CoBreeding Plants DMP

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - DMP template (english)" fourni par Agence nationale de la recherche (ANR).

Plan Details

Plan title	CoBreeding Plants DMP
Deliverable	D0.1
Version	First version
Fields of science and technology (from OECD classification)	Agriculture, forestry, and fisheries
Language	eng
Creation date	2023-03-29
Last modification date	2023-06-05

Project Details

Project title	Co-design of animal and plant breeding schemes with multi-performance objectives (economic, social and environmental) to develop agroecological production
Acronym	CoBreeding
Abstract	New ways of mobilizing genetic diversity must be explored to respond better and faster to the challenges of the agroecological transition, climate change (adaptation, mitigation), global health and animal welfare, environmentally friendly plant cultivation practices, and consumer expectations. Breeding programs need to evolve so that larger numbers of diverse and complementary genotypes become available to fit the full range of current and future environmental conditions and production systems. To develop and implement new genetic improvement strategies, systemic approaches are needed to revisit the technical and socio-economic elements that determine the choices of populations/breeds/lines/varieties and individuals and their orientations in the different animal and/or plant production systems. The main levers are (i) the development of “co-design” approaches including the various actors concerned by these production systems to favor disruptive innovation for agroecology for building new social agro-ecosystems and defining innovating living entities ,(ii) the ability to evaluate multi-performance objectives (economic, social, environmental), (iii) diversification to improve the overall resilience of production systems, that imply to design new breeding strategies, but also new populations and farming practices, and to better characterize

genotype x environment x farming practices interactions. All these issues are common to the genetic improvement of animals and that of cultivated plants and will advantageously be addressed together in the CoBreeding project. To meet these challenges, the CoBreeding project is structured around three main axes. Axis A will put the emphasis on the process of collective design, considering heterogeneous actor communities. Axis B will concern the enrichment in multicriteria genetic evaluation methodologies, including environmental complexity. Axis C will focus on diversification and breeding strategies. By combining the results of experiments and field data, newly developed digital methods based on mathematics and computer modeling will highlight the synergies and antagonisms between the technical, economic, environmental and societal components of a system's performance. They will be used to optimize production systems and implement case studies for developing innovative lines/cultivars and associations for selected animal and plant species. Some data collection funded in project I.2 will be required to get the proofs of concept of the feasibility and impact of our innovations in breeding scheme design. We will also benefit from resources collected in flagship projects II.1, II.2 and infrastructures Liph4SAS and AGROECOPHENO to collect new genotypes and phenotypes. The objectives will be to co-produce knowledge and generic tools facilitating efficient and dynamic genetic management of livestock populations and crops, with the participation of selection operators, agricultural technical services, and direct or indirect users of the innovations produced (farmers, consumers, citizens) to meet the objectives of agroecological transformation. The CoBreeding project requires a funding of 3 M€ and brings together over 80 recognized specialists in biology (genetics, animal and plant sciences), data sciences (bio-informatics, mathematics, statistics) and social sciences from INRAE, INRIA, AgroParisTech, AgroCampus-Ouest, Université Paris-Saclay, Mines Paris - PSL, ENSFEA and Oniris who will devote 360 working months of permanent staff to the project.

Funding

- Agence Nationale de la Recherche : ANR-22-PEAE-0003

Start date 2023-01-01

End date 2027-12-31

Partners

- Institut national de recherche pour l'agriculture, l'alimentation et l'environnement https://ror.org/003vg9w96
- Institut national de recherche en informatique et en automatique https://ror.org/02kvxyf05
- université Paris sciences et lettres https://ror.org/013cjyk83
- Institut des sciences et industries du vivant et de l'environnement (AgroParisTech) https://ror.org/02kbmgc12
- Université Paris-Saclay https://ror.org/03xjwb503

Research outputs :

1. Simulation results output (Dataset)

2. Sequencing data output (Dataset)
3. Genotyping data output (Dataset)
4. Interaction matrix output (Dataset)
5. Observation and phenotyping data output (Dataset)
6. GWAS and genomic prediction results output (Dataset)

Contributors

Name	Affiliation	Roles
De Oliveira Yannick		- DMP manager - Personne contact pour les données (GWAS and GP, Sequences, Genotyping, Phenotyping, Interaction matrix, Simulation results)
Phocas Florence - 0000-003-1161-3665	INRAE	- Project coordinator

Droits d'auteur :

CoBreeding Plants DMP

1. Data description and collection or re-use of existing data

Simulation results output

1a. How will new data be collected or produced and/or how will existing data be re-used?

In task B1a, simulations will need soil characterization and historical climatic series datasets from existing databases. Data will be produced by an ecophysiological model.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Data will be collected in CSV text files. Results won't exceed some Mo

Sequencing data output

1a. How will new data be collected or produced and/or how will existing data be re-used?

In task B1b samples will be collected from field and AvrStb genes will be targeted sequenced.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Sequences are mainly text files in Fasta (other format ?) format.Volume ?

Genotyping data output

1a. How will new data be collected or produced and/or how will existing data be re-used?

In Task B2c a 245-offspring progeny of a polycross between 20 plants from three accessions and 1 200 plants from contrasted populations will be genotyped

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Text files in VCF format will be used

Interaction matrix output

1a. How will new data be collected or produced and/or how will existing data be re-used?

Interaction matrices will be acquired by performing disease tests under controlled conditions with wheat genotypes and Z. tritici strains representative of the haplotypic diversity at specific R-AVR couples

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Text files in CSV format

No more than 1Mo per matrice file.

Observation and phenotyping data output

1a. How will new data be collected or produced and/or how will existing data be re-used?

Phenotyping/Observation data will be collected in task B1c, B2c and C1c.

These data will be collected by observation or measuring devices at the field or technical rooms. Data can be recorded with electronic device.

In task A2b samples from previous projects (yet to be determined) could be used to validate model outputs.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Text files in CSV format.

The volume of this kind of data do not exceed 1Mo per file.

In task C1c, images from UAVs (unmanned aerial vehicles) will be produced in the TIFF/JPEG format. The volume of this kind of data is a few Gb per file. In total there will be a few To of such data.

GWAS and genomic prediction results output

1a. How will new data be collected or produced and/or how will existing data be re-used?

In tasks B2c and C1c, genome-wide association study will be performed on yield components as well as traits involved in competition for light (height and earliness)

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Data produced (GWAS results) will be text file in CSV format

2. Documentation and data quality

Simulation results output

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

The metadata will fit at minima the Standard Metadata template of RechercheDataGouv dataverse instance.

2b. What data quality control measures will be used?

In task A2b, simulation outputs will be controled by a combination of data to be acquired in the project from external samples, external data from previous studies and literature reviews, and expertise when no factual data can be found.

Sequencing data output

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Sequences will be documented as most as possible regarding the European Nucleotide Archive metadata model (https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html)

2b. What data quality control measures will be used?

Quality control tests are carried out after each stage of sequencing, as required by providers of kits and recommendations from sequencing apparatus

Genotyping data output

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Genotyping data will be documented with metadata as it is described in the VCF specification (https://samtools.github.io/hts-specs/VCFv4.3.pdf)

2b. What data quality control measures will be used?

Quality control tests are carried out after each stage of genotyping, from DNA extraction (control of DNA quantity and quality) to the final end of genotyping, depending on the methodology (array, GBS, etc).

Interaction matrix output

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

The metadata will fit at minima the Standard Metadata template of RechercheDataGouv dataverse instance.

2b. What data quality control measures will be used?

Need help here for quality control ...

Observation and phenotyping data output

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Phenotyping data will be formated according the MIAPPE standard (https://www.miappe.org/). Traits will be defined using ontologies.

The CropOntology (https://cropontology.org/) such as the Maize ontology (CO_322) and the wheat ontology (CO_321) will be evaluated and may be used to fit this purpose.

In task C1c, a set of metadata is automatically recorded for each session of UAV flight.

2b. What data quality control measures will be used?

Quality of data will be evaluated regarding established scales (it can be the ones defined in the CropOntology). Data will be scanned by a script using these scales for each Trait measured in phenotyping experiment. A report will highlight outlier data.

Use of validated software and/or procedures, comparison of several methods, application of stringent thresholds.

GWAS and genomic prediction results output

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

The metadata will fit at minima the Standard Metadata template of RechercheDataGouv dataverse instance.

GWAS results will also refer to input data such as genotyping, phenotyping datasets used

2b. What data quality control measures will be used?

Use of validated software and/or procedures, comparison of several methods, application of stringent thresholds

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during the research?

Data can be stored by three ways :

- on the NextCloud instance of INRAE. A specific CoBreeding space has been created. Metadata will be stored with the data file. The backup of data is ensured by information systems department at INRAE. The versioning system of NextCloud has been activated ensuring file versioning on 45 days.
- on local servers of partners, data backup will be ensured by local IT Team (will be the case for UAV images of task C1c)
- on the forge MIA hosting git repositories with authentication

3b. How will data security and protection of sensitive data be taken care during the research

- On NextCloud and forgeMIA security will be ensured by INRAE authentication system. Data will be shared with all CoBreeding partners.
- On local servers, data security will be ensured by local IT Team. CoBreeding partners can access to this data on demand.

4. Legal and ethical requirements, code of conduct

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

No personal data collected for this type of data.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

None

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

None

5. Data sharing and long-term preservation

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Data access will be limited to CoBreeding partners during the project.
Data will be shared on adapted data warehouse after the project in a time that remains to define.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Datasets will be archived on adapted data warehouse :
The European Nucleotid Arichive (https://www.ebi.ac.uk/ena/browser/home) for sequence data
The European Variant Archive (https://www.ebi.ac.uk/eva/) for genotyping data
The Recherche Data Gouv platform (https://recherche.data.gouv.fr/fr) for data that do not need specific data warehouse

5c. What methods or software tools are needed to access and use data?

All data will be formated with an open text format that do not need specific tools to be read

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

A DOI will be provided to all datasets published in RDG.

An URI will be assigned to datasets published in ENA and EVA. If needed these datasets can be referenced as external sources in RDG to generate a DOI for theses resources.

6. Data management responsibilities and resources

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Justin Blancon will be in charge of data produced in task B.1.a
Thierry Marcel will be in charge of data produced in task B.1.b
Eric Tannier will be in charge of data produced in task B.1.c
Bernadette Julier will be in charge of data produced in task B.2.c
Tristan Mary Huard and Alain Charcosset will be in charge of data produced in task C.1.a
Cyrille Saintenac will be in charge of data produced in task C.1.b
Timothée Flutre and Jean-Marc Gilliot will be in charge of data produced in task C.1.c
Sophie Bouchet will be in charge of data produced in task C.2.b

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Reproductible research methodological principles will be applied through the project by each of the members. If needed, training sessions to these practices will be proposed thanks to BreIF project in the PEPR AgroEcoNum.