PRO-GRACE

PROMOTING A PLANT GENETIC RESOURCE COMMUNITY FOR EUROPE

DELIVERABLE 7.4

Data management plan

Call identifier: HORIZON-INFRA-2022-DEV-01-01

PRO-GRACE

Grant agreement no: 101094738

Promoting a plant genetic resource community for Europe

Deliverable No. D7.4

Data management plan

Contractual delivery date:

M6

Actual delivery date:

M6

Responsible partner:

ENEA

Contributing partners:

IPK, WR, MPG, INRAE, IPGRI, MAICH

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094738.

Grant agreement no.	Horizon Europe – 101094738
Project full title	PRO-GRACE – Promoting a plant genetic resource community for Europe

Deliverable number	**D7.4**
Deliverable title	**Data management plan**
Type	DMP
Dissemination level	PU
Work package number	7
Author(s)	Giovanni Giuliano
Keywords	

The research leading to these results has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094738**.**

Index

1. Executive summary

This report describes the initial Data Management Plan (DMP) for the PRO-GRACE project, funded by the EU's Horizon Europe Programme under Grant Agreement number 101094738. The purpose of the DMP is to provide an overview of all datasets collected and generated by the project and to define the PRO-GRACE consortium's data management policy that will be used with regard to these datasets.

This DMP follows the structure of the Horizon Europe DMP template[1] with few modifications. It describes the types of data that is collected, processed or generated and the general principles and standards through which the data will be made Findable, Accessible, Interoperable, and Reusable and through which standards and repositories, as well as the policies for sharing the benefits deriving from these data with relevant stakeholders and protecting data privacy.

This initial version of the DMP defines the general policy and approach to data management in PRO-GRACE. A final version will refine and enhance policy aspects and will go into more detail regarding the datasets collected and produced.

2. Definitions and Acronyms.

CBD	Convention on Biological Diversity
dbSNP	NCBI Single Nucleotide Polymorphism Database
DEC	Dissemination and Exploitation Committee
DGVa	EBI Database of Genomic Variants archive
DISSCO	Distributed System of Scientific Collections
DMP	Data Management Plan
ENA	European Nucleotide Archive
EVA	EBI European Variation Archive
FAIR	Findable, Accessible, Interoperable, and Reusable
ITPGRFA	International Treaty on Plant Genetic Resources for Food and Agriculture
MIAPPE	Minimum Information About Plant Phenotyping Experiments
MINSEQE	Minimum Information About a Next-generation Sequencing Experiment
PGR	Plant Genetic Resources
RI	European Research Infrastructure
UPOV	International Union for the Protection of New Varieties of Plants

3. Introduction and Project background.

[1] https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf

Plants are the basis of all food, feed and renewable bioenergy production and are essential for the transition from a fossil-based to a bio-based economy. Plant genetic resources (PGR) play a key role in ensuring this transition, as well as food security and climate mitigation. More than 2 million plant accessions are preserved *ex situ* in 410 institutes in Europe and associated countries and listed in the EURISCO database; even more diversity is found *in situ* in European farmlands and wild habitats, where it contributes significantly to agricultural resilience and climate mitigation. Detailed information on *ex situ* accessions is, at best, fragmentary, while for *in situ* accessions it is almost non-existent. A considerable part of these resources could be lost over the coming decade due to limitations in the *ex situ* infrastructure and management, climate change, habitat loss, and invasive/alien species. The roadmap 2016 of the European Strategy Forum on Research Infrastructures (ESFRI) identifies a clear gap in the sector "Plant facilities – unlocking green power", i.e. the lack of a European Research Infrastructure (RI) specifically dedicated to PGRs. PRO-GRACE will undertake the first step to fill this this gap, by developing the concept of a novel (RI) dedicated to the conservation and study of PGRs. The concept will describe the proposed distributed structure, governance, economic plan and scientific services of the proposed RI, and will be the basis for a full proposal at the next ESFRI call. If implemented, this new RI will aim to catalog, describe, preserve and enhance European plant agrobiodiversity, and translate the results into conservation practices and agricultural innovation, and will collaborate with global organizations dedicated to PGR and with other established ESFRI RIs working on complementary fields. (eg ELIXIR, EMPHASYS, DISSCO, LIFEWATCH, MIRRI).

4. Aims and applicable principles

The aim of the DMP is to lay out a strategy and tools to make the project data and metadata **Findable, Accessible, Interoperable, and Reusable (FAIR)**[2] by project participants, stakeholders (e.g. plant scientists, seed companies, plant breeders, farmers, seed conservation networks, national and international agencies dealing with plant biodiversity, political decision makers) and, when appropriate, the general public.

Prepublication genomic data sharing has been recommended by the **Toronto statement**, in which authors of large-scale genomic studies are encouraged to deposit their data pre-publication on open access databases, retaining the priority right to genome-scale analyses[3]. The Toronto statement recommends extending this policy to other large-scale -omics datasets, such as Polymorphism

[2] Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9

[3] Toronto International Data Release Workshop Authors. „Prepublication data sharing." *Nature*, *461* (2009): 168-170.

Discovery, Genetic association studies, Somatic mutation discovery, Microbiome Studies, RNA profiling, Proteomic studies, Metabolomic studies, RNAi or chemical library screens, and 3D-structure elucidation.

Finally, in the case of PGRs, it is also important to **reconcile open data sharing with the equitable sharing of benefits deriving from their use**, in the frame of the Multilateral System of Access and Benefit Sharing under the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA)[4].

5. Data summary

The types of data managed by the PRO-GRACE project partners are illustrated in the table below.

Table 1. Types of data managed in PRO-GRACE and their main characteristics and curating partners

Data type	Permanent identifier	Public Repository	(meta)data format	Freedom to use?	WPs involved	Curating partners
DNA sequence	GenBank or ENA unique IDs	GenBank, ENA	Fastq, BAM, MINSEQE compliant	Yes	3	ENEA, INRAE, IPK, CREA, CSIC, UNITO, WR
DNA sequence variation	dbSNP, DGVa, EVA unique IDs	dbSNP, DGVa, EVA	VCF, Minseque compliant	Yes	3	ENEA, INRAE, IPK, CREA, CSIC, UNITO, WR
Metabolomic	Metabolights unique IDs	Metabolights (EMBL-EBI)	Various (depending on technology used), to be defined	Yes	3	MPG, WR
Passport	DOI recommended for accessions	Local Genbank databases, EURISCO, GENESYS	EURISCO Multi Crop Passport Data compliant exchange format	Yes	1,2	Genebank curators
Images	To be decided	Local Genebank databases, DISSCO Database	high quality jpg/png/Raw, to be defined	Yes	1,2,4	Genebank curators, *in situ* conservation networks
Phenotypic	Recommended for datasets	EMPHASIS nodes recommended databases	MIAPPE compliant	Yes	4	INRAE, CREA, IPGRI

[4] https://www.fao.org/plant-treaty/en/

Text	yes	Zenodo, BiorXiv	Accessible PDF	Yes	All	All partners
Software	Yes	Zenodo, GitHub, Sourceforge	Various	Yes	1,3	IPK, UniTO, ENEA, MPG
In situ / on-farm curatorial data	To be established	EURISCO	To be established	Yes	1,2	IPGRI, UOB, IPK, RSR, PSR

6. FAIR data

6.1 Making data findable, including provisions for metadata

- Efforts will be undertaken to assign globally unique and persistent identifiers to data and metadata. This is already routine for DNA sequence, DNA sequence variation, metabolomic and text data deposited in permanent public databases, while it must still be partially implemented for passport, image and phenotypic data associated with single accessions.
- Standards for associating rich metadata to the existing data will be elaborated, especially for passport, image and phenotypic data.
- Metadata will clearly and explicitly include the identifier of the data they describe
- (Meta)data will be registered or indexed in searchable online platforms.

6.2 Making data accessible

- (Meta)data will be retrievable by their identifier using a standardized communications protocol, which is open, free, and universally implementable. This is already routine for DNA sequence, DNA sequence variation, metabolomic and text data deposited in permanent public databases, while it must still be partially implemented for passport, image and phenotypic data associated with single accessions.
- The protocol will allow for an authentication and authorization procedure, where necessary
- The data will be deposited into an appropriate, publicly accessible repository (see Table 1). In the case of data for which such a repository is not defined yet, a prototype database will be proposed for implementation as part of the future GRACE-RI.

6.3 Making data interoperable

(Meta)data will use a formal, accessible, shared, and broadly applicable language for knowledge representation (See Table 1).

- (Meta)data will use suitable ontologies and standard vocabularies that follow FAIR principles
- (Meta)data will include qualified references to other (meta)data

6.4 Increase data reuse

- (Meta)data will be richly described with a plurality of accurate and relevant attributes
- (Meta)data will be released with a clear and accessible data usage license
- (Meta)data will be associated with detailed provenance
- (Meta)data will meet domain-relevant community standards

7. Benefit sharing

In order to reconcile the Open and FAIR data management with sharing of benefits deriving from their use, PRO-GRACE will implement the following actions:

- First, identify the **relevant stakeholders involved in the data ecosystem**, including data providers, users, and affected communities, understand their interests, needs, and concerns regarding open data and benefit sharing, and **involve them** in the decision-making process regarding data sharing through the project workshops, where diverse voices can be heard and considered.
- Create **robust governance mechanisms,** proposed by the project's **Steering Committee and the Dissemination and Exploitation committee** and approved by the **General assembly**, that will define the rules and principles for open data sharing. These frameworks will explicitly address the issue of benefit sharing, outlining how benefits derived from the data will be distributed among stakeholders.
- Promote **data literacy and capacity building** among all stakeholders, through the project workshops and training schools, ensuring that stakeholders acquire the knowledge and skills to understand the value of data and advocate for their fair share of benefits.
- Emphasize the **social impact and public interest in the use of open data**.
- Implement **safeguards for privacy and data protection,** to not compromise individuals' privacy and data protection rights.

8. Other research outputs

Experimental protocols, software: These will be published in refereed, open access, indexed journals. The software and codes will be made available through open access repositories like SourceForge[5].

New plant varieties: These will be prioritized for IP protection, preferably through the UPOV route[6]. When such varieties are generated using materials protected by the CBD or the ITPGRFA, clear

[5] https://sourceforge.net/
[6] https://www.upov.int/portal/index.html.en

guidelines for sharing the economic benefits arising from the exploitation of such varieties will be put in place by the DEC.

9. Allocation of resources

All genebank partners have institutional funding for generating, storing and updating the passport data associated with their holdings. Novel data generated by the project will mainly consist in -omics (genomic, metabolomic, DNA sequence, phenotypic, images, scientific publications). Enough financial resources have been allocated on each partners' budget to make these data openly accessible (including open access publications) and/or deposit them in open access, secure, permanent repositories. For data for which such accepted standards/repository does not yet exist (e.g., phenotypic data, images), at first the creation of common standards/repositories will be discussed with the relevant, exiting RIs (ELIXIR, EMPHASIS, DISSCO). If necessary, the custom gateways provided by OpenAIRE[7] will be used.

10. Data security

The security of project data is primarily concerned with preserving data from accidental deletion or loss, as well as preventing unauthorized access.

Risk Assessment:
The majority of data types will be stored in digital format. The following are some of the typical risks associated with this situation:
File loss due to accidental deletion or hardware damage
Unauthorized access to specific datasets that may be considered "access limited"

Data Security Measures:
Each type of data will be stored in a specific public online database. Most of them already implement security measures such as: Cloud storage with redundancy;; Controlled data access, in some cases security is enforced using required login credentials, specific user access-control lists and encrypted connections.

11. Ethics

No personal, or otherwise sensitive data will be generated/managed in PRO-GRACE. The issue of benefit sharing has been already discussed in section 7. In the case of management of data regarding rare, endangered genotypes growing in the wild, adequate measures will be put in place to protect/anonymize the exact geographical location of such genotypes, thus preventing illicit collection/biopiracy.

12. Other Issues

Not applicable.

[7] https://catalogue.openaire.eu/service/openaire.research_community_dashboard/overview