European Plant Phenotyping Network

EPPN 2020

D3.1 Data management plan

P. Neveu, B Usadel, F Tardieu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731013. This publication reflects only the view of the author, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

Document information

EU Project N°	731013	**Acronym**	EPPN2020
Full Title	European Plant Phenotyping Network 2020		
Project website	www.eppn2020.plant-phenotyping.eu		

Deliverable	**N°**	D3.1	**Title**	**Data management plan**
Work Package	**N°**	3	**Title**	**Building a consistent Information System in the different nodes and defining standardisation strategies**

Date of delivery	**Contractual**		31/10/2017 (Month 6)	**Actual**	23/10/2017 (Month 6)
Dissemination level	X	**PU Public, fully open, e.g. web**			
		CO Confidential, restricted under conditions set out in Model Grant Agreement			
		CI Classified, information as referred to in Commission Decision 2001/844/EC.			

Authors (Partner)	Pascal Neveu (INRA), Bjorn Usadel (FZJ), François Tardieu (INRA)			
Responsible Author	**Name**	Neveu Pascal	**Email**	Pascal.neveu@inra.fr

Version log			
Issue Date	**Revision N°**	**Author**	**Change**
04/09/2010	1	Executive Committee	first version/
06/10/2017	2	All partners	Second version
23/10/2017	3	Executive Committee	Final version

Table of contents

Executive Summary

This document presents the plan for managing the datasets generated and processed during and after the Transnational Access (TNA) experiments of EPPN2020. The main objective is that datasets are findable, accessible, interoperable and reusable (FAIR standard), in such a way that the datasets collected in the TNA framework can be analysed by several groups inside and outside of the EPPN2020 consortium. The data management plan presents the different data categories, data sources and how data will be collected, structured, stored and made accessible for the purpose of analysis and reuse. A plan is presented for unambiguous identification of all objects involved in experiments (e.g. plants, sensors, cameras) in such a way that all these objects can be traced in further analyses. Data will be stored safely in servers with backups. Standards for data identification, file formats, and workflows are presented. The plan also presents the ownership of the different categories of data collected in a TNA experiment, with the rationale of optimizing analyses by involved groups (users and providers), and by other groups of the scientific community. Briefly, phenotypic data belong to users and environmental data, design procedures, workflows and calibrations belong to the provider.

This document has been written by the leaders of JRA3 ‘Building a consistent Information System in the different nodes and defining standardisation strategies’ and by the coordinator, based on principles that have been discussed in the consortium during the kick-off meeting. A first version has been amended by the Executive Committee and sent to the whole consortium. Interesting remarks have been suggested by members of the consortium, and integrated in the text.

The Data Management Plan (DMP) describes the management of datasets generated and processed during and after the Transnational Access experiments (TNA) of EPPN2020. This DMP will help partners to manage data, meet funder requirements, and facilitate multiple use of data by the scientific community. For that, it aims at the standard “findable, accessible, interoperable and reusable” (FAIR). The DMP includes information on data handling, data property and on the standards used for storing, curating and sharing data.

1. Data description

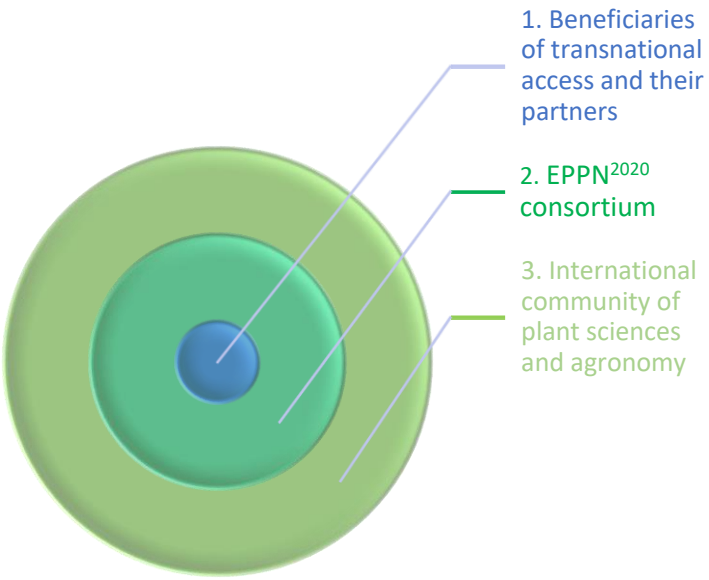
Data categories and sources that need to be managed are:

- Resource: species, genotype, seed origin, accession.
- Facilities: installations, sensors, cameras, vectors (e.g. conveyors), specific devices.
- Characteristics of experiments e.g. design, protocol and organisation.
- Phenotypic data at plant or population level, e.g. image-based traits, phenological stages, manual measurements.
- Environmental conditions as collected by sensors (e.g. soil water status, air temperature or evaporative demand).
- Date and description of management events
- Workflows: sensor and image analysis methods and software tools used to extract traits from raw image and other data.

Collected data can be numerical data, images, documents, texts or manual measurements. The data volume will be over hundreds of terabytes. In each experiment, data are collected for typical periods of 20-100 days, including raw data (sensor or camera outputs) and curated data (validated at further steps of analyses).

The project will work to reach the objective that, at the end of the project, data are interoperable between platforms and that querying mechanism are standardized, in the context of the Joint Research Activity 3 (JRA3) *‘Building a consistent Information System in the different nodes and defining standardisation strategies’*. All platforms will collect data in such a way that this is eventually possible, and tools will be progressively deployed in the project sites.

We distinguish three circles of users, namely beneficiaries of access and their partners, the EPPN2020 consortium and the international community of plant sciences and agronomy. The objective of EPPN2020 is that the three categories of users potentially have access to all datasets, with the necessary metadata and information for the datasets to be reusable. An intermediate objective is that some members of the EPPN2020 consortium can test the information system before giving it a larger diffusion, in a joint effort with the project EMPHASIS PREP.



The scientific leaders of phenotyping platforms will be responsible for managing the data and they will ensure that the data management plan is carried out.

2. Data organization, storage and archiving

An integrated information system is currently under development in the frame of EPPN2020. It aims at hosting datasets produced in all categories of platforms in EMPHASIS (see WP2).

Its main features are that

- The references and names of all objects involved in TNA experiments (e.g. plants, sensors or images) are standardized and unambiguous. This is essential to trace these objects in further analyses, including those performed by groups not involved in experiments. For instance, it is essential to trace the x-y-z position of sensors in a platform or in a field, and their successive calibrations to correctly estimate environmental conditions during experiments. The same applies to the x-y position of plants in platform experiments and of plots in field experiments. Another important information is the movement of plants between facilities, especially in the case of perennial plants that spend a small part of their lives in a platform.
Identification systems will be based on persistent unique identifiers for the objects mentioned above. The preferred technical solution is the use of URI and DOI, which will progressively be deployed in all sites receiving TNA. Files and folders will be versioned and structured by using a name convention.

- In all sites participating to TNA, data will be stored in servers with duplication in companion servers located in a different buildings, different sites when possible, in such a way that they are not lost in case of server failure. Data will be saved daily with backup on the separate server. Backup will be checked at intervals of two weeks.

For longer-term storage and data sharing EPPN2020 will progressively use the European e-infrastructure European Grid Infrastructure. We aim at preserving datasets for at least ten years, in the frame of the EMPHASIS[1] data management plan (ESFRI project). Associated costs for dataset preparation for archiving will be covered by the project itself for the first 4 years.

[1] https://emphasis.plant-phenotyping.eu/

3. Data ownership and data sharing

The ownership and rules of release of the datasets generated during a transnational access will depend on categories of data. The rationale of these rules is to optimize data use and to facilitate meta-analyses.

- For scientific projects, phenotypic data (e.g. images, measurements, observations) belong to the beneficiary of the access. Data will be published whenever possible, and made available either after publication or at the latest three years after the last day of the experiment. An advisable procedure is that the access provider group is associated with data analysis and publication in order to obtain the best possible analyses. Detailed procedures will be established on this base in bilateral conventions between users and providers.

- For technological projects, data will be published whenever feasible. The main results will be made available to the EPPN2020 partnership and to a broader community after publication. SMEs are not obliged to diffuse their data to a larger circle but this will be encouraged.

- Four categories of data will remain the property of the providers and made available to the users.

(i) Environmental data collected during the experiment will belong to the provider, in such a way that this group can perform meta-analyses of environmental data over seasons and years.

(ii) Calibration results and calibration procedures for sensors and cameras need to be analysed across experiments and years, so they belong to the provider.

(iii) Trait recovery workflows and procedures used to extract phenotypic measurements from raw sensor data

(iv) Experimental designs within the installation and the statistical tools to optimize them are developed in JRA2 and belong to providers and partners of JRA2. All this information will be made available to the users and, once published, to the whole phenotyping community.

When relevant, partners will share datasets in a publicly accessible disciplinary repository using descriptive metadata such as MIAPPE. Additional metadata will be stored and made available within a separate XML or RDF file in a standardised way by using machine readable schema or ontologies. Keywords will be added by using standardized controlled vocabularies.

Public groups will publish software codes along with datasets in a disciplinary repository. Whenever possible, analysis will be performed using freely available open source software tools. It is planned to make our dataset publicly available in a disciplinary research data repository along with scholarly journal and open access publications after the primary publications, in all cases three years after the end of the experiments. All effort will be dedicated to make datasets understandable for other researchers by using standards and metadata.

4. File formats

We aim at producing data files with the following characteristics:

- Non-proprietary
- Open, documented standard
- Common usage by research community
- Standard representation (ASCII, Unicode)
- Unencrypted
- Checksum to ensure integrity

Preferred file format choices will include:

- ODF, PDF (not Word)
- CSV, ASCII (not Excel)
- PNG, TIFF, JPEG2000, GIF (classical JPG to be preferably avoided because of a risk of obsolescence and of loss of information)
- JSON, XML or RDF