Data Management Plan

Deliverable 7.2

Author's name: Mercedes Beltrán (UU) and Corné Pieterse (UU)
Date: 04-07-2024

Document ID	**D7.2**
Due date	30 June 2024
Submission date	04 July 2024
Deliverable type	DMP
Dissemination level	Public
Work Package	WP7
Keywords	data management, FAIR data
Author(s)	Mercedes Beltrán (UU), Corné Pieterse (UU)
Contributors	Hanna Koch (AIT), Qi Wang (MPI), Julian Drausinger (LVA), Andrea Monti (UNIBO), Paul O'Toole (UCC), Clodagh Corcoran (ATL), Ariane Kaper (DAN-TR), Viggó Marteinsson (Matis)
Document version	**1.0**
Grant agreement	101083671
Duration	60 months
Start date	January 2024
End date	December 2028

Project information

Project number	**101083671**
Acronym	**MICROBIOMES4SOY**
Full name	Healthier diets and sustainable food and feed systems through employing microbiomes for soya production and further use
Project URL	https://www.microbiomes4soy.eu
EU project officer	Celine Doyendaz

Table of contents

Executive summary

This deliverable presents the Data Management Plan (DMP) for the Microbiomes4Soy project. The DMP addresses the relevant aspects of the management of data and other outputs in line with FAIR and ethical principles as outlined in the Microbiomes4Soy project plan. This DMP is a living document that will be regularly amended and updated.

1. Introduction

All staff of the Microbiomes4Soy project are individually responsible for the for the correct management of the research data according to institutional regulations, ethical and security policies, next to the guidelines established in the Microbiomes4Soy data management plan. In the case of a publication, researchers themselves are responsible for depositing the necessary data with the relevant bodies. The data summary table **[Table 1]** indicates responsible partners for data curation. A complete list with the institutional data management referents for the consortium partners can be found in **Annex 1.**

2. Data Summary

During the course of MICROBIOMES4SOY, different types of data will be generated related to microbiomes and their impact on their host along the food system. Furthermore, the project will collect information reflecting knowledge, needs and ambitions of relevant stakeholders. In addition, a prospective dietary intervention will be carried out (SOYBIOM study), collecting health data from human subjects in order to assess the effect of replacing animal protein with soya-derived protein on the human gut microbiome composition and function, and elaborated metabolome. **Table 1 (a - e)** presents an overview of the data generated in each work package.
In addition, the project will reuse previously collected data**.** AIT will compare sequencing data generated in WP2 with sequencing data collected in previous research. Reused data concerns public domain data, available in archives such as NCBI, and there are no foreseen restrictions to its reuse.

Data generated by this project is relevant the project partners and the broader scientific community, next to potential utility for the farming community (farmers and advisors), industry sector (biofertilizers/agro-input producers, food industry, feed industry, aquaculture), healthcare providers and general public. Project results from WP5 activities will be provided in a form of sensible advisory packages that can be utilised by enterprises and advisory groups. Towards the end of the project the data delivering groups will evaluate the data regarding their validity over time and their potential interest for the farming and food sector.

Table 1

Table 1a

WP1: Harnessing soybean microbiome potential for microbiome-optimized European soybean crop performance (Leader: UU)					
WP Tasks	**Data type**	**Data description**	**File Formats**	**Estimated data size**	**Data curation**
T1.1	Tabular/Numerical	Raw microbiome and microbial genome sequence data	.fastq	1 TB	UU
T1.1	Imaging	Raw root and shoot phenotype images	.tiff	100 TB	UU
T1.1	Tabular/Numerical	Processed microbiome and microbial genome sequence data	.xlsx; .csv	1 TB	UU
T1.1	Tabular/Numerical	Processed_ root and shoot phenotype images	.xlsx; .csv	50 TB	UU
T1.1	Tabular/Numerical	Field data (phenotypic scoring data)	.xlsx; .csv	0,1 TB	UU
T1.1	Tabular/Numerical	Survey on agricultural practices *	.xlsx; .csv	10 MB	UU
T1.1	Tabular/Numerical	Soil samples_ chemical composition	.xlsx; .csv	10 MB	UU
T1.1	Tabular/Numerical	(Recruitment information) Farm geolocation and certification for soil sampling	.xlsx; .csv	1 MB	DAN-TR / UU
T1.1	Physical samples	Soil samples	-	3000 kg	UNIBO/ UU
T1.2	Physical Samples	Glycerol stock of bacterial culture collection from soybean root samples	Cryo tubes	268 tubes	MPI
T1.2	Text	Dataset of culture collection	.xlsx; .csv	1 Mb	MPI
T1.2	Text	Dataset of phosphate-solubilizing bacteria screening	.xlsx; .csv	1 Mb	MPI
T1.2	Scripts/Code	R script for analysing the phosphate-solubilizing bacteria	.r	1 Mb	MPI
T1.3	Physical samples	Collection of bacterial strains	Cryo tubes	1000 tubes	UU
T1.3	Tabular/Numerical	Bacterial phenotyping	.xlsx; .csv	10 MB	UU
T1.3	Tabular/Numerical	Validation bioassays	.xlsx; .csv	10 MB	UU
T1.1 to T1.3	Scripts/Code	Scripts data analysis	.R/.py	10 MB	UU

Table 1_a. Overview of generated research data for WP1. A (*) indicates that personal data is collected.

Table 1b

WP2: Improving soybean production and seed quality by microbiome modulation under field conditions (Leader: AIT)					
WP Tasks	**Data type**	**Data description**	**File Formats**	**Estimated data size**	**Data curation**
T2.1 to T2.3	Tabular/Numerical	Seed phenotypic data	.csv	100 - 300 KB	AIT
T2.1 to T2.3	Tabular/Numerical	Microbiome amplicon sequencing data	.fastq .fasta	50 - 150 GB	AIT
T2.1 to T2.3	Tabular/Numerical	Processed microbiome amplicon sequence data	.csv	100 - 300 KB	AIT

T2.1 and T2.2	Tabular/Numerical	Microbiome metagenomic sequencing data	.fastq; .fasta	500 GB – 1TB	AIT
T2.1 and T2.2	Tabular/Numerical	Processed microbiome metagenomic sequence data	.csv	100 - 200 KB	AIT
T2.2	Tabular/Numerical	Strains genome data	.fasta	5 GB	AIT
T2.2	Tabular/Numerical	Processed strains genome data	.csv	100 KB	AIT
T2.1 to T2.4	Tabular/Numerical	Parameters surveyed during crop cycle	.xlsx; .csv	1 – 4 GB	UNIBO
T2.1 to T2.4	Multimedia (audiovisual)	Photos and video of the trials	.jpeg; .tiff	1 – 4 GB	UNIBO
T2.1 and T2.2	Physical Samples	Plant, root and seed samples	-	10 - 20 kg	UNIBO
T2.3 and T2.4	Physical Samples	Seed samples	-	10 - 20 kg	UNIBO

Table 1_b. Overview of generated research data for WP2.

Table 1c

WP3: Assessing the benefits of a plant-based protein diet for human health (Leader: UCC)					
WP Tasks	**Data type**	**Data description**	**File Formats**	**Estimated data size**	**Data curation**
T3.1	Tabular/numerical	Analysis values from sample analysis	.pdf; .xslx; .csv	100 MB	LVA
T3.2	Physical Samples	Gut Microbiome Composition	-	-	ATL/UCC
T3.2	Physical Samples	Stool meta Transcriptomics	-	-	ATL/UCC
T3.2	Physical Samples	Faecal Metabolome	-	-	ATL/UCC
T3.2	Physical Samples	Urine Metabolome	-	-	ATL/UCC
T3.2	Physical Samples	Serum Inflammatory Cytokines	-	-	ATL/UCC
T3.2	Tabular/Numerical	Lipid Profile	.csv	100 KB	ATL/UCC
T3.2	Tabular/Numerical	Blood Pressure	.csv	100 KB	ATL/UCC
T3.2	Tabular/Numerical	Percentage Product Compliance	.csv	100 KB	ATL/UCC
T3.2	Tabular/Numerical	Adverse Events	.csv	100 KB	ATL/UCC
T3.2	Tabular/Numerical	Vital Parameters	.csv	100 KB	ATL/UCC
T3.2	Tabular/Numerical	Serious Adverse Events	.csv	100 KB	ATL/UCC
T3.2	Tabular/Numerical	microbiome metagenomic sequencing data	.fastq .fasta	500 GB	AIT
T3.2	Tabular/Numerical	processed microbiome metagenomic sequence data	.csv	100 KB	AIT
T3.2	Tabular/Numerical	Analysis values	.xlsx; .csv	5 TB	UCC
T3.2	Tabular/Numerical	Analysis values	.xlsx; .csv	2 TB	UCC

Table 1_c. Overview of generated research data for WP3.

Table 1d

WP4: Effect of animal vs. plant-based protein diet on fish health & microbiome (Leader: MATIS)					
WP Tasks	**Data type**	**Data description**	**File Formats**	**Estimated data size**	**Data curation**
T4.1	Physical Samples	Collection of bacterial strains from ISCAR	Cryo tubes	40-80 tubes	MATIS
T4.2	Tabular/Numerical	Formulation of fish feed	.xlsx; .csv	1 MB	MATIS
T4.3	Tabular/Numerical	Raw microbiome and microbial genome sequence data	.fastq	1 TB	MATIS
T4.3	Tabular/Numerical	Processed microbiome and microbial genome sequence data	.xlsx; .csv	1 TB	MATIS

Table 1_d. Overview of generated research data for WP4.

Table 1e

WP5: Transition pathways towards a sustainable food system (Leader: AIT)					
WP Tasks	**Data type**	**Data description**	**File Formats**	**Estimated data size**	**Data curation**
T5.1	Text	Workshop documentation	.docx .pdf	500 MB	AIT
T5.2	Multimedia (audiovisual)	Webinars and factsheets	-	-	AIT
T5.3	Multimedia (audiovisual)	Webinars and factsheets	-	-	EQY
T5.4	Text	Workshop documentation	.docx .pdf	500 MB	AIT

Table 1_e. Overview of generated research data or other outputs for WP5.

3. FAIR Data

3.1 Making data findable, including provisions for metadata

Due to the diversity of the datasets generated by this project, data will be published using a multirepository strategy, opting for disciplinary repositories where available.

Datasets, protocols and scripts will be published via relevant disciplinary repositories according to the data types and will be uniquely and persistently identifiable. In cases where a DOI is not granted by the repository (e.g., Sequence Read Archive and European Genome-Phenome Archive), datasets are identified with an internal ID and a unique URL that resolves to the digital object.
To enhance findability, published (meta)data, protocols and software will be interlinked where relevant, making the related identifier explicit in metadata form. In addition, all persistent identifiers for datasets, publications and other outputs will be listed in the project website.

Published datasets will be accompanied by rich discoverability metadata according to international schemas [Table 2] to ensure datasets indexing and discoverability. In addition, specific keywords will accompany each published dataset to enhance semantic discoverability.

Project metadata will follow the Investigation/Study/Assay (ISA) tab-delimited (TAB) format, a general purpose framework with which to capture and communicate the complex metadata required to interpret experiments employing combinations of technologies, and the associated data files.

Table 2

Repositories and Metadata				
WP	**Datasets**	**Repository**	**Metadata standards**	**Access conditions**
WPs 1-4	Microbiome and microbial genome sequence data	Sequence Read Archive (SRA) of NCBI	Discoverability metadata will follow the SRA scheme (https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/). Data object level metadata: Sequencing data will be described with minimum information metadata standards developed by the Genomic Standards Consortium, such as minimum information about genomes (MIGS), metagenomes (MIMS), metagenome-assembled genome (MIMAG).	Open
WP1	Root and shoot phenotype images	European Genome-Phenome Archive (EGA)	Discoverability metadata will follow the EGA metadata schema based on the International Nucleotide Sequence Database Collaboration. During the project data object level metadata will comply with the Minimum Information standard for plant phenotyping standardization (MIAPPE - https://www.miappe.org/)[1].	Open
WP1	Stakeholders Survey Data	Yoda UU/ Dataverse NL	Discoverability metadata will follow Metadata Scheme Datacite 4.0	Restricted /Open*
WPs 1-2	Field data: Phenotypic scoring data	European Genome-Phenome Archive (EGA)	Discoverability metadata will follow the EGA metadata schema based on the International Nucleotide Sequence Database Collaboration. During the project data object level metadata will comply with the MIAPPE standardization.	Open
WP3	Microbiome data from dietary intervention	European Nucleotide Archive	Discoverability metadata will follow repository minimum metadata values in accordance with MIxS standards. In addition (meta)data will follow standards from the International Human Microbiome Standards consortium (IHMS).	Restricted : data openly available after a one-year embargo

[1] Taylor, C., Field, D., Sansone, SA. et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26, 889–896 (2008). https://doi.org/10.1038/nbt.1411

WP3	Microbiome data from in vitro colon model	European Nucleotide Archive	Discoverability metadata will follow repository minimum metadata values in accordance with MIxS standards. In addition (meta)data will follow standards from the International Human Microbiome Standards consortium (IHMS).	Restricted : data openly available after embargo
WP4	Aquaculture feed data	Matís /Laxá Dataverse	The formulations will follow the guidelines listed in: Nutrient Requirements of Fish and Shrimp. National Research Council. 2011. Nutrient Requirements of Fish and Shrimp. Washington, DC: The National Academies Press. https://doi.org/10.17226/13039	Restricted : data openly available after a five-year embargo
WP 1– 5	Scripts, models and code	Github and Zenodo	Discoverability metadata will follow the standards used by Zenodo: DataCite 4.3 and Dublin Core.	Open

Table 2: Repositories and metadata standards per data type for the Microbiomes4Soy project. A (*) indicates that personal data is collected. Dataset will be de-identified and aggregated before made open in compliance with the GDPR.

3.2 Making data accessible

As a guiding principle, in order to allow research dissemination and reproducibility, Microbiomes4Soy seeks to make all produced research datasets and software openly available and retrievable through free and standardized access protocols. Unless where stated otherwise in this DMP, datasets are made available immediately at the time of publication of public reports and scientific papers.

In this line, datasets will be published via relevant disciplinary repositories according to the data types [Table 2] and assigned an unique/persistent identifier. Alternatively, if no disciplinary repository is suitable, datasets will be published using generalist trusted repositories as Zenodo or via institutional repositories. Where relevant, protocols and reports will be published via Zenodo. Scripts, models and code will be made available in GitHub and published via Zenodo, obtaining therefore a DOI. Relevant machine-readable metadata will link different types of data and the tools needed to interpret the data.

Where datasets cannot be openly available due to compliance with data protection regulations (personal data, e.g., nutritional and dietary data and clinical data from WP3), technical and procedural safeguards will be in place to ensure protection of the data subjects. Whenever possible, data will be deidentified followed an anonymization plan and shared in a pseudonomyzed and aggregated format. Personal identifiers will not be shared. Should access to these datasets need be restricted, comprehensive project metadata will be made openly available, with information regarding the access procedure, to ensure data FAIRness.

In case of re-used data (raw data collected from public databases, such as NCBI), detailed provenance information will be provided, together with the specification of the sources where the original datasets are available.

Project documentation (e.g., recruitment information, workshops documentation) will have access restricted to project partners. Workshop documentation (WP5) will also be shared directly by workshops participants.

3.3 Making data interoperable

Microbiomes4Soy will ensure the interoperability of the data and code, following disciplinary metadata standards [See **Table 2**] at a data object level and accompanying datasets with comprehensive data documentation. Next to machine-readable metadata, all published datasets will be accompanied as a minimum by a README file, including a codebook, to promote re-use of the datasets and guarantee reproducibility of research results. In addition, a description of the software and tools required or recommended for data processing will be included in the documentation where necessary.
Open and preferred formats will be used for data sharing, reuse and preservation of data files following international recommendations of the Digital Curation Centre [see for details **Table 1**]. Where working data files use a proprietary file format, files will be migrated to a recommended format before preservation. If migration to an open format is not possible, a link to the software needed to open files will be included in the documentation, along with instructions to enable data reuse.
Where possible, controlled vocabularies and ontologies will be used to document the data, according to disciplinary standards (e.g., Open Biological and Biomedical Ontology Foundry, Standards from the International Human Microbiome Standards consortium).

3.4 Increase data re-use

Data will be accompanied by comprehensive documentation. README files will include data provenance information, naming conventions used and a codebook clarifying concepts and/or variables that are present in the datafiles. In addition, protocols for data acquisition will be published and linked to the corresponding datasets metadata. A general README template for the project can be found in Annex 4.
Published data will be licensed under a Creative Commons license (e.g., CC0/CC-BY), clarifying conditions for reuse.

All data processing will follow quality assurance procedures in line with the standard operating procedures (SOPs) and established protocols at each institution. Measurements will include technical and when feasible biological replicates, in addition to calibration curves and appropriate controls.
For WP3, SOPs from the IHMS (human-microbiome.org) are used. The clinical study (ATL) is aligned with the principles of ICH Good Clinical Practice and follows a Data Management SOP (See Annex 2 of deliverable 7.1 Ethics Plan). Dat collection takes place via an eCRF (electronic Case Report Form) which is under ongoing review via the query management system as flagged by preprogramed edit checks. The data management team of ATL performs periodic data reviews at set timepoints during the study to manually flag findings to the study team.
Received analysis values by LVA are elaborated under ISO17.025 accreditation.
The work package teams will have regular meetings to sustain quality benchmarks and to review the data and its interpretation.

4. Other research outputs

The management of other digital and physical outputs has been included in section 2 and 3 of this DMP. This includes the management of software, workflows, protocols, models, workshop documentation and physical samples.

5. Allocation of resources

All data underlying publications will be preserved for a minimum period of 10 years. Where feasible research data and other relevant project outputs will be made openly available is suitable repositories (see section 3. FAIR Data) and linked via the project website.
All staff of the Microbiomes4Soy project are individually responsible for the for the correct management of the research data according to institutional regulations, ethical and security policies, next to the guidelines established in the Microbiomes4Soy data management plan.
Foreseen costs for data storage and management are accounted for and will be covered by each responsible partner.

6. Data security

As stated above, members of the Microbiomes4Soy project will manage research data according to institutional regulations and security policies. For the duration of the project, researchers will ensure safe storage and backup of data.
All digital data, including raw and processed data will be stored on internal institutional servers or enterprise-standard cloud storage with regular backups (where possible in different locations) to prevent data loss.
All resources are managed in compliance with institutional security policies, regularly subjected to backup procedures and protected with institutional authentication credentials.
During the project access to data will be restricted to authorized members of the research teams.
Where datasets include personal data, additional safeguards will be in place to include privacy by design in the research process and in all used research tools, in compliance with international data protection regulations (GDPR) and institutional policies. [See also deliverable 7.1 Ethics Plan].

7. Ethics

Microbioms4Soy data, include data from human participants taking part of a dietary study (WP3), next to data from discussions (meetings, events, workshops) and utilisation of diverse questionnaires and surveys.

Microbiomes4Soy is committed to an ethical management of data in compliance with ethical standards and the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679).
Next to data-oriented strategies (data minimisation, separation, abstraction and hiding), procedural and structural safeguards will be in place to guarantee the protection of personal data.
All project activities, including data handling, will be undertaken only after approval from responsible bodies regulating the welfare of animals in research (WP4) or after approval by an Ethics Review Board in case of data derived from human subjects (WP3).
Further details can be found in deliverable 7.1 Ethics Plan, including its annexes:
Development of submissions to the IRB/IEC;
Data Management SOPs of the CRO Atlantia;
Participant information and informed consent form for dietary study;
Participant information and informed consent for events and surveys.

8. Other issues

Data management will be done following institutional regulations, ethical and data protection standards as indicated above.
This plan will be updated yearly.

Contact and details

Coordinator and WP7 Leader

First and last name	Angela Sessitsch
Organization / University	AIT Austrian Institute of Technology GmbH
Abbreviation	AIT
Phone number	+43 50550 3509
Mail address	angela.sessitsch@ait.ac.at

Annex

Participating Partners Microbiomes4Soy

Participating Partners and Data Management referents Microbiomes4Soy			
#	**Participating Organisation [Acronym]**	**Country**	**Institutional Data Management Referent [Name and contact information]**
1	AIT Austrian Institute of Technology GmbH [AIT]	AT	Hanna Koch [hanna.koch@ait.ac.at]
2	Universiteit Utrecht [UU]	NL	Corné Pieterse [c.m.j.pieterse@uu.nl]
3	University College Cork [UCC]	IE	Paul O'Toole [pwotoole@ucc.ie]
4	MATIS OHF [MATIS]	IS	Viggó Marteinsson [viggo@matis.is]
5	European Food Information Council [EUFIC]	BE	Darya Silchenko [darya.silchenko@eufic.org]
6	Alma Mater Studiorum - Università di Bologna [UNIBO]	IT	Andrea Monti [a.monti@unibo.it]
7	Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. [MPI]	DE	Qi Wang [qwang2@mpipz.mpg.de]
8	Fodurverksmidjan Laxa HF [LAXA]	IS	Gunnar Kristjánsson [laxa@laxa.is]
9	Donau Soja Gemeinnützige Gesellschaft mit beschränkter Haftung [DS]	AT	Jasmin Karer [karer@donausoja.org]
10	LVA GmbH [LVA]	AT	Julian Drausinger [julian.drausinger@lva.at]
11	Euroquality SAS [EQY]	FR	Antonia Nikolova [antonia.nikolova@euroquality.fr]
12	Agrifutur SRL [AGF]	IT	Roberto Kron Morelli [rkm@agrifutur.com]
13	DANTRADE BV [DAN-TR]	NL	Ariane Kaper [ariane.kaper@danone.com]
14	Plant Republic GmbH [PR]	AT	Catharina Werl [Catharina.Werl@plant-republic.eu]
15	Atlantia Food Clinical Trials Limited [ATL]	IE	Clodagh Corcoran [ccorcoran@atlantiatrials.com]
16	Institut National De La Recherche Scientifique [INRS]	CA	--
17	Michigan State University [MSU]	US	--

Project: Microbiomes4Soy
Title Dataset :
Author:
Author Affiliation:
Date data collection:
Date publication of data package:

General
Brief description of the dataset.

Folder directory
Explanation on folder structure [.....]

File formats
Indicate what type of file formats are encountered in the dataset_[.....]_

Software required
Indicate whether there is specific software required to open files in the dataset_[.....]_

Abbreviations
List of used abbreviations used in files, columns, and in filenames. Clarify where needed the naming convention of your files.

Codebook
Explain the concepts and/or variables that are present in the datafiles

Data provenance
Description on how the data was attained or which settings were used on equipment to attain the data. This should also include any processing of data that took place to attain the files.
Analysis scripts are well documented either by comments in the script or in an additional readme file in GitHub.

Data re-use
If not stated in the machine-readable metadata, add information on license for data reuse and data request procedure if needed.

Related outputs
If not stated in the machine-readable metadata, add links to other related datasets, protocols and outputs.