DOMINO

Funded by the European Union

UKRI

DELIVERABLE

D.3.1 – Data Management Plan

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060218

DOMINO

Data Management Plan

Deliverable Nr.	D3.1
Due date	August 2023
Submission date	August 2023
Deliverable type	DMP
Dissemination level	Public
Work package	WP 3
Author(s)	Valentin Loux, Hélène Chiapello

Document version	0.3		
Grant agreement	101060218	**Duration**	60 months
Start Date	March 2023	**End date**	February 2028

DOMINO

Contributors

NAME	ORGANISATION
Orla O'Sullivan	Teagasc
Gary Frost	ICL
Federica Pinto	Unitn
John Kenny	Teagasc
Jutta Roosen	TUM
Debora Serra	EUFIC
Raivo Vilu	TFTAK
Jekaterina Kazantseva	TFTAK
Hélène Chiapello	INRAE
Valentin Loux	INRAE

Revision history

VERSION	DATE	REVIEWER	MODIFICATIONS
0.1	07/11/23	Valentin Loux	First Draft
0.2	08/03/23	Valentin Loux	Second draft
0.3	08/31/23	Valentin Loux, Stéphane Chaillou, John Kenny	First public version

Table of content

1. Executive summary

Deliverable D3.1 is the Data Management Plan (DMP) for the DOMINO project (Grant Agreement No. 101060218), which is funded by the European Commission's Horizon Europe Research and Innovation action Programme. The DMP gives an overview of all datasets that are planned to be collected and generated in the DOMINO project and details how the data will be managed, exploited and made accessible for verification and reuse in the life-time of the project and beyond the end of the project (post February 2028).

DOMINO is a Research and Innovation action project running from March 2023 to February 2028. The DMP is a living document, this is the first version of the document, covering the data management for the first months of the project. It details principles of the project in data management and main data sets that will be generated. Subsequent versions of the DMP are planned at M24 (D3.3), M48 (D3.5) and M60 (final version, D3.9) to reflect the evolution of data management within the DOMINO project.

2.Introduction

2.1. DOMINO Project Overview

2.1.1. DOMINO Objectives

The overall aim of DOMINO is to attribute health benefits to traditional Fermented Foods (FF), whilst developing novel FF which address the changing societal demands.

First, we intend to demonstrate the health impacts of FF-based diet on a healthy population to establish how fermented food consumption shapes the gut microbiome and provides health benefits to consumers using a longitudinal human nutritional trial. We will use milk kefir, a traditional, widely consumed animal-based FF to perform this trial. In parallel, we will target adults suffering from metabolic syndrome to better focus on health biomarkers and establish if consumption of this FF leads to improved clinical outcomes in this cohort.

Second, we intend to demonstrate that innovative strategies to tackle sustainability and nutritional health can be developed through the reasoned design of tailor-made FF-derived microbial consortia. For this, six food case studies selected as representative of the wide diversity of plant-based FF prototypes will be developed. This aims to move toward environmentally-sustainable food sources, while maintaining a healthy nutritional diet in the transition from animal to plant-based products.

Finally, we aim at demonstrating that restoring trust in the food system is possible if relevant actors are actively engaged in the co-design and pilot production process. Several living labs will be set up to ensure citizen engagement and a targeted diversity of actors from the food system with the objective to co-create and validate that DOMINO outcomes are relevant to society. Furthermore, for pilot-scale and business exploitation purposes, several Small and medium enterprises (SMEs) with a proven record of innovation in the food sector are members of the consortium. The multi-actor approach will guide our activities, design, dissemination and communication, to engage actors in a targeted and interactive way that fosters dialogue. This strong societal engagement will improve the uptake of the project outputs. To reach the overall aim of DOMINO, the project has set the following objectives:

Objective #1: Establishing the impact of FF consumption in a healthy population and in people with metabolic syndrome.

Objective #2: Providing tailor-made microbial solutions to address the challenges associated with sustainable food production and healthy nutrition.

Objective #3: Articulating our scientific approaches around an integrated 'omics' strategy and computational biology modeling for a better understanding and prediction of the role of FFs in nutrition.

Objective #4: Organizing multi-actor engagement in a holistic context encompassing primary food production, processing, marketing, consumption and the role of society in several EU countries with their own characteristic food cultures.

2.1.2. DOMINO Concepts

DOMINO is built on the concept (Figure 1) that leveraging food microbial diversity can be a solution to achieving not only healthier diets but also creating innovative sustainable fermented foods.

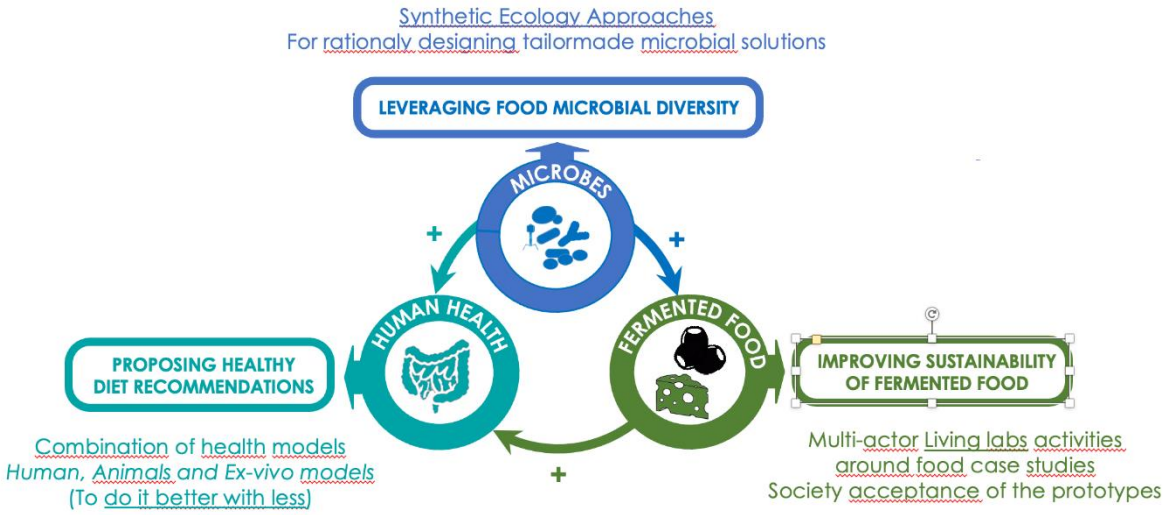

Figure 1 Concepts of DOMINO project

The DOMINO project is organized under 7 workpackages (WPs) detailed in figure 2.

WP1 : project coordination and management

WP2 : Health assessment of fermented food-based diet

WP3 : Open Food Microbiome Database (FMD) and computational tools

WP4 : Design of microbial solutions for sustainable and healthy fermented food prototypes

WP5 : Living Labs for healthy and sustainable diets

WP6 : Communication, Dissemination, Exploitation

WP7 : Ethics

WP2 to WP5 produce research data and are concerned by the DMP.

WP1 and WP7 produce documents destinated to communication (internal and external) and are concerned by the DMP and Communication,Dissemination, Exploitation Plan of DOMINO (deliverable D6.1)

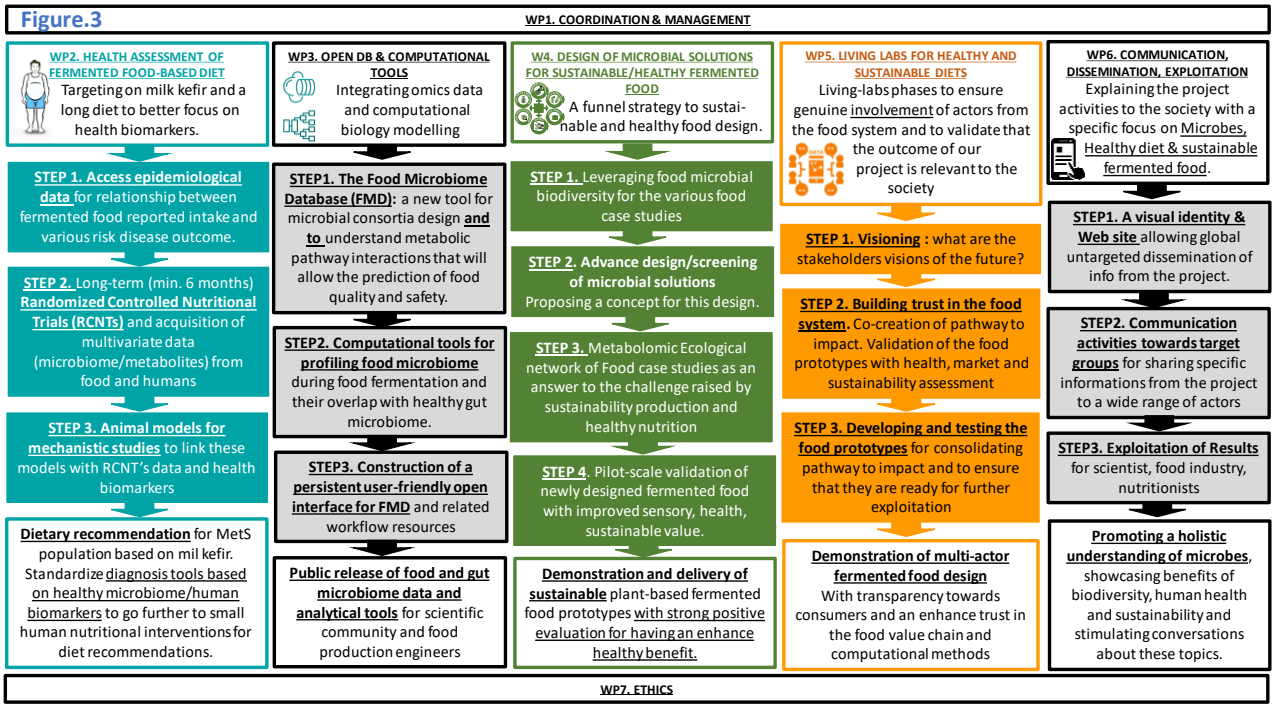

Figure 2 DOMINO work packages

2.2. DOMINO and Open Science

Annex 5 - Article 17 of the DOMINO Grant Agreement states the following concerning **Open Science and data management.**

Open science: open access to scientific publications

The beneficiaries must ensure open access to peer-reviewed scientific publications relating to their results. In particular, they must ensure that:

- at the latest at the time of publication, a machine-readable electronic copy of the published version or the final peer-reviewed manuscript accepted for publication, is deposited in a trusted repository for scientific publications

- immediate open access is provided to the deposited publication via the repository, under the latest available version of the Creative Commons Attribution International Public Licence (CC BY) or a licence with equivalent rights; for monographs and other long-text formats, the licence may exclude commercial uses and derivative works (e.g. CC BY-NC, CC BY-ND) and

- information is given via the repository about any research output or any other tools and instruments needed to validate the conclusions of the scientific publication.

Beneficiaries (or authors) must retain sufficient intellectual property rights to comply with the open access requirements.

Metadata of deposited publications must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent, in line with the FAIR principles (in particular machine- actionable) and provide information at least about the following: publication (author(s), title, date of publication, publication venue); Horizon Europe or Euratom funding; grant project name, acronym

and number; licensing terms; persistent identifiers for the publication, the authors involved in the action and, if possible, for their organizations and the grant. Where applicable, the metadata must include persistent identifiers for any research output or any other tools and instruments needed to validate the conclusions of the publication.

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)*
  Associated with document Ref. Ares(2022)8369065 - 02/12/2022
- *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 organizations and the grant. Where applicable, the metadata must include persistent identifiers for related publications and other research outputs.

DOMINO Communication,Dissemination, Exploitation Plan

Communication, Dissemination and Exploitation plan (CDE) of DOMINO (D6.1) presents a comprehensive plan of the communication, dissemination, and exploitation (CDE) activities planned for the DOMINO project, as well as the associated actions that will be carried out during the project. It also presents an operational process to define how the CE will be managed and implemented.

3. DOMINO Data Management

A data management plan (DMP) is a key element of good data management. Our DMP defines the specific data sets being collected, used and re-used during the DOMINO project and describes the life-cycle of the data.

To ensures the data is FAIR (Findable, Accessible, Interoperable and Reusable) our DMP includes information on:

- what data will be collected, processed and/or generated
- which methodology and standards will be applied, regarding formats, metadata, repositories
- whether data will be shared or made open
- the handling of research data during and after the project
- how data will be curated and preserved (including at the end of the project)

The DOMINO DMP answers questions of the Horizon Europe template model provided by the EC for Horizon 2020 funded projects.

The DMP presented here is a living-document. Several versions of the DOMINO DMP will be delivered during the project. A first version at M6 of the project, and revisions each year to reflect the evolution of data management within the project.

Representatives of all WP have contributed to and reviewed the DMP. Successive versions of the DMP have been submitted to the Executive Committee (ExCom) since 08/03/2023 after initial agreement on the process of validation of the document on 07/06/2023 ExCom meeting.

3.1. Research Outputs

This DMP describes the data that will be collected, used and reused during the DOMINO project, how it will be stored and preserved during and after the project's lifetime.

According to Horizon 2020 Online Manual [1]:

Research data refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation.

In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form.

Users can normally access, mine, exploit, reproduce and disseminate openly accessible research data free of charge.

In accordance with the second French plan for Open Science[2], we also have considered source codes, software and interfaces as research data.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm
[2] https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2021-10/second-french-plan-for-open-science-13715.pdf

Research Object Title	Research Object description	Concerned workpackage	Data Contact
Project Documentation	Internal and External DOMINO documents	WP1 and WP6	Agathe Renard, Debora Serra
Clinical Trials	Dietary records associated with microbiome metagenomics and metabolomics data	WP2	Orla O' Sullivan, Gary Frost
Food Microbiome Database Data Content	Previously generated data (MASTER project, public data) mixed with DOMINO data	WP2	Federica Pinto
Computational Workflows	Source codes used to analyze human data and data to construct microbial consortia and Metabolic ecological networks	WP2, WP4	Federica Pinto, John Kenny
Multi-omics Data collected from Human	Metagenomic and metabolomics data generated from the stool samples collected during the nutritional trial performed in WP2	WP3	Federica Pinto
Multi-omics Data collected from food	Metagenomic and metabolomics data generated from food samples related to the 6 food case studies of WP4	WP4	John Kenny
Food Microbiome Database Interface	Source codes of the FMD interface	WP3	Federica Pinto
Living Lab Manual	Manual with scripts and reporting templates for living lab sessions and results of rapid literature reviews.	WP5	Jutta Rosen
Sample & Strain data	Samples and new microbial strains related to the milk kefir and the 6 food	WP2, WP4	John Kenny

	case studies of the project		
Food models and data	Report describing tailor made microbial consortia and simplified food model for each food case study	WP4	John Kenny
Living Lab outputs	Reports of the six living-labs results regarding consumer acceptance	WP5	Jutta Rosen
Consumer Surveys	Results of consumer surveys in 6 countries	WP5	Jutta Rosen
Mintel Data	Data extracted from Mintel's Global New Product Database (GNPD) to evaluate current market developments and innovation potential of FF products	WP5	Jutta Rosen

3.2. Data Summary

- Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.
- What types and formats of data will the project generate or re-use?What is the purpose of the data generation or re-use and its relation to the objectives of the project?
- What is the expected size of the data that you intend to generate or re-use?What is the origin/provenance of the data, either generated or re-used?To whom might your data be useful ('data utility'), outside your project?

Research Object Title	Data summary
Project Documentation	Internal and external DOMINO documents for administrative, organization and communication purposes. Those documents can be in many formats : • text: txt, docx , odt or pdf, • images : png or jpg • web pages : html

	• presentations : ppt or odp • audio or video documents

Clinical Trials	Some DOMINO activities are building on existing public data and knowledge on the microbiomes associated with fermented foods with regards to human health. Using these pre-existing data will ensure further scientific merit is gained from the already generated data (generating more exploitable results).
	
	All pre-existing data are anonymized. Where required, ethics are in place from previous studies. The types of pre-existing data that will be used include:
	
	• Dietary records (e.g. from food frequency questionnaire, food diaries).
	• Microbiome and metabolome datasets
	• Subjects anthropometrics and basic metadata
	• Genomes and metagenome-assembled genomes from public repositories (e.g. NCBI, EBI, PATRIC).
	
	New data to be collected will be in digital and non-digital format and will be collected as administrative, technical and sequencing/metagenomic data. Our research data will exist in a number of states (e.g. raw, cleaned, processed, analysed, archived) and take a number of forms including:
	
	• Results of experiments, trials and studies in excel files and word files
	• Laboratory notebooks (e.g. details of experiments, measurements, etc)
	• Sequence data in the form of nucleotide/metabolomic sequences from:
	  • Shotgun metagenomics
	  • Metaproteomes
	  • Metabolomes
	  • Static intermediate files generated from the analyses (e.g., contigs from sequence assembly)
	  • Outputs from bioinformatic analyses
	  • Analytical pipelines
	  • Outputs from statistical analyses
	
	• Metadata from clinical studies
	
	Around 850 stool samples collected from three different locations will be sequenced for metagenomics analysis. Samples will also be analyzed for metabolic profiling, gut microbial metabolites generating around 6000 metabolomics datasets.

Food Microbiome Database Data Content	The Food Microbiome Database (FMD) will be built on: i) the curatedFoodMetagenomic data of MASTER project in which some of the partners have been involved in the last 4 years (over 2500 metagenomes), ii) publicly available data by an extensive survey of all the literature on fermented food microbes and microbiomes; ii) the sequencing data generated within DOMINO of isolates (around 230) and food microbiomes (around 200). All the metadata associated to the metagenomic samples will be also retrieved and curated in order to integrate into species-level genome bins all the genome-resolved data directly available in FMD based on a validated (meta)genomic assembly and binning approach. FMD will be open for the Consortium from the beginning and open access by the end of the project. The FMD is a database, file format used is tsv,csv. The expected size of the data depends on the extent and nature of the data generated and retrieved. This information will be updated as the project progresses.
Computational Workflows	DOMINO analysis workflows will be based on existing workflows already published by partners and public workflows and tools distributed under open-source licenses.
Multi-omics Data collected from Human	The metagenomic data will be generated from the stool samples collected during the nutritional trial performed in WP2. The trial consists in collecting around 850 stool samples in three locations Italy, France and UK. The trial aim is to identify metabolic and microbiome markers of health post milk kefir consumption. For more detail see Research Object Clinical Trials. The sequencing step will produce approximately 10 x 10^9 bases (nucleotides) of raw data (output: FASTQ file). Each sample is associated with metadata, which includes parameters of the volunteers (e.g. BMI, age, gender), dietary information for each adult volunteer, and information about health (e.g. previous pathology, chronic disease, any medical treatment). Metadata and samples are pseudonymized with a unique identifier (see below). The bridge between sample identifier and any other information about the Data Subjects (donator) will be kept on our secure server and only authorized researchers have access to. In brief, these pseudonymized data will be stored on our (University of Trento, INRAE, TEAGASC) secure and backed-up high-performance storage facility and will be uploaded as soon as they are analyzed to the public repository (NCBI SRA, ENA) in anonymized form. Metabolomics data will be collected in the form of CSV files.
Multi-omics Data collected from Food	The metagenomic, metatranscriptomic and metabolomic data generated from food samples related to the case studies of WP4 and from the pitched kefir used in WP2. WP4 has the objective to implement synthetic ecology experimental strategies for designing microbial solutions for six food case studies of the project. The

	considered food sources for the six case studies are: fermented cereals/legumes, water kefir, olives, fermented vegetables, apple pomace. The estimated number of samples are around 200 for metagenomes and around 230 isolates for genome sequencing. The estimated number of metatranscriptomes will be around 240 samples (around 40 samples per case study). The sequencing step for genome and metagenome will produce 5 to 10 x 10^9 bases (nucleotides) of raw data (output: FASTQ file). The sequencing step for metatranscriptome will produce 20 to 50 x 10^9 bases (nucleotides) of raw data (output: FASTQ file). The estimated number of samples is 2500 for metabolomics data. Metabolomics data will be collected in the form of CSV files. Each sample is associated with metadata, which includes parameters of the sources, fermentation procedure, location. Metadata and samples will be assigned with a unique identifier These data will be stored on our (University of Trento, INRAE, TEAGASC) secure and backed-up high-performance storage facility and by the single beneficiaries.
Food Microbiome Database Interface	The user-friendly interface to query the FMD database, and (re)launch analysis workflows will be developed as a set of Galaxy components and workflows. Galaxy is an open-source software (Academic Free License version 3.0).
Living Lab Manual	No existing data will be re-used. Living lab manual will be distributed as a document on the internal website of DOMINO.
Sample and strain data	Around 240 isolates will be sequenced for each of the six uses-cases. Each sample is associated with metadata, which includes parameters of the sources, fermentation procedure, location. Metadata and samples will be assigned with a unique identifier
Food models and data	No re-use of data is expected as input for WP4. The data generated as part of WP4 will be used to design the microbial consortia for the case study food products using standard genomics, metabolomics and microbiology information and data types that would be useful and interpretable by experts in the area of food and microbiome research.
Living Lab outputs	New data will be collected in living labs that are case-study focused (WP 4) and unique. They will involve interactions with stakeholders.

	Living lab output could be of different format: text documents, video, images, presentations.
Consumer Surveys	The project will conduct online surveys among consumers in project-related countries on FF acceptance and trust in the food system. The information will help to assess the market potential for FF in Europe and to build trust in the food system. We expect to collect between 2000 and 10000 responses on multiple questions. The questionnaires will be developed in close connection to the case studies so that the data is useful to generate knowledge, but likely of little interest for secondary use of the data.
Mintel Data	Most of the data is being newly retrieved from Mintel's Global New Product Database (GNPD) which is comprised of product (launch) data including product details like ingredients and FOP information. The data is downloaded from GNPD based on customized searches in the database. The format is a MS Excel file. Purpose of this generation is related to WP5 ("Evaluation of (MS3) and Report on current market developments and innovation potential of FF products" (D5.3)): By assessing product launch data conclusions can be drawn on, among others, marketing practices (FOP claims) as well as nutritional composition of the foods. Assessment of the data is to be done in a format suitable for statistical analysis (e.g., Stata). Generated and analyzed data may take up a few gigabytes. Data (and analysis) will be made available on aggregated level only, also due to licensing reasons for using the GNPD data.

3.3. FAIR Data

Making data findable, including provisions for metadata

- *Will data be identified by a persistent identifier?*
- *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.*
- *Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?*
- *Will metadata be offered in such a way that it can be harvested and indexed?*

Research Object Title	Making Data Findable
Project Documentation	The shared sharepoint website to store internal documentation is organized by workpackage, with a naming convention of the directories. Deliverables and milestones documents have to be

	stored on this website. A naming convention has been proposed to name those documents. Keywords and hashtags to facilitate indexing and discovery of external documents (site, tweets...) are described in DOMINO's CDE.
Clinical Trials	Metadata from clinical trials will include parameters of the volunteers (e.g. BMI, age, gender), dietary information for each adult volunteer, and information about health (e.g. previous pathology, chronic disease, any medical treatment).
Food Microbiome Database Data Content	Twill be made public in a repository. File nameMs will be unique, descriptive, naturally orderDed and consistent within WPs and should describe the project, file contents, date, researcher's initials and a version number. The earcher's initials and a version number. The wMDill include metagenomic samples with a unique identifiers and rich metadata associated to them. The metadata could contain description of the context of the data, quality and condition of the data included, place of origin of the samples. Ontologies developed in MASTER pre-existing project will be used and, if needed, extended. Further categories could be included. The dataset version will be documented along the project. Released version of software developed during the project will be will be versioned in a gitlab repository and archived archived in Software Heritage[3] for long term archiving and DOI assignment.
Computational Workflows	Released version of software developed during the project will be will be versioned in a gitlab repository and aarchived in Software Heritage for long term archiving and DOI assignment.
Metagenomics Data collected from Human	The sequencing datasets will be stored locally until the data and metadata are submitted to public repository (NCBI SRA or ENA) and made publicly available. Data that will be put in public repository include raw sequencing data without human DNA and metadata. In those repositories, both the overall dataset and the single samples are assigned a unique persistent identifier for the raw data and metadata. Such an identifier is searchable by metadata, submitter, and publication, and is also put in each publication making use of the data. The data and metadata will be separate files.

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

Metagenomics collected from food	The sequencing datasets will be stored locally until the data and metadata are submitted to NCBI and made publicly available, when possible. Data will be put in the FMD database (built within DOMINO) and in NCBI SRA or ENA (raw sequencing data without human DNA plus metadata). Doth the overall dataset and the single samples are assigned a unique persistent identifier for the raw data and metadata. Such an identifier is searchable by metadata, submitter, and publication, and is also put in each publication making use of the data. The data and metadata will be separate files. However, due to the nature of DOMINO (Research and Innovation Actions) not all data will be openly accessible due to IP concerns and potential for exploitation of results by the project partners. All beneficiaries are responsible for the data collected at their own site as part of the project and will make the required data accessible to all other partners, as required for the purpose of the project. Research data will be made public as soon as possible, subject to partner approval, as industry partners will want to commercialize certain aspects of the project results.
Food Microbiome DatabaseInterface	The interface source code will be versioned and stored in the INRAE GitLab instance. Release version of software developed during the project will be archived in Software Heritage for long term archiving and DOI assignment.
Living Lab Manual	
Sample and strain data	The data will have an identifier linked to the sample when made publicly available, prior to that the groups working on each case study will hold the information. No standard exists for this type of food case study. Metadata linked the sample will explain the source of the data. Spreadsheets including relevant sample metadata including Raw material, food processing description, name of matrix, date of sampling, fermentation time and temperature, sample storage conditions, physicochemical parameters, geographical information, date of analyses, type of analyses performed.
Food models and data	The data will have an identifier linked to the sample when made publicly available, prior to that the groups working on each case study will hold the information. No standard exists for this type of food case study. Metadata linked the sample will explain the source of the data. Spreadsheets including relevant sample metadata including Raw material, food processing description, name of matrix, date of sampling, fermentation time and temperature, sample storage conditions, physicochemical parameters, geographical information, date of analyses, type of analyses performed. All information will be stored in word or csv files.

Living Lab outputs	
Consumer Surveys	The data will be indicated in the publication with a persistent identifier. Metadata will include information on the data collection (timing, sampling etc.) according to consumer research standards
Mintel Data	Raw data from Mintel's GNPD is in MS Excel format (one line per product launch with one unique record ID provided by Mintel). Data structure and metadata is also from GNPD. However, analysis and publication will be done on an aggregated level only. For statistical analysis, manipulation of data as well as assessment will be transparent by utilizing do-files/log files.

Making data accessible : Repository

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Research Object Title	Repository
Project Documentation	Internal documents will be stored in an intranet sharepoint website operated by INRAE Deliverables and milestones documents will be deposited in EC website.
Clinical Trials	Personal Data collected during clinical trial will be stored on an eCRF (electronic Case Report Form) provided by INRAE, compliant to EU regulation (GDPR, Certified Health data Host).
Food Microbiome Database Data Content	The FMD will be available for all DOMINO Partners within the project platform and publicly available by the need of the project.
Computational Workflows	The workflow will be available on project code repository for members of the project until released openly.
Multi-omics Data collected from Human	The metagenomic data collected to human will be accessible to all partners as soon as they are produced. However, due to the nature of DOMINO project, the data will be publicly available upon the exploitation of results. The primary repository for the data and metadata will be the FMD database created within DOMINO and

	secondly public sequences repository (ENA ou NCBI SRA). Those repositories have the option to have the data stored private or publicly available but we will always make the data publicly available upon partners approval and as soon as the corresponding report is uploaded to accepted for publication at an open-access journal (notice that it is a requirement for all journals that the datasets of the publications are freely available). Metadata will be available except for the information that would permit de-anonymization of the subjects.
Multi-omics Data collected from Food	The metagenomic and metatranscriptomic data will be accessible to all partners as soon as they are produced. However, due to the nature of DOMINO project, the data will be publicly available upon the exploitation of results. The primary repository for the data and metadata will be the FMD database created within DOMINO and secondly public sequences repository (ENA ou NCBI SRA). Those repositories have the option to have the data stored privately or publicly available but we will always make the data publicly available upon partners approval and as soon as the corresponding report is uploaded to be accepted for publication at an open-access journal (notice that it is a requirement for all journals that the datasets of the publications are freely available). Metadata will be also publicly available. All the data stored in public repositories can be easily downloaded using the browser or using the software tools available such as "SRA Toolkit" [4]
Food Microbiome Database Interface	The source codes will be made accessible to the community at the end of the database development.
Living Lab	The living lab manual guides the establishment and running of the living labs in the different countries by case study leaders. The living lab manual will be published as deliverable
Sample and strain data	Each site involved in the data generation will hold the data for their case study in line with the local best practice of that Company/RPO using their servers, PCs and freezers. At the time of making the data public, subject to agreement by the associated groups involved, it will be uploaded to the appropriate open access repository (such as Zenodo[5]) and included in research publications. The data will be is assigned an identifier in the repository.
Food models and data	Each site involved in the data generation will hold the data for their case study in line with the local best practice of that Company/RPO

[4] https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/
[5] https://zenodo.org

	using their servers, PCs and freezers. At the time of making the data public, subject to agreement by the associated groups involved, it will be uploaded to the appropriate open access repository and included in research publications. The data will be is assigned an identifier in the repository.
Living Lab	The qualitative data will be highly context specific, regarding countries and case studies. Data will not be shared systematically. Informed consent will be obtained prior to the interviews. Collected data, interview transcripts, and recordings etc. will be stored on TUM servers. .
Consumer Surveys	Data will be deposited in a trusted repository after publication, e.g., at the TUM university library[6].
Mintel Data	The raw data will not be deposited in a trusted repository, due to licensing reasons for using the GNPD data

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.

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.

Will the data be accessible through a free and standardized access protocol?

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

How will the identity of the person accessing the data be ascertained?

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

Research Object Title	Data

[6] More information can be found here: https://www.ub.tum.de/en/research-data

Project Documentation	Internal documentation is, by definition, restricted in access to the members of the project. Public data available on the project website is accessible to all.
Clinical Trials	Personal Data collected during clinical trial will be stored on an eCRF provided by INRAE. This software is compliant with EU regulations concerning health and personal, mainly GDPR and certification associated with software security (Certified Health data Host). Access to health data will be limited to authorized person within partners collecting data through personal accounts on eCRF. Clinical trial promotor (INRAE) will be in charge of approving access to authorized members of the project. Metadata and samples are pseudonymized with a unique identifier. The bridge between sample identifier and any other information about the Data Subjects (donator) will be kept eCRF and only authorized researchers have access to. Only pseudonymized data will be shared between partners.
Food Microbiome Database Data Content	The CFD will be available for all DOMINO Partners within the project platform and publicly available by the end of the project. Due to the nature of DOMINO as an Innovation Action, not all data will be immediately openly accessible due to IP concerns and the potential for exploitation of results by the project partners. **Embargoed access to data**: Selected results, data and metadata will be accessible after a defined duration. **Limited accessibility to data**: remains to be defined as the DMP develops. **Data available on request**: remains to be defined as the DMP develops. **Closed access to data**: Results, data, and metadata of a sensitive nature in relation to IP, patenting, commercialization and exploitation will not be made openly available.
Computational Workflows	The workflow will be available on Github/Gitlab for members of the project until released under an open-source license.
Multi-omics Data collected from Human	The primary repository for the data and metadata will be public sequences repository (ENA ou NCBI SRA). Those repositories have the option to have the data stored privately or publicly available but we will always make the data publicly available upon partners approval and as soon as the corresponding report is uploaded to a pre-print server such as bioRxiv or is accepted for publication at an open-access journal (notice that it is a requirement for all journals that the

	datasets of the publications are freely available). Metadata will be also publicly available. All the data stored in public repositories can be easily downloaded using the browser or using dedicated software.
Multi-omics collected from food	The primary repository for the data and metadata will be public sequences repository (ENA ou NCBI SRA). Those repositories have the option to have the data stored privately or publicly available but we will always make the data publicly available upon partners approval and as soon as the corresponding report is uploaded to a pre-print server such as bioRxiv or is accepted for publication at an open-access journal (notice that it is a requirement for all journals that the datasets of the publications are freely available). Metadata will be also publicly available. All the data stored in public repositories can be easily downloaded using the browser or using dedicated software.
Food Microbiome Database Interface	A user agreement will be produced to address the question of privacy and restricted access of data.
Living Lab Manual	
Sample and strain data	Data will be made publicly available. Depending on the nature of the results generated, it may be that some approvals (e.g., intellectual property protections) need to be agreed upon by the groups working on a project and put in place prior to making the data open access.
Food models and data	Data will be made publicly available. Depending on the nature of the results generated, it may be that some approvals (e.g., intellectual property protections) need to be agreed upon by the groups working on a project and put in place prior to making the data open access. 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
Living Lab outputs	
Consumer Surveys	Data will be made openly available after publication of primary project results. Data will be freely accessible. As the data repository is open access, the identity of the persons accessing the data cannot be identified. There is no need for a data access committee.
Mintel Data	The raw data will not be openly available, also due to licensing reasons for using the GNPD data. However, aggregated and assessed data will

	be available throughout the project, e.g., through project reports and scientific publications.

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?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

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)?

Research Object Title	Metadata
Project Documentation	
Clinical Trials	Meta-data on clinical trials, excluding personal informations will be made available along publication.
Food Microbiome Database Data Content	
Computational Workflows	Computational workflows will be documented and opened with LICENSE and AUTHORS files.
Multi-omics Data collected from Human	As for the data, metadata will be available except the information that would permit de-anonymization of the subjects. The data and metadata will be separate files and linked by referring to the dataset's unique and persistent identifier in the metadata. Metadata will follow field-specific standards as established by the Genomics Standards Consortium (GSC) and the Minimum Information for Biological and Biomedical Investigations (MIBBI).
Multi-omics Data collected from Food	As for the data, metadata will be available except the information that would permit de-anonymization of the subjects. The data and metadata will be separate files and linked by referring to the dataset's unique and persistent identifier in the metadata. Metadata will follow field-specific standards as established by the Genomics Standards Consortium (GSC) and the Minimum Information for Biological and Biomedical Investigations (MIBBI).

Food Microbiome Database Interface	Interface source code will be will be documented and opened with LICENSE and AUTHORS files.
Living Lab Manual	
Sample and strain data	After agreement by the relevant parties involved, the metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement. It is expected that metadata will contain information to enable the user to access the data. The data will remain available and findable for as long as the repository is operational. This is expected to mean indefinitely, for example when using a resource like the European Nucleotide Archive or a journal website, but is outside the control of the consortium. It is envisaged that open access software options will be available to access or read the data. Any software used by the consortium for analysis will be described in the associated open access publications.
Food models and data	After agreement by the relevant parties involved, the metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement. It is expected that metadata will contain information to enable the user to access the data. The data will remain available and findable for as long as the repository is operational. This is expected to mean indefinitely, for example when using a resource like the European Nucleotide Archive or a journal website, but is outside the control of the consortium. It is envisaged that open access software options will be available to access or read the data. Any software used by the consortium for analysis will be described in the associated open access publications.
Living Lab outputs	
Consumer Surveys	Metadata will be included in the paper (method section) and be openly available. Data will be saved and uploaded in widely available software (csv) with appropriate documentation in pdf.
Mintel Data	The meta-data will not be openly available.

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?

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?

Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

Research Object Title	Make Data Interoperable
Project Documentation	
Clinical Trials	
Food Microbiome Database Data Content	The FMD will be built in order to be used on different systems and technologies. We will use common formats and standards and we will provide a project specific vocabulary and ontologies. The metadata will include reference/cross reference to other (meta)data to create as many links as possible between (meta)data resources, All datasets will be properly cited by including their globally unique and persistent identifiers.
Computational Workflows	
Multi-omics Data collected from Human	FastQ is the default format for shotgun metagenomic and metatranscriptomic sequencing and it is the standard format to allow interoperability (https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/). We will use the FastQ format as the default format for all our samples both locally and when deposited to the public domain. The metadata in NCBI SRA or ENA is made available using the vocabulary and the specification provided with the EXPERIMENT format (https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/).
Metagenomics collected from food	The data will be accessible and usable on different systems and technologies. DOMINO researchers will maximize the interoperability of the generated data to facilitate data exchange and re-use between researchers, organizations and countries. They will adhere to standards for formats and compliant with available software applications.
Food Microbiome Database Interface	Once released publicly, FMD will provide API (Application programming Interfaces) that permit access to public resources
Living Lab Manual	
Sample and strain data	As is standard in the field of research, strain and microbiology data will be recorded on csv format, using standard language of the field

	There should be no need for uncommon language in the production, genomics, or microbiology contexts. The publications from this project will refer to previously published research as required.
Food models and data	As is standard in the field of research, strain and microbiology data will be recorded on csv format, using standard language of the field. There should be no need for uncommon language in the production, genomics, or microbiology contexts. The publications from this project will refer to previously published research as required.
Living Lab outputs	
Consumer Surveys	Data will follow common standards in consumer research.
Mintel Data	Data will not be made interoperable due to the above limitations (licensing).

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.)?

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 licenses, in line with the obligations set out in the Grant Agreement?

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes.

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.

Research Object Title	Increase Data re-use
Project Documentation	Public website will promote DOMINO results and make link to Data and metadata produced within the project.
Clinical trials	Re-use of trials data will be restricted as described on ethics application of the trial.
Food Microbiome Database Data Content	MDAs soon as the data will be released with full open access upon deposition of a report in pre-print server such as bioRxiv or acceptance for publication in a scientific journal, the CFD will be accessible to be re-used by the community. CFD is the effort from the

	project to make available the profiled samples with metadata in the easiest and most re-usable way. In the package our data can be analyzed and re-used in the framework of many other available datasets that we put in the same DB. The will remain re-usable for over 10 years beyond the end of the project.
Computational Workflows	Workflows will be publicly available under an open-source license. It will be documented and fully re-usable.
Multi-omics Data collected from Human	The data will be usable on different systems and different technologies. Data will be readable for machines without the need for specialised algorithms, translators or maps. As soon as the data will be publicly available, there will be no embargo and all our data will be released with full open access upon acceptance for publication in a scientific journal. NCBI SRA is mirrored to EBI and both repositories are keeping the data available indefinitely. Data uploaded into NCBI SRA passes several quality control steps and data on the technology that produced them is stored so the data can be re-used in the long term.
Multi-omics Data collected from Food	The FMD database has the aim to make available the profiled samples with metadata in the easiest and most re-usable way. Once agreement is reached the data will be published in open access journals and repositories, except in case IP, patenting or commercialization sensitivities. When the DOMINO partners wish to protect their deposited research data they can do so with the policy of the repository to where the data is being submitted or with one of the appropriate Creative Commons licences for other research data. The data will remain re-usable for >10 years beyond the end of the project.
Food Microbiome Database Interface	Source code will be publicly available under an open-source license. It will be documented and fully re-usable.
Living Lab Manual	
Sample and strain data	All data will be made publicly available on open access data repositories and through publication in open access journals. This will detail the provenance of the materials and data involved, permitting use by anyone wishing to access the data, and in line with FAIR principles.
Food models and data	All data will be made publicly available on open access data repositories and through publication in open access journals. This will detail the provenance of the materials and data involved, permitting

	use by anyone wishing to access the data, and in line with FAIR principles.
Living Lab outputs	
Consumer Surveys	Data will be available in the public domain. Some functionalities (i.e., linking variable definitions to variable labels) will require appropriate commercial statistical software. The code for analysis as used for publication of results will be associated with the data.
Mintel Data	Methodology, documentation etc. will be described in respective project reports. Any data manipulation/assessment is validated by at least two researchers.

3.4. 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.).

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 re-use, in line with the FAIR principles.

Research Object Title	Other research outputs
Project Documentation	
Clinical Trials	
Food Microbiome Database Data Content	
Computational Workflows	
Multi-omics Data collected from Human	
Metagenomics collected from food	

Food Microbiome DatabaseInterface	
Living Lab Manual	
Sample and strain data	Biological samples used for generation of digital data (e/g, sequence or metabolomics data) will be stored in the research labs for a period of time, but due to degradation associated with biological samples will not be housed indefinitely as they will cease to be useful. SOPs generated for the project will be made publicly available, as part of publications or on the DOMINO website allowing free access to any interested parties.
Food models and data	Biological samples used for generation of digital data (e/g, sequence or metabolomics data) will be stored in the research labs for a period of time, but due to degradation associated with biological samples will not be housed indefinitely as they will cease to be useful. SOPs generated for the project will be made publicly available, as part of publications or on the DOMINO website allowing free access to any interested parties. Specific combinations of microbes will be tested and approved for the generation of the foods in the different case studies. Information on those communities will be disclosed as part of the publications.
Living Lab outputs	
Consumer Surveys	
Mintel Data	

3.5. 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.)?

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)

Who will be responsible for data management in your project?

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)?

Research Object Title	Allocation of resources

Project Documentation	Cost associated with internal (sharepoint) or external (web site) documentation are covered by the project. Website will stay online 3 years after the end of the project (March 2031).
Clinical trials	Cost associated with eCRF, clinical trials data archiving are covered by the project.
Food Microbiome Database Data Content	The FMD will be made public in free repository, as OSF or GitHub. In case we will have sensitive data and any sequence data or metadata that we cannot make freely available (we are not foreseeing such cases), we will keep them at one of our high performance storage infrastructure of one of the partner (University of Trento, INRAE or Teagasc) which is redundant, mirrored, and automatically backed up in a distinct location. This data will be kept on the Linux filesystems with the permits allowing the access to the data only to the project partners. Costs related to open access to research date in Horizon Europe are eligible for reimbursement during the duration of DOMINO. Funding for open access publication of scientific manuscripts generated from the results of the MASTER project is already accounted for in the budget.
Computational Workflows	The source code of workflows will be made available in a free repository (Github, Gitlab).
Multi-omics Data collected from Human	Our main data repositories (NCBI SRA or ENA) , are fully free and does not have any temporal limitation on the availability of the data. Those repositories have robust internal backup and recovery infrastructure. ENA and NCBI are mirrored one with the other. We can thus assume that the deposited data is extremely safe. In case there are sensitive data and sequence data or metadata that cannot make freely available (we are not foreseeing such cases), They will be stored internally on high performance storage infrastructure at the University of Trento, INRAE and TEAGASC. DOMINO has been allocated budget for open access publication.
Multi-omics Data collected from Food	Our main data repositories (NCBI SRA or ENA) , are fully free and does not have any temporal limitation on the availability of the data. Those repositories have robust internal backup and recovery infrastructure. ENA and NCBI are mirrored one with the other. We can thus assume that the deposited data is extremely safe. . In case we will have sensitive data and any sequence data or metadata that we cannot make freely available (we are not foreseeing such cases), we will keep them at one of our high performance storage infrastructure of one of the partner (University of Trento, INRAE or Teagasc) which is redundant, mirrored, and automatically backed up in a distinct location. DOMINO has been allocated budget for open access publication.

	Each partner associated with a case study will ensure appropriate data management at their location in line with the principles described in the DMP.
Food Microbiome DatabaseInterface	The source code of interface will be made available in a free repository (Github, Gitlab).
Living Lab Manual	
Sample and strain data	Costs for ICT infrastructure for the project will be covered by overheads from the grant budget. Each RPO will be responsible for the data at their organisation prior to being made public. As described above, long term data storage will be ensured by depositing the data on open access sites such as the ENA, or through publication in open access journal. Local shorter and interim level storage prior to making the data public will be the responsibility of each RPO, and must align with their local rules set in place by their ICT departments.
Food models and data	Costs for ICT infrastructure for the project will be covered by overheads from the grant budget. Each RPO will be responsible for the data at their organisation prior to being made public. As described above, long term data storage will be ensured by depositing the data on open access sites such as the ENA, or through publication in open access journal. Local shorter and interim level storage prior to making the data public will be the responsibility of each RPO, and must align with their local rules set in place by their ICT departments.
Living Lab outputs	
Consumer Surveys	Personnel costs in the form of 2 person days.
Mintel Data	No costs are associated with this data generation/assessment. Responsibility for the data management lies within this WP. (Raw) data will be held available throughout the duration of the project as well as afterwards.

3.6. 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)?

Will the data be safely stored in trusted repositories for long term preservation and curation?

Research Object Title	Data security
Project Documentation	Internal documentation is stored on INRAE's secure and backed-up storage facility. Website is secured and backed-up.

Clinical Trials	The eCRF provide by INRAE is labelled "Certified Health data Host".
Food Microbiome Database Data Content	FMD will be stored on University of Trento secure and backed-up high-performance storage facility.
Computational Workflows	Computational workflows will be stored on Gitlab server.
Multi-omics Data collected from Human	Data will be stored and managed locally on the computers of researchers, students and other staff involved in the DOMINO project. Data will also be store locally on the beneficiary institutions storage infrastructure which are redundant, mirrored, and automatically backed up in a distinct location. Each beneficiary will be responsible for the data they produce within DOMINO.
Multi-omics Data collected from Food	Data will be stored and managed locally on the computers of researchers, students and other staff involved in the DOMINO project. Data will also be store locally on the beneficiary institutions storage infrastructure which are redundant, mirrored, and automatically backed up in a distinct location. Each beneficiary will be responsible for the data they produce within DOMINO.
Food Microbiome Database Interface	FMD interface will be stored on Gitlab server.
Living Lab Manual	
Sample and strain data	All RPOs have their own ICT security teams and protocols in place. These teams oversee the back-up, security and maintenance of data as part of their procedures. As described above, long term data storage will be ensured by depositing the data on open access sites such as the ENA, or through publication in open access journal.
Food models and data	All RPOs have their own ICT security teams and protocols in place. These teams oversee the back-up, security and maintenace of data as part of their procedures. As described above, long term data storage will be ensured by depositing the data on open access sites such as the ENA, or through publication in open access journal.
Living Lab outputs	
Consumer Surveys	The data is securely stored on TUM university servers.
Mintel Data	For the time being, data is stored on servers of Technical University of Munich, where general technical and organizational measures apply.

3.7. 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).

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

Research Object Title	Ethics
Project Documentation	
Clinical Trials	Ethics issues associated with clicial trials (data conservation, informed consent,...) will be detailed on ethic application. UK being not engaged with GDPR regulation anymore, a specific attention will be made to clarification of exchanges of material and data between EU and UK partners.
Food Microbiome Database Data Content	
Computational Workflows	
Multi-omics Data collected from Human	All the ethics and ethical documents related to the production of this research output refer to the Clinical trials research object.
Metagenomics collected from food	
Food Microbiome Database Interface	
Living Lab Manual	
Sample and strain data	No ethical issues are expected to be associated with the sample data for the food case studies under WP4. The strains used for the case study production will align with the Nagoya Protocol for biological materials.
Food models and data	No ethical issues are expected to be associated with the sample data for the food case studies under WP4. The strains used for the case study production will align with the Nagoya Protocol for biological materials.
Living Lab outputs	
Consumer Surveys	Approval by ethics committee of TUM will be sought. There is no ethics deliverable. Informed consent for data sharing and long-term

	preservation of pseudo-anonymized data will be included in questionnaires.
Mintel Data	For this data generation/assessment no ethical questions are concerned as only product (launch) data is being analyzed. Personal data is not concerned within this analysis.

3.8. 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 additional issues are foreseen concerning data management in the project.

A subsequent version of the DMP should take into account additional data that will be produced in task 2.4 (animal exprimentation) , 4.1, 4.3 and 4.5 (microbial diversity of soils, ex-vivo models).