partners have commercial interest in this.
- Data on the exact procedure of culture production used for remediation at sites as industry partners have commercial interest in this.
- Protocols used to generate evolution data will be restricted due to commercial interests of industry partners. They will not be accessible within the consortium nor outside the consortium. On a case-specific basis we will evaluate if it is possible to publish data-subsets, like part of the protocol.
- Data on adaptive laboratory evolution (key performance indicator, diversity data - sequencing data, pictures - protocols used to generate the evolution data) will be restricted as this is commercially-sensitive information.
- Restrictions may apply to data collected by MIBIREM partners as well as data from 3rd parties (i.e. contaminated field site owners) when there is no consent for public disclosure of such data. We will consider making part of the data available or after anonymising site specific data. The process will be evaluated in adherence to General Data Protection Regulation (GDPR).

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

When a produced dataset applies to restricted access conditions, these will be explained and put into context with constraints as per the Grant Agreement and Consortium Agreement (4.5 GDPR compliance) in the detailed data description in the Table of produced data (Annex).

Making data accessible – Data: If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Time embargos for data publication or protection of intellectual property (e.g. patents) are not expected for most of the produced research data.

However, for all results the possibility to protect Intellectual Property (IP) and commercially exploit them will be analysed before they are disseminated – this also includes data generated within the project. IP management will not only look at key exploitable results (KERs), but at all results generated in MIBIREM.

We currently anticipate few cases of data produced in the project which will be made accessible after a time embargo as they are of interest for patents which need time to carefully evaluate. The reason for the embargo is thus commercial exploitation. This concerns:

- soil bac-traps to capture pollutant-degrading bacteria
- modified microbiomes for bioremediation
- direct evolution equipment and developments on adaptive laboratory evolution (key performance indicator, diversity data - sequencing data, pictures - protocols used to generate the evolution data)
- single bacterial strains (has to be evaluated under cost/benefit considerations)

Making data accessible – Data: Will the data be accessible through a free and standardised access protocol?

As we plan to publish research data in established repositories, there will be a free and standardised access protocol.

For instance data published in the Zenodo community can be accessed directly through downloading using the DOI and link.

Making data accessible – Data: If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

For produced research data which is subject to restrictions on use, the data set owner will manage the access to the data. For restricted data sets we use the storage facilities specified below. They are all based on the principle of "need-to-know", meaning that only users who require access to the data for their work or tasks are authorised to access it. Every data user can access-log in these facilities with personal-institutional credentials. The owner of the restricted data set will decide how to share the data at different levels (e.g. read only, editor, co-owner) by providing access if requested.

We follow this procedure already for the project documentation/management data which is restricted to project members. Access is by registration on the google drive and authorization through the project manager.

When a dataset applies to restricted access conditions, access arrangement during and after the end of the project will be specified for the data set. A summary of that will be included in the Table "DMP_Data_Produced.xlsx" (Annex).

Making data accessible – Data: How will the identity of the person accessing the data be ascertained?

Open-access data sets will not request the identity of the person accessing the data.

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

For restricted data sets we use the storage facilities specified below. They are all based on the principle of "need-to-know", meaning that only users who require access to the data for their work or tasks are authorised to access it. Every data user can access-log in these facilities with personal-institutional credentials. The owner of the data set will provide authorization.

Making data accessible – Data: Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

We do not expect personal or sensitive (research) data to be produced in the project. If data applies to restricted access this will be handled as outlined before. We thus do not see the need for a data access committee.

Making data accessible – Metadata: Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

We plan to make metadata openly available and licensed under a public domain dedication CC0. At this stage, we do not foresee exceptions. In case of any exception we will clarify why. We will use metadata community standards which will enable users to access the data and increase its findability.

Making data accessible – Metadata: How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Published data will be available and findable long-term. This will be ensured by publishing it in trusted long-term maintained repositories and journals. The same holds for meta-data.

Making data accessible – Metadata: Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

We expect a large amount of data which requires software to access and efficiently read the data (e.g. sequencing data). This will be clearly documented and referenced in the metadata.

In case the needed software is open source and published (e.g. on GitHub, referenced with DOI), we will provide reference. If the software which is needed to access the data, is developed alongside the data it will be either published separately and referenced or included (if not published stand-alone). If the respective software is a commercial software or subject to restricted licence, it will not be possible to include it.

5.3 Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

As part of the project task DMP, we will collaboratively set up data file templates. We seek to use preferable formats not restricted to commercial software such as .txt/.pdf/.odt for Text, .csv./Rdata for quantitative data, .jpg/.tiff for images.

We will use metadata standards of the individual disciplines to guarantee that variables are exactly defined:

- Common metadata standards such as Datacite 4.0 & Dublin Core
- ENA metadata standard
- For genomics data we will follow the MIxS checklist (https://gensc.org/mixs/).

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

The meta data standard is also depending on the repository where data will be preserved or shared. For instance, we will follow Zenodo's best effort principles. For making data interoperable this includes that metadata uses a formal, accessible, shared, and broadly applicable language for knowledge representation such as the JSON scheme as internal representation of metadata.

The use of established metadata standards (including vocabularies, formats and methodologies) will also ensure that data is interoperable between project partner and other interested user.

Making data interoperable: In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

We do not expect the need to use uncommon ontologies nor are we planning to produce specific ontologies or vocabularies in this project.

Making data interoperable: Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

[1]A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)

Where applicable, produced research data will include qualified references to other data.

When research results/data builds up on each other, contextual knowledge will be provided. This holds e.g. for publication of results in scientific publications. Their used and produced data will be linked and with qualified references.

We will also use qualified references and cross-references to outline chains of data dependency and linked data. This will be facilitated by publishing the data within one Zenodo community which is planned for a large part of the data.

5.4 Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Depending on the data type we will provide documentation which facilitates validation of data analysis and re-use. We will make use of readme files, codebooks and other measures for data management. Details on the methodology will be also documented in research papers. Most information will be provided in the method sections. We will also make use of Supporting Material which are common for scientific articles to provide detailed documentation if required. This way we ensure that research data can be validated and is reusable.

Increase data re-use: Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licences, in line with the obligations set out in the Grant Agreement?

We expect most of the data to be freely available in the public domain (open-access repositories) to allow re-use. We will only restrict data if it is subject to commercial use. Details on restricted data are outlined above.

Licences for data sharing and re-use remain subject to the project partners who produced the data. Data licensing will be arranged in accordance with the Grant Agreement following e.g. Creative Commons, Open Data Commons. We expect:

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- CC-BY or CC-BY-NC-SA: for sequencing data
- MIT for software/simulation data, e.g. produced by UGent and UU

Increase data re-use: Will the data produced in the project be usable by third parties, in particular after the end of the project?

Openly published data will be usable by third parties, during and after the end of the project. This will apply to most of the processed data (results and visualisations) as well as developed software and simulation data.

Raw data produced during the project with restricted re-usability (depending on licence) will not be usable by third parties. This will e.g. apply to field observation data from specific contaminated field sites where site owners hold the rights and refrain from data publication for specific reasons, like safety issues.

Raw data on microbiome analysis will be published and reusable by third parties during and after the end of the project.

Increase data re-use: Will the provenance of the data be thoroughly documented using the appropriate standards?

The provenance of the data will be thoroughly documented using the appropriate standards. This will be documented in READMEs, publications as well as in the meta data provided to the published data set on repositories. For instance all data and metadata uploaded to Zenodo is traceable to a registered Zenodo user. But we will make sure that the original authors of the published work are described in the metadata as well.

Increase data re-use: Describe all relevant data quality assurance processes.

We will assure data quality through

- data collection following the handbook for sampling and sample treatment (developed during the project WP1, T1.1)
- sample analysis following a unified procedure (preferably all performed at the same lab)
- use of data templates (developed during the project, WP1, T1.4)
- documentation of data with metadata following community standards as well as meta-data standards of targeted repositories for publication
- use of FAIRness assessment tools for improving FAIRness of data
- review of data within the DM subgroup

Increase data re-use: Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

Besides the management of research data, we will also address the management of other research outputs, such as conference/workshop contributions, software development and physical research output like bacteria strains .

We will also address aspects related to allocation of resources and data security. Ethical aspects do not apply here, as the research data is not related to personal data. See section 6.

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

6 Other aspects

6.1 Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

The principles outlined in the DMP for research data will similarly applied to digital research output (not being data), such as:

- protocols: expected results are the "Handbook for sampling and sample treatment" developed in WP1 (T1.1)
- software: expected results are the prediction tool developed in WP4.
- models: will be partially developed for the sites under investigation, also in co-development of the prediction tool (WP4)
- workflows: expected results are the bioremediation tool box (MIBIREM-TOOLBOX innovation) in WP6.

Physical research output we expect includes:

- physical samples, such as soil samples (WP1)
- laboratory samples, such as microcosms (W2)
- bacteria strains (WP3)

A selection of strains that are active degraders will be publicly deposited in the BCCM/LMG Bacteria Collection. Internet links between the culture collection catalogue and the genome database and the genome sequences will be provided.

Whole consortia that actively degrade selected pollutants will be publicly deposited in the BCCM/LMG Bacteria Collection.

The procedure leading to deposition (including whole genome sequence determination and analysis) will be based on standards developed within the project (Task 3.1 of WP3). This includes setting quality-control standards for public deposit, preservation and distribution of bioremediation microbiomes.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for reuse, in line with the FAIR principles.

The questions addressing the FAIR principles will be similarly applied to digital research output. This digital research output will be published, similar to data, in trusted repositories, with unique identifier (e.g. DOI), open access, with metadata and proper documentation (README etc.). Code and software will be created following good coding practice to make it reusable and interoperable. For instance it will be developed via GitHub and documented including Readmes. Data and software are covered by copyright; most are expected to be made freely available under specific licences (e.g. Creative Commons CC-BY). The Prediction tool software developed in WP4 will be freely available but licensed for use to end-users.

For the physical output of bacteria strains, we will consider the applicable elements of the FAIR principles during the process of preparation and public deposition in the BCCM/LMG Bacteria Collection. Project task 3.1 (WP3) will focus on setting quality-control standards for public deposit, preservation and distribution of bioremediation microbiomes. There the elements of the FAIR principles will find consideration. For instance, management of metadata is done to conform with the MIRRI (Microbial Resource Research Infrastructure) principles (https://www.mirri.org/mirri-microbial-resources/).

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

6.2 Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

Several project teams were awarded with money for open access publication costs. This applies to:

- CNRS: 10 000€ for 4 articles at 2500€ each.
- UGent: 10 000€ for 4 articles at 2500€ each.
- Sensatec: 2000 € for 1 article.
- Utrecht University: 10 000€ for 4 articles at 2500€ each.

Utrecht University was awarded with 7500€ for data management to cover expenses for hardware, software and personnel to support sharing and preservation of the data. From this amount, we expect at least 2000€ to be spent on hardware expenses for providing a central project storage. One option is the UU-institutional repository YODA which can be also used by external partners. (Decision on central project data storage in M13 as part of Milestone MS3). Extending the default storage capacity of 1TB cost 4€ per TB and month. Thus, to store e.g. additional 10TB of data we need 2160€ (10TB for 54 months at 4€/TB/Month).

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

The cost of data publication will be covered by the money granted to the project as specified above. The same holds for the cost specified above on hardware, software and personnel to support sharing and preservation of the data.

Who will be responsible for data management in your project?

All project groups will be involved in the data management. However, the lead on the DMP is by Utrecht University through Alraune Zech. She will be directly supported through a DMP work group consisting of one representative for each project partner. In addition specialised personnel (data stewards) from Utrecht University will provide assistance to the general data management tasks of this project and will ensure the compliance of the data with the Open Science and FAIR principles.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

We will ensure long term preservation of research data by publishing it on trusted and long-time operating repositories. As a consortium we agree to use Zenodo (multi-purpose repository) and NCBI (domain specific repository) as both provide different advantages in use (see above). If unforeseen circumstances require partners to publish in other repositories, this will be discussed in the data management subgroup.

For Zenodo a project community has already been established. This allows all project partners to publish and deposit data without data size restriction. For both Zenodo and NCBI, we therefore do not expect any additional costs for the publication and long time preservation.

We formed a Data management group where each project partner is represented. We meet regularly to assess the status of the DMP, assess the dissemination of data and its potential publication. We will assist researchers in the project to prepare the data according to project needs as well as following the FAIR principles. This way one person from each research group takes the responsibility of administering and supervising the whole process of data management.

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

We are going to use open formats e.g. .csv, .txt because these formats facilitate the interoperability and re-usability of our data by others which will also contribute to long-term preservation. File formats will also be specified during development of standard file formats for sample data as part of the data management (Task 1.4, WP1).

Before each research publication, a selection will be made with regard to which data is useful for long-time preservation. The published datasets will be linked with the according publication and will be given a persistent identifier e.g. DOI. A data selection will be made based on reproducibility and re-use purposes. During our project we have agreed on following the next guidelines for each data publication:

- Raw and processed data included in:
    - The data in the Zenodo repository will be preserved for the lifetime of the repository. As stated by Zenodo on their "retention period": *Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.* (https://about.zenodo.org/policies/)

    - The data deposited to NCBI or ENA will be preserved for the lifetime of the repository/archive. As stated by ENA " All database records submitted to the INSDC [International Nucleotide Sequence Database Collaboration, including NCBI Genbank and ENA] will remain permanently accessible as part of the scientific record. (https://www.ebi.ac.uk/ena/browser/about/policies)

- Physical samples will be stored in the storage of the sampling partner and/or the partner performing the analysis of the data. For instance, physical samples collected along evolution experiments and analysed by Altar, will be stored at Altar's location and preserved for the duration of the project, plus another 2 years after project end. Physical samples processed by UGent (enriched cultures for isolation in WP2 and microbial consortium in WP3) will be stored in the lab and preserved indefinitely (or until use) as part of the UGent database.
- Bacteria strains will be publicly deposited in the BCCM/LMG Bacteria Collection and will also be preserved indefinitely.
- Digital research output will be published on Zenodo following the preservation specifics as outlined above.

6.3 Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Produced or processed research data (during the production process, i.e. premature to publication) from this project will be securely stored by all project partners in individual storage facility solutions, such

- cloud solutions
- institutional repositories
- institutional servers/databases

Cloud solutions used by partners are:

- Microsoft OneDrive (AIT, UU, UniPi)
- Microsoft Sharepoint (UGent, DND, TAUW)
- Google Drive (UHasselt, RTDS ,ALTAR)

All the partners using cloud solutions have special agreements with providers on privacy and information security legislation and guidelines to comply with the GDPR, e.g. storage of data on European servers.

Institutional repositories, clouds, servers and databases:

- YODA (UU)
- OTELo Cloud NC (CNRS)

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

- MySQL servers (UGent)
- Sensatec server (Sensatec)
- ALTAR server (ALTAR)

The institutional storage solutions are part of the University/company infrastructure and comply with privacy and security policies of the concerning institutes.

These facilities allow storage of large data volumes that are backed up on a daily basis, in specialised data centres, not only ensuring the preservation of the data in a trustworthy and dependable location, but also effectively minimising the risk of data loss in the event of any system malfunction or crash.

The used storage facility solutions handle data access based on the principle of "need-to-know". This means that only users who require access to the data for their work or tasks are authorised to access it. Every data user can access-log in these facilities with personal-institutional credentials.

The analysed or resulted data of the project will be published on well-known and trusted repositories (i.e. Zenodo, NCBI etc., see above) and therefore they will be secure through the security provisions of these platforms.

While transfer of sensitive research data does not apply in the project, the use of secure institutional systems can provide assurance against any unanticipated attempts at data tampering, whether deliberate or accidental in nature.

6.4 Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

We foresee no ethical issues for data sharing within the project. As there are no human participants involved in this project, the project will not be submitted for ethical assessments.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

This is not applicable in this project as we are not planning to conduct any surveys.

6.5 Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

No. We only follow the data management procedures provided by the European Commission (Horizon Europe Template).

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

7 Activity plan for data management

Task	Activities	Resp.	By when / Status
T1.4 Data management (M1-M54), UU **Contributors:** **All**	Develop a data management plan in collaboration with all consortium members	UU	March 2023 (M6)
	setup of data file templates (collaborative), including metadata standards aligned with T1.1	UU, all	September 2023 (M12)
	setup an central data storage solution as platform for the data exchange and input of the bioremediation prediction tool	UU	September 2023 (M12)
	Data management plan update	UU	November 2024 (M26)
	Final data management plan	UU	March 2027 (M54)

Deliverables:

D1.2 Data management plan, version 1 – M6 (UU)

D1.3 Data management plan update, version 2 – M26 (UU)

D1.5 Final data management plan, version 3 – M54 (UU)

Milestones:

MS3 Standard data formats implemented and database setup – M13 (UU) (Partners agree on data formats and transfer of data into database.)

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

8 Annex

8.1 Tables of produced and re-used data

Produced data will be listed with all necessary information in an excel-table, named *DMP_Data_Produced_VX.xlsx* (where the VX stands for version number X). The table will be regularly updated and given a new version number. The table is available to all consortium members in the GoogleDrive Intranet. The template of the table for produced data sets has the form:

	A	B	C	D	E	F	G	H
1	**type**	**description of data set**	**owner**	**source**	**dependence on re-used data**	**role in MIBIREM project**	**formats**	**software**
2			Who produced the data set and holds the licence?	How has the data been produced? E.g. experiments (raw data); interpretations/processing of experimental data; implementation of algorithms; modelling output	Has data been reused to produce this data set? If yes provide information on the re-used data.	What is the context of the data set in the MIBIREM project? E.g. In which work package and task was is produced? In which work package will it be used? Has it lead to a deliverable?	What is the format of the data? E.g. csv/.biom/.fastq/.fasta/.xlsx/.txt/.doc/.docx/.tsv/.tiff/.jpeg/.png/.pdf/.svg/.rmd/.Rdata/.qmd/.r/.py/.sh	Is there a special software needed to open/read the data set?
3								

I	J	K	L	M	N	O	P	Q
max size	**number**	**purpose of use in the MIBIREM project**	**meta data standard**	**storage/ repository**	**findable/link**	**publication**	**dissemination/ accessibility**	**exploitation/ licence**
What is the size of the data set? Particularly important is the order of magnitude, i.e. kb,Mb,Gb,Tb	Number of files (if applicable)	Why are we re-using this data set? E.g. the data is need for performing task X.X of the project.	What metadata standard is used.	Where is the data stored? Provide name and type of repository (e.g. type: domain, institutional, general purpose). How long will data be stored?	Where is the data set located? Does a persistent identifier like doi exist? Provide a link/doi to the data.	Has this data led to a publication? If YES, provide link/doi to publication.	Is the data openly accessible? If the data is restricted, why?; How will you restrict access? How will you enable access to those authorized?	Terms of conditions for data use. If not open access explain why.

Re-used data is listed with all necessary information in an excel-table, named *DMP_Data_Reused_VX.xlsx*. (where the VX stands for version number X). The current version is V1, which will be updated if necessary. The table is available to all consortium members in the GoogleDrive Intranet. The current table for re-used data sets has currently the form:

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**type**	**description of data set**	**owner**	**source**	**formats**	**software**	**Max size**	**number**	**purpose of use in the MIBIREM project**	**meta data standard**	**storage/ repository**	**findable/ link**	**Publication**	**dissemination/ accessibility**	**Exploitation/ licence**
2			Who produced the data set and holds the licence?	How has the data been produced?	What is the format of the data?	Is there a special software needed to set? open/read the data set?	What is the size of the data	Number of files (if applicable)	In which WP/task will the data be used?	What metadata standard is used.	Where is the data stored? State repository.	Where is the data set located? (provide persistent identifier like doi)	Has this data led to a publication? (link/doi)	Is the data openly accessible or restricted?	Terms of conditions for data use.
3	qPCR standards and standard curves	qPCR amplification results (Ct, calibration line coefficients, efficiency) for 16S rDNA, PAH-RHDα GN and GP, alkB, Catechol 1,2 and 2,3dioxygenase genes	Aurélie Cébron (CNRS)	raw data	.csv/.pdf		500kb	6	T2.1, T2.3 and T3.1	no standard for metadata	locally			open	
4	protocole for DNA-SIP	DNA-SIP incubation and treatment (CsCl ultra-centrifugation, fractionation, validation) procedure	Aurélie Cébron (CNRS)	protocole	.txt/.pdf		500kb	1	T2.3	no standard for metadata	locally			open	
5	BactoTraits database use	Scripts to infer bacterial traits from bacterial diversity data	Aurélie Cébron (CNRS)	R scripts	.R		200ko	10	T2.1	no standard for metadata	locally			open	
6	BactoTraits database	Bacterial traits database	Aurélie Cébron (CNRS)	data table	.csv/.xlsx		100Mo	1	T2.1	no standard for metadata	locally	https://doi.o		open	
7	Field measurement and description of site (existing before project)	soil and groundwater characteristicspad contaminant physico-chemical and measurements	MIBIREM & site owner	report on sampling campaign	.pdf		kb	1 per site	WP1	no standard for metadata	locally			open	
8	Field measurements groundwater (existing before project)	pH, EC, ntu of groundwater while sampling	MIBIREM & site owner	data table	.csv/.xlsx		kB	1 per sampling location	T1.3, T1.4, T4.1-T4.4, T5.2-T5.4	no standard for metadata	locally			open	
9	Soil characteristics (existing before project)	Type of soil, geological information, location-specific characteristics	MIBIREM & site owner	data table	.csv/.xlsx		MB	1 per sampling location	T1.3, T1.4, T4.1-T4.4, T5.2-T5.4	no standard for metadata	locally			open	
10	Coordinates (existing before project)	X/Y coordinates of sampling points	MIBIREM & site owner	data table	.csv/.xlsx		kB	1 per sampling location	T1.2, T5.1-T5.4	no standard for metadata	locally			open	
11	Pictures (existing before project)	Photos of the sites and sampling points	MIBIREM & site owner	camera & smart phone	.jpg		MB	>10 pictures per site	T5.1-T5.4, T7.2	no standard for metadata	locally			open	
12	field/soil characteristics (existing before	drilling profiles	MIBIREM & site owner	processing of data	.pdf		MB	< 10 profiles per site	T1.2, T1.3, T5.2-T5.4	no standard for metadata	locally			open	

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
12	type	description of data set	owner	source	formats	software	Max size	number	purpose of use in the MIBIREM project	meta data standard	storage/ repository	findable/ link	Publication	dissemination/ accessibility	Exploitation/ licence
13	experimental data (existing before project)	data of former degradation studies and remediation projects	MIBIREM & site owner	experiments+interpretation&proc	.csv/.xlsx		kB	5	T2.2, T2.3, T5.2-T5.4	no standard for metadata	locally			open	
14	Microbiome data (existing before project)	microbiome data of existing PHC culture	MIBIREM & site owner	interpretation/ processing	.pdf		kb	1	T2.1	no standard for metadata	locally			open	
15	SILVA database	16S/18S SSU ribosomal database, here taxonomic markers for SSU are deposited	Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures	scientific community	.gz		264 Mb	1	T2.1, T4	no standard for metadata	arb-silva.de	https://www		open	cc-by4.0
16	SILVA database	23S/28S LSU ribosomal database version 138.1 non redundant 99%; here taxonomic markers for LSU are deposited	Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures	scientific community	.gz	64 Mb	1	T4, T3.1	no standard for metadata	arb-silva.de	https://www		open	cc-by4.0	
17	UNITE database	ITS markers	unite community (non profit)	scientific community	.gz		120 Mb	1	T2.1, maybe t4	no standard for metadata	unite.ut.ee d	https://doi.o		open	CC BY-SA
18	Kraken 2 db	Curated set of prokaryotic and eukaryotic genome indexes.	Langmead lab, johns Hopkins University.	Whole genome sequences freely available on NCBI GenBank	.tar.gz		246 GB	>100k	T2.1, T2.3, T2.4	NCBI GenBank, ENA		https://ben	https://doi.o	open	CC BY 4.0
19	GTDB (Genome taxonomy database)	Curated archaeal and bacterial genome sequences and taxonomic information.	The University of Queensland, Australia.	Whole genome sequencing.	.tar.gz		66 GB	>300k	T2.1, T2.3, T2.4	GTDB		https://gtdb	https://doi.o	open	CC BY 4.0
20	MALDI-TOF MS Database	BDAL: reference library used for identification of isolates by pairwise comparison of measured spectrum to database	Bruker Daltonics (Germany)	processing of experimental MALDI-TOF MS data	.btmsp (Bruker file type), .txt, .xlsx, .xml	Bruker Daltonics software: Flex Analysis and MBT Compass Explorer	MB	>10K	T2.4	no standard for metadata	No identifier			Commercial database under license, restricted to UM-UGent staff	

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
20	type	description of data set	owner	source	formats	software	Max size	number	purpose of use in the MIBIREM project	meta data standard	storage/ repository	findable/ link	Publica-tion	dissemi nation/ tion/ accessi bility	Exploita- licence
21	MALDI-TOF MS Database	LM UGENT ID: reference library used for identification of isolates by pairwise comparison of measured spectrum to database	LM-UGent	processing of experiment al MALDI-TOF MS data	.btmsp (Bruker file type), .txt, .xlsx, .xml	Bruker Daltonics software: Flex Analysis and MBT Compass Explorer	MB	>5K	T2.4	no standard for metadata	No identifier			In-house database ' restricted to LM-UGent staff	
22	EzBioCloud database	Standardized 16S rRNA gene sequence database of reference taxa	CJ Bioscience, Inc. (Korea)	Processing of whole genome sequencing data					T2.4	NCBI	EzBioCloud web server	https://www.https://doi.o	open for research ers affiliated with non-profit academi c institutio ns		
23	checkM database	Curated set of marker genes from reference genomes	The University of Queensland, Australia.	Processing of whole genome sequencing data	.tar.gz		GB	33	T3.2	no standard for metadata		https://doi.dhttp://doi.o	open		CC BY 4.0
24	TYGS	Curated set of type strain genomes	Leibniz Institute DSMZ (Germany)	Processing of whole genome sequencing data					T3.2	no standard for metadata	TYGS web server	https://tygs https://doi.o And https://doi.o	open for academi c purposes		
25	LPSN	List of Prokaryotic names with Standing in Nomenclature	Leibniz Institute DSMZ (Germany)	data table	.csv		MB	1	T3.2	no standard for metadata		https://lpsn https://doi.o And https://doi.o	open		CC BY-NC 4.0
26	UniProt	Database of protein sequences	EMBL-EBI	Processing of experiment al data and whole genome sequencing data	.gz		MB		T3.2	no standard for metadata		https://ftp.uhttps://doi.o			CC BY 4.0
27	antismash	Database of secondary metabolite gene clusters	Technical University of Denmark (DTU) and University of Wageningen (Netherlands)	Processing of whole genome sequencing data	.tar.gz		GB		T3.2, T5.1	no standard for metadata		https://dl.sshttps://doi.o	open		

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

	type	description of data set	owner	source	formats	software	Max size	number	purpose of use in the MIBIREM project	meta data standard	storage/ repository	findable/ link	Publica-tion	dissemi nation/ accessi bility	Exploita-tion/ licence
27	eggnog-mapper	Database of functional annotations and orthology relationships	EMBL Heidelberg (Germany)	Processing of whole genome sequencing data	.db, .pkl, .d md		GB	3	T3.2, T5.1	no standard for metadata		http://eggn	https://doi.c	open	
28	CARD (comprehensiv e antibiotic resistance database)	Database of antibiotic resistance genes	McMaster University	Processing of whole genome sequencing data	.tar.bz2		MB		T3.2, T5.1	no standard for metadata		https://card	https://doi.c	open for academi c purposes	
29	AMRFinderPlus database	Database of antibiotic resistance genes	NCBI	Processing of whole genome sequencing data	.tar.gz		MB		T3.2, T5.1	no standard for metadata		https://www	https://doi.c	open	
30	Virulence Factor database (VFDB)	Database of virulence genes	Institute of Pathogen Biology (China)	Processing of whole genome sequencing data	.gz		MB		T3.2, T5.1	no standard for metadata		http://www	https://doi.c	open	
31															

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

8.2 Acronyms

CA	Consortium Agreement
DDBJ	DNA DataBank of Japan
DM/DMP	data management/ data management plan
ENA	European Nucleotide Archive
FAIR	findable, accessible, interoperable, reusable
GA	Grant Agreement
GDPR	General Data Protection Regulation (of EU)
IP	Intellectual property
INSDC	International Nucleotide Sequence Database Collaboration
KER	key exploitable results
NCBI	National center for biotechnology information
ROL	results ownership list
SRA	Sequence Read Archive (of NCBI)
WP	work package

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 [EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)]. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union