
# DATA MANAGEMENT PLAN

[ 27-10-2022]

TRACE-RICE with Grant nº 1934, is part of the PRIMA Programme supported under Horizon 2020, the European Union's Framework Programme for Research and Innovation

PRIMA
PARTNERSHIP FOR RESEARCH AND INNOVATION IN THE MEDITERRANEAN AREA

http://trace-rice.eu

# TECHNICAL REFERENCES

| | |
|---|---|
| **Project Acronym** | TRACE-RICE |
| **Project Title** | |
| **Project Coordinator** | Carla Moita Brites<br>carla.brites@iniav.pt |
| **Project Duration** | September 2020 – August 2024 (48 months) |

| | |
|---|---|
| **Deliverable No.** | D7.3 |
| **Dissemination level*** | CONFIDENTIAL |
| **Work Package** | WP 7 - PROJECT MANAGEMENT |
| **Task** | T7.5 – Data management plan elaboration |
| **Lead beneficiary** | INIAV |
| **Contributing beneficiary/ies** | UNL-ITQB, M.Dynamics |
| **Due date of deliverable** | 28 February 2021 |
| **Actual submission date** | 27 October 2022 |

| HISTORY OF CHANGES | | | |
|---|---|---|---|
| Date | Beneficiary | Version | Change |
| 12 May 2022 | INIAV | | Deliverable 7.1 is delayed due to COVID pandemics problems, project coordinator requests the UNL-ITQB engagement. |
| 2 June 2022 | INIAV | | Initial Proposal of Deliverable Structure |
| 8 July 2022 | UNL-ITQB | | Proposal of BioData platform and MIAPPE model specification |
| 20 Sept 2022 | UNL-ITQB | V1 | Version 1 of the deliverable |
| 12 Oct 2022 | UNL-ITQB | V2 | Revised version (version 2) sent to Project Coordinator |
| 27 Oct 2022 | INIAV | Final | Final version approved by project coordinator |

# EXECUTIVE SUMMARY

The present deliverable D7.1 –Data Management Plan (DMP) – establishes the framework under which TRACE-RICE consortium will monitor/generate, process and collect data during the deployment, integration and after the project is completed. Current document corresponds to the output of BioData.pt Data Stewardship Wizard and a first version of TRACE-RICE DMP. Therefore, DMP is composed by preliminary information and frameworks that will be followed, in the near future, to collect and characterise project's data sets (which will be presented in final report), as well how the data will be exploited or made accessible for verification and re-use and how the data will be curated and preserved after the project is completed. This document has been built based on the Horizon 2020 DMP Template v2.0 which promotes the principle "as open as possible, as closed as necessary" together with the project ambition of "Open by Default". Moreover, the DMP details the standards and instruments that will be used, how the research data will be preserved and what data sets will be published for verification or reuse as open access.
The purpose of the DMP is to contribute to good data handling through indicating what research data the project expects to generate and describe which parts of the data that can be shared with the public. Furthermore, a SOP (Standard Operating Procedure) for file naming and shared spreadsheet denominated miappev1.1_rice.xlsx is already created with the sheets of MIAPPE (Minimum Information About a Plant Phenotyping Experiment) data model specification. MIAPPE incorporates several metadata standards and practices, in order to ensure interoperability and avoid remodelling and redefining aspects that are already well established. MIAPPE includes precise definitions and examples for each of its fields, with recommendations for the use of controlled vocabularies, ontologies and ISO norms whenever appropriate (e.g. ISO 8601 for dates, Crop Ontology terms in observed variable section and Plant Ontology terms for characterising samples). These definitions and recommendations promoting compliance with the FAIR principles. Each dataset will be given a persistent identifier (Digital Object Identifier, DOI), supplied with relevant metadata and linked to the project name.

The TRACE-RICE DMP in annex describes:

- **Data Summary**: Overview of what data will be gathered and processed in the project
- How data will be stored and processed according to the H2020 **FAIR Data Management principles**, making data: finable, accessible, interoperable, and reusable. (FAIRness metrics).
- **Allocation of resources**: The costs and human resources of making data FAIR
- **Data Security**: How we intend to keep the data secure
- **Ethical aspects**: A summary of the ethics and privacy strategy in TRACE-RICE

However, the DMP must be consistent with exploitation and Intellectual Property Rights (IPR) requirements. Hence, research data linked to exploitable results will not be put into the open domain. Particularly, sensitive data provided by consortium partners for the demonstration scenarios and personal data of the users will be kept strictly confidential and anonymized, i.e., secured in a way to maintain compliance with General Data Protection Regulation (GDPR).

The DMP is a living document and will be updated at the end of the project to reflect the actual research data generated during the project and include updated instructions for how to access open data.

**ANNEX:**

First version of TRACE-RICE DMP, as preview output of BioData.pt Data Stewardship Wizard

# TRACE-RICE

## TRACE-RICE Consortium

Instituto Nacional de Investigação Agrária e Veterinária, I.P.

iBET Instituto de Biologia Experimental e Tecnológica

ITQB Universidade Nova de Lisboa

Casa do Arroz

ernesto morgad desde 1920

Mater Dynamics

BGI www.bgi.pt

Grupo Desarrollo

iata Instituto de Agroquímica y Tecnología de Alimentos

Alexandria University

Data Management Plan

# Trace-Rice

**Following the Horizon 2020 DMP Template v2.0**


Contact person: ***Pedro Sampaio*** *(pedro.sampaio@ulusofona.pt,*
*0000-0003-2917-4904)*
*Universidade Lusófona*


Based on: *Life Sciences DSW Knowledge Model, 2.4.0 (dsw:lifesciences:2.4.0)*

Project phase: *Before Submitting the DMP*


Created by: ***Ines Chaves*** *(ichaves@itqb.unl.pt)*
*ITQB NOVA/IBET*


Generated on: *27 Oct 2022*

*Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»*

# Projects

We will be working on the following projects and for those are the data and work described in this DMP.

## Tracing rice and valorizing side streams along mediterranean blockchain

Acronym: *TRACE-RICE*

Start date: *2020-09-01*

End date: *2024-08-31*

Funding: *H2020 Societal Challenges: Grant nº 1934, (call 2019, section 1 Agrofood) is part of the PRIMA Programme (granted)*

The overall objective of TRACE-RICE is to achieve a multidisciplinary critical mass to address pressing challenges to the Mediterranean rice sector by enabling the transfer of competences, technologies and organizational innovation among the partners. The work plan will facilitate adoption of cost-efficient and environmentally safe tools and technologies and promote innovative business models, improving quality, sustainability and opening new markets, with a focus on authenticity, for rice produced in the Mediterranean. Activities are focused on creation of expedite tools for rice authenticity, innovative solutions for contaminant mitigation, product traceability, conversion of by-products to innovative natural functional ingredients and creation of robust models from unstructured data. TRACE-RICE pays particular attention to implementation of RFID tags for wireless and real time supply chain integration, enabling a more sustainable use of resources and minimizing waste. The project will also target upscaling and replication of successful case studies in two different rural territories, taking account of their diversity. Liaison with standardization bodies will allow a large dissemination of validation data obtained by their inclusion in harmonized protocols. The dissemination and adoption of techniques along the rice Mediterreanean value chain will be achieved via a communication plan for the 3 countries involved and further linking to the European Federation of the sector for a wider dissemination. TRACE-RICE will identify business models with high potential for empowering rural communities to take advantage of the opportunities arising from improved rice value chain optimisation and new markets. It will directly support the creation of sustainable jobs and growth and in the long-term strengthen rural economic diversification, supporting Regional and Rural Development policy.

# 1. Data Summary

### Instrument datasets

The following instrument datasets will be acquired in the project:

- **Rice nutritional and organoleptic quality traits**

  This dataset will be collected by experts in the project, with our own equipment.

  The equipment is very well described and known.

  Other researchers working in the same field of research could be interested in using this data.

- **Rice whole-genome sequencing**

  This dataset will be collected by experts in the project, at a specialized infrastructure.

  The equipment is very well described and known.

### Non-equipment datasets

We also collect data from questionnaires. The non-equipment datasets are:

- **Sensory analysis with a trained panel of tasters** – Assessment of the organoleptic attributes of a product by the human senses.

### Data formats and types

We will be using the following data formats and types:

- **European Registry of Materials Identifier**

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

- **Informed Consent Ontology**

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

- **IUPAC Compendium of Chemical Terminology**

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

- **Minimum Information for Reporting Next Generation Sequencing Genotyping**

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

# 2. FAIR Data

## 2.1. Making data findable, including provisions for metadata

- **Rice nutritional and organoleptic quality traits** (published)
  The distributions will be stored in:
  - Domain-specific repository: dmportal.biodata.pt. We have already contacted the repository.
- **Rice whole-genome sequencing** (published)
  The distributions will be stored in:
  - Domain-specific repository: European Nucleotide Archive. We don't need to contact the repository because it is a routine for us.

  There will be different versions of this data over time; the versions will be numbered.
  We will be adding a reference to the published data to at least one data catalogue.

There are the following 'Minimal Metadata About ...' (MIA...) standards for our experiments:

- European Registry of Materials Identifier
- 
- Informed Consent Ontology
- 
- IUPAC Compendium of Chemical Terminology
- 
- Minimum Information for Reporting Next Generation Sequencing Genotyping
- 
- Minimum Information about Plant Phenotyping Experiment
- 

We will use lab notebooks to make sure that there is good provenance of the data analysis.

The provenance will be captured using W3C PROV.

We made a SOP (Standard Operating Procedure) for file naming. Unique and shared spreadsheet denominated miappev1.1_rice.xlsx is created with the sheets of miappe data model specification. We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

## 2.2. Making data openly accessible

We will be working with the philosophy *as open as possible* for our data.

The data cannot become completely open immediately because of:

- patent-related business reasons
- we want to publish a paper first

Data that is not legally restrained will be released after a fixed time period (Five years), unconditionally.

Metadata will be openly available. Metadata will available in a form that can be harvested and indexed (managed by the used repository / repositories).

We have a consortium agreement that arranges Intellectual Property.

For our produced data, conditions are as follows:

- **Rice nutritional and organoleptic quality traits** (published)
  The distributions will be stored in:
  - Domain-specific repository: dmportal.biodata.pt. We have already contacted the repository. It will be *Shared* with a predefined list of people.

  The dataset will published after all our processing has finished.
- **Rice whole-genome sequencing** (published)
  The distributions will be stored in:
  - Domain-specific repository: European Nucleotide Archive. We don't need to contact the repository because it is a routine for us. It will be *Open* (shared with anyone). The distribution will be available under the following license:
    - Freely available for any use (public domain or CC0).

  A user of this data can use it without any specific software.
  The dataset will published after an embargo.

## 2.3. Making data interoperable

We will be using the following data formats and types:

- **European Registry of Materials Identifier**

  It is a standardized format.

- **Informed Consent Ontology**

  It is a standardized format.

- **IUPAC Compendium of Chemical Terminology**

  It is a standardized format.

- **Minimum Information for Reporting Next Generation Sequencing Genotyping**

  It is a standardized format.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- **European Registry of Materials Identifier**
- **Rice Ontology**
- **Gramene Taxonomy Ontology**
- **ICIS Germplasm Methods Ontology**
- **Country and Location Ontology**

## 2.4. Increase data re-use (through clarifying licenses)

The metadata for our produced data will be kept as follows:

- **Rice nutritional and organoleptic quality traits** (published) – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.
- **Rice whole-genome sequencing** (published) – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.

As explained in Section 2.2, our data cannot become completely open immediately.

We will be archiving data (using so-called *cold storage*) for long term preservation already during the project. The data are expected to be still understandable and reusable after a long time.

To validate the integrity of the results, the following will be done:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors.
- We will run part of the data set repeatedly to catch unexpected changes in results.

# 3. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

We will be archiving data (using so-called 'cold storage') for long term preservation already during the project.

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication.

Ines Chaves is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Pedro Sampaio, Inês Chaves, and Maria Beatriz Teixeira Vieira are responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Pedro Barros and Maria Beatriz Teixeira Vieira are responsible for finding, gathering, and collecting data.

Pedro Barros, Pedro Sampaio, Inês Chaves, and Maria Beatriz Teixeira Vieira are responsible for maintaining the finished resource.

Pedro Sampaio and Inês Chaves are responsible for the management and proficiency of data including data processing, data policies, data guidelines, and data availability.

To execute the DMP, additional specialist expertise is required and we have such trained support staff available.

We do not require any hardware or software in addition to what is usually

available in the institute.

# 4. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers.They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The risk of information leak in the project or organization is acceptably low. The possible impact to the project or organization if information is vandalised is small.

We are not using any personal information.

The archive will be stored in a remote location to protect the data against disasters. The archive need to be protected against loss or theft. It is clear who has physical access to the archives.

We are running the project in a collaboration between different groups and institutes. A collaboration agreement that describes who can have access to what data in the project is set.

# 5. Ethical aspects

For the data we produce, the ethical aspects are as follows:

- **Rice nutritional and organoleptic quality traits**
  - It does not contain personal data.
  - It does not contain sensitive data.
- **Rice whole-genome sequencing**
  - It does not contain personal data.
  - It does not contain sensitive data.

**Data we collect**

We will not collect any data connected to a person, i.e. "personal data".

# 6. Other issues

We use the Data Stewardship Wizard with its *Life Sciences DSW Knowledge Model* (ID: dsw:lifesciences:2.4.0) knowledge model to make our DMP. More specifically, we use the https://biodata-pt.ds-wizard.org DSW instance where the project has direct URL: https://biodata-pt.ds-wizard.org/projects/3e9bafa2-01d1-4732-b5a2-db8250e59328.