European Plant Phenotyping Network

EPPN 2020

D2.4. Methodology and software for integration and standardization for platform data, based on explicit environmental characterization, dynamic biological models and statistical methods for analysis of single and multiple trials.

Francois Tardieu (INRAE), Emilie Millet (WUR), Xavier Draye (UCL), Fred van Eeuwijk (WUR)

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°**	D2.4	**Title**	Methodology and software for integration and standardization for platform data, based on explicit environmental characterization, dynamic biological models and statistical methods for analysis of single and multiple trials
Work Package	**N°**	JRA2	**Title**	Design and analysis of phenotyping experiments across multiple platforms, scales of plant organisation, traits and management conditions

Date of delivery	**Contractual**	30/04/2020 (Month 36)	**Actual**	29/05/2020 (Month 37)
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)	INRAe, UCL, WUR			
Responsible Author	**Name**	François Tardieu	**Email**	francois.tardieu@inrae.fr

Version log			
Issue Date	**Revision N°**	**Author**	**Change**
15/05/2020	1	Francois Tardieu	Shared with co-authors
20/05/2020	2	Cloé Paul-Victor	Reviewed by Project Manager

Executive Summary

In a context of climate change, a single phenomic experiment comparing genotypes cannot address the genetic improvement of crop resilience to abiotic stresses and new pests. Joint analysis of experiments carried out in different installations, at different time resolutions from minutes to months, for contrasting environmental conditions (current or future) is an essential component of the EPPN2020 rationale. This multi-scale strategy allows one to understand and combine traits that provide positive features for environmental scenario and links physiological mechanisms from the molecular to field scales.

This document presents how Joint Research Activities (JRA), Transnational Accesses, networking activities (NA) and a joint trans-platform trans-JRA experiment collectively contribute to integration and standardization of datasets. Although *stricto-sensu* activities of EPPN2020 are limited to controlled conditions, we take into account the integration with field datasets to set the scene for future activities that involve field experiments.

After surveys performed in JRA and NA, two levels were defined for every partner installation. A first level was considered as a necessary condition for the considered platform to host transnational access experiments subsidized by the project. A second level was considered as the objective to reach at the end of EPPN2020. This labelling of installations was a tool for action and interactions between JRAs and installations.

Ten topics were defined for improving integration and standardization, with two levels each when relevant, and actions are presented to reach level 1 and level 2 for each of them. (i) A consistent environmental characterization was reached by defining the environmental variables and the good practices to measure them. Level 2 consists in a mapping of these conditions for each plant in the platform. (ii) A consistent calibration process was reached by regression between image analysis outputs (e.g. pixel number) and traits in each platform. Level 2 consists in sharing virtual images to perform a common calibration at consortium level (and beyond it). (iii) Outlier identification is necessary but a source of divergence between groups. Methods were designed, tested and diffused for the identification of outlier points in growth curves and of outlier plants in panels of genotypes. (iv) Joint principles for experimental designs are essential in a phenotyping community. Training sessions, protocols and distribution of software were performed to reach this objective. (v) Methods for integrated data analysis were designed and tested in the consortium, namely a 2D dynamic characterization of the spatial variability of phenotypes, the comparison of time course curves and mixed models for multi-site experiments. (vi) Reaching FAIR-compatible datasets require that measured variables have reproducible names in each installation (level 1), while level 2 consists in designing, testing and diffusing software elements for generating machine-readable variable names connected with public ontologies. (vii) In the same way, FAIR datasets require the use of an information system allowing one to trace, organize and visualize phenotypic data together with the necessary metadata. Of particular importance is the identification of all sensors, vectors and plants allowing one to organize the metadata in an efficient way. A combination of training sessions, hands-on courses, case studies and local implementation resulted in common practices (level 1) and in local information systems in nine local infrastructures (level 2). (viii) The next step webservices that relate local information systems. This is at the stage of use case. (ix) Integration is also encouraged in TransNational Accesses (TNA): the existence of previous datasets with the studied genetic material and the strategy for integrating those datasets with TNA outputs is an explicit criterion for project evaluation by referees. (x) A trans-platform trans-JRA experiments is carried out in 13 platforms, and used for testing all methods designed in JRAs, for addressing methodological issues in phenomics and for exchanges with other communities (e.g. ELIXIR, AGMIP).

Table of contents

Document information 2
Executive Summary 3
Table of contents 4
1. Introduction 5
1.1. Scientific context 5
1.2. Strategy for integration and standardization 5
1.3. EPPN2020 labelling of platforms into "level 1" and "level 2" : a tool for actions and interactions between JRAs and platforms 6
2. Consistent methods, software and tools for characterizing environment and plant measurements across platforms, including model-assisted phenotyping (JRA1) 9
2.1. A consistent environmental characterization 9
2.2. A consistent calibration process across platforms 9
3. Consistent experimental designs, analysis tools ad methods across multiple platforms (JRA2) 10
3.1. Common tools for quality control and annotation protocols 10
3.2. Common tools for Experimental designs 11
3.3. Data analysis, methods for trans-platform analysis 12
4. Building a consistent information system in the different nodes and defining standardization strategies (JRA3) 13
4.1. Data interoperability 13
4.2. Local data management 14
4.3. Distributed information system 14
5. Encouraging integrative projects for transnational accesses (TNA and NA) 15
6. A trans platform trans jra experiment for testing methods and addressing methodological issues (jra1) 15
7. Conclusion 16

1. INTRODUCTION

1.1. Scientific context

In a context of climate change, a single phenomic experiment comparing genotypes cannot address the challenges associated with the genetic improvement of crop resilience to abiotic stresses and new pests. Indeed, this challenge does not consist of solely associating a genotype to one phenotype in a controlled environment, but rather in getting insight into the plasticity of the plant phenome when exposed to a range of environmental conditions in the field. Two plants sharing the same genotype when placed in contrasting environmental conditions can display markedly different plant structures, functions and metabolisms. Genotypes whose plants perform best in one condition are not those with best performance in another condition (genotype x environment interaction). In a study (Millet *et al.*, 2016) that aimed at identifying genomic regions associated with yield (quantitative trait lici, QTLs), based on a panel of 250 genotypes in 36 fields, all identified QTLs had positive, neutral or negative effects on yield depending on the environmental scenario in the considered experiment (e.g. dry and hot during flowering time vs dry and cool during the same period of time). Hence, the role of phenomic experiments in controlled conditions is not to replace field studies, nor to perform experiments mimicking fields, but to expose genotypes to defined conditions under which measurement with higher precision are possible. Platform measurements provide the necessary elements for understanding and exploiting the genotype x environment interaction in genetic and agronomic studies.

Analysing plant phenotypes also faces the challenge of how to deal with rapidly varying environmental conditions and plant adaptation mechanisms. During a summer day, a plant can be at 11°C with a favourable water status in the early morning, but then experience 36°C and suffer severe water stress six hours later, triggering spectacular changes in plant phenotype such as wilting or rapid variation of leaf growth rate (Caldeira *et al*., 2014). Because one aims at linking plant performance (integrated over months) with mechanisms of plant adaptation to changing conditions (measured over minutes), it is essential that experiments performed at different time scales can be analysed jointly.

1.2. Strategy for integration and standardization

Based on the above two paragraphs, the overall rationale of EPPN2020 is to provide to a broad scientific community the equipment, tools and analysis methods allowing joint analysis of experiments carried out in different installations, at different time resolutions from minutes to months, for contrasting environmental conditions (current or future). This multi-scale strategy allows one to understand and combine traits that provide positive features for each (main) environmental scenario. It links physiological mechanisms with plant performance across genotypes and environments from the molecular to field scales. However, EPPN2020 is limited to controlled conditions, due to the topic released by the European Commission in 2015 for an 'advanced community'. We took care of setting the scene for combining experiments in controlled conditions at different scales with series of field experiments.

This deliverable presents our strategy for integration and standardization of methods, software and rules across platforms, and how it connects with international initiatives (e.g. MIAPPE, ELIXIR, Crop Ontology). We present here the "broad picture" whereas details are presented in specific deliverables that all aim at integration and standardization across platforms. It presents this strategy through (i) a suite of methods, common rules and software developed in the tasks of each of the three Joint Research Activities (JRA), namely tasks 1.1, 1.3, 1.4, 2.1, 2.2, 2.3, 2.5, 3.3, 3.4, (ii) a priority given, in TransNational Accesses, to those applicants who analysed their genetic material in field experiments prior to the TNA experiment in controlled conditions (NA1.1, 1.5, 2.2, 2.3) (iii) on a trans-platform, trans-JRA experiment with a common genetic material grown in 11 indoor and two field platforms of EPPN2020 (JRA1 task 4), plus in two field experiments including one brought to the project by a partner, (iv) networking activities aiming at diffusing the methods designed in JRAs.

1.3. EPPN2020 labelling of platforms into "level 1" and "level 2": a tool for actions and interactions between JRAs and platforms

The mapping and characterization of installations belonging to the EPPN2020 consortium was key to integration and standardization. It was performed for each JRA, in such a way that a precise idea of the situation for each aspect was collected at the end of year 1. Then, two levels were defined for each platform and each criterion.

- A first level was considered as a necessary condition for the considered platform to host transnational access experiments subsidized by the project,
- A second level was considered as the objective to reach at the end of EPPN2020 for the considered criterion and platform, and a necessary condition for this platform to join the future EMPHASIS European infrastructure.

The different actions for reaching standardization and facilitating integration were identified in the DoA. They correspond to most of the tasks of every JRA or NA. Table 1 distributes them in topics to be addressed, with levels 1 vs 2 when relevant. The different actions that were taken are presented for each topic and level. They collectively aimed to help every platform to reach the first level as rapidly as possible, and the second level at the end of the project. The text which follows presents more detail on these actions and refers to relevant deliverables or to sections of the first periodic report.

Table 1. *Summary of the tasks in deliverable D2.4. The table is organised in tasks, themselves divided in tasks for reaching level 1 (necessary for the corresponding platform to host experiments of TNA) and level 2 (desired level at the end of EPPN2020). Actions are described for each task, and the current status of the task is shown in the right column*

Topic for reaching standardization	Action1, initial steps	Action 2, progress	Action 3, final steps	Current status
1. Consistent environmental characterization				
Level 1: Common variables measured and good practices	Meeting for sharing the variables	Good practices distributed to the consortium, with names of recommended	(NA) recommendations for good practices	Completed
Level 2: Modelling environmental conditions as sensed by every plant	Training session for presenting the problem and relevant tools	Software developped, with tutorial, distributed to the consortium	Publications	In progress
2. Consistent calibration process				
Level 1 Regressions between pixel number and traits with different objects in each site	Survey of local practices		Suggested corrections if necessary	Completed
Level 2: Calibration with common objects	Measuring characteristics of each imaging cabin, generating images as should be seen in the cabin	Calibration in each platform with distributed images, test with new images	Joint analysis by JRA and each platform	In progress
3. Outlier identification				
Level 1: Outlier points in a time course	Software designed and distributed to consortium	Case studies distributed to the consortim	Bilateral + consortium discussion	Completed
Level 2: Outlier plants in a panel	Training session for presenting the problem and relevant tools	Software + Case studies distributed to the consortim	Publication + open discussion	In progress
4. Experimental design				
Level 1: Identification of sources of variation in each platform	Survey by JRA (in situ), protocol for identification of environmental gradients	Training session: spatial variability, interpretation for indoor experiments	Protocols for JRA with designs taking gradients into account, bilateral discussions	Completed
Level 2: Automatic generator of experimental design	Software designed and tested in 3 platforms	Modified software distibuted and tested by the consortium	Software used for the design of joint experiments (see	In progress

Topic for reaching standardization	Action1, initial steps	Action 2, progress	Action 3 final steps	Current status
5. Data analysis, methods for trans-platform analysis				
Level 2: 2D characterization of spatial variability of phenotypes	Software designed (SPATS)	Software distributed and tested by consortium members	Software used for the analysis of joint experiments	In progress
Level 2: comparison of curves	Software designed	Software distributed to voluntary platforms		In progress
Level 2: Mixed models for multi site analysis	Software designed	Software distributed to voluntary platforms		In progress
6. Data interoperability				
Level 1: Reproducible names for variables	Survey of local practices: on line + bilateral interactions	Training session: the need for traceability of what was measured	Rules for variable naming and meta data diffused and used in the consortium	Completed
Level 2: Semi automated naming, Correspondence with public ontologies and initiatives	Design of a reproducible method for naming of variables	A software distributed and tested for variable naming	Methods for mapping variables with existing public ontologies	In progress
7. Local data management				
Level 1: Identification of plants, sensors and objects (URIs)	Survey of local practices: on line + bilateral interactions	Training session: the need for traceability of plants, vectors, sensors and events	Interaction with JRA: Plants, vectors, sensors and events traced with site-specific names and identifiers	Completed
Level 2: Local information systems	Design and diffusion of a generator for Uniform Resource Identifyers (URI)	Training sessions for deploying local information sysetms	Local information systems installed in 9 installations	In progress
8. Distributed information system	Design webservice allowing access to different local information systems	Test for three sites	Diffusion to a wide community	In progress
9. Transnational access, integrative projects encouraged	Advertisment for TNA includes encouragement for field + indoor analyses	Rules distributed to applicants and reviewers, for taking integration in the review process	TNAs combine previous datasets with that collected in TNA, some with field + controlled conditions	In progress
10 Trans platform trans JRA experiment	Design of objectives and protocols for testing integration across platforms and JRA	Common protocol agreed, genetic material agreed and distributed, 13 experiments performed (2 in field)	Data analysis with (i) test of methods in JRAs, (ii) addressing methodological questions, (iii) common analysis with other initiatives (MIAPPE-Elixir, ICASA)	In progress

2. CONSISTENT METHODS, SOFTWARE AND TOOLS FOR CHARACTERIZING ENVIRONMENT AND PLANT MEASUREMENTS ACROSS PLATFORMS, INCLUDING MODEL-ASSISTED PHENOTYPING (JRA1)

2.1. A consistent environmental characterization

A common environmental characterization is a necessary condition for joint analysis of datasets originating from different platforms. At the start of EPPN2020, a clear lack of appropriate methods was made obvious for the characterization of environmental conditions in phenomic installations, including spatial and temporal variations, and for calibration procedures. This was evidenced by the outcome of the joint experiments in the first EPPN project (2012-2015), and by a survey of all installations performed by JRA1 at the beginning of the project.

Actions were taken for platforms to rapidly reach level 1 (Table 1). This level involves a common list of environmental conditions to be measured in all platforms with appropriate practices. We began with a training session that presented the bases of environmental characterization and the necessary variables for modelling and for a relevant clustering of environmental scenarios. The consortium then agreed on a list of environmental variables and on good practices to measure these variables. It involves the use of several light and temperature sensors and (at least) one humidity sensor and one CO_2 sensor within each compartment (growth chamber, greenhouse or FACE field), with measurements every hour. Documents were made available to the consortium, including the names of recommended sensors and an annually updated. These recommendations are now posted, in a simplified way, on the project website for a large diffusion.

The second level is defined by the computation of radiation and temperature maps that allow the reconstruction of microclimate at the individual plant level. The same training session was necessary to make partners aware of the huge spatial variability for light and temperature in a greenhouse or a growth chamber (e.g. 92% variability of cumulative light on a given day in a greenhouse). A method and a software were developed for a simple characterization of the amount of light sensed by each plant of the platform (Cabrera-Bosquet *et al.,* 2016), the same was performed for temperature (see highlight JRA1 in the EPPN2020 first periodic report), both based on a combination of sensor networks and of modelling. Corresponding software elements and tutorials were distributed and tested by consortium members. This resulted in changes in methods and software, through individual interactions and group discussions.

2.2. A consistent calibration process across platforms

Cameras in each installation provide images that cannot easily be compared between installations. However, joint analyses across platforms are based on traits, not on images, thereby requiring a calibration process (e.g. transforming every day the number of pixels of segmented objects into leaf area, plant biovolume or root length). For traits to be compared across platforms, this calibration needs to be consistent over the consortium, in

such a way that joint analyses are not biased by differences in calibration. The first EPPN project launched a series of measurements of common plant-like objects, which did not result in sufficient reproducibility. The EPPN2020 strategy for a common calibration is presented in deliverable D1.1. As above, two levels were defined.

Level 1 consists of the independent adoption by each platform of procedures that ensure the correspondence between pixel number and object sizes. A survey showed that all platforms perform regressions between relevant traits and pixel number (or other variables given by standard image analysis packages). Interaction with platforms allowed improvement of these methods, but this level was reached in all platforms at the end of year 1. Note, however, that Level 1 calibrations are based on plants grown locally in each installation, hence not common across the consortium.

Level 2 was difficult to define, because most obvious methods in the literature were found to be either impractical or inaccurate, whereas those methods that were *a priori* acceptable by each platform did not provide sufficient reproducibility. An intense process of discussion and 'trial and error' was necessary for identifying appropriate methods, presented in deliverable D1.1. Briefly, the strategy is that (i) partners perform geometrical characterization of each imaging cabin and of cameras and send results to JRA1, (ii) JRA1 sends back model-based 'ground truth' images of a set of common plants as they would be captured by the considered platform, (iii) partners measure these images with their normal procedure and perform a calibration (iv) Partner and JRA1 jointly analyse the resulting calibration and, potentially, correct them. Level 2 calibration therefore replaces regression over data acquired from locally grown plants with regression over data acquired from a common plant set. This process will be used and fully tested during the "trans platform trans JRA experiment", and then extended to platforms outside the consortium in the frame of companion projects.

3. CONSISTENT EXPERIMENTAL DESIGNS, ANALYSIS TOOLS AD METHODS ACROSS MULTIPLE PLATFORMS (JRA2)

At the start of EPPN2020, there was a clear demand for a unified set of tools and methods to analyse data from platforms. Indeed, large spatial heterogeneities, the presence of outliers and of confounding of effects also exist in controlled conditions, although this is often less recognised than in the case of field conditions. The diversity of phenotyping techniques, the collection of data over time, and the increased amount of data points make it difficult for platform users to directly apply designs, models and analysis methods originally developed for field trials.

3.1. Common tools for quality control and annotation protocols

Phenomic experiments with thousands of plants inevitably face disorders in sensors or cameras, but also errors in plant labelling or incorrect generation of seeds due to undesired pollen diffusion. This generates outlier points or plants. Identification of these outliers is relatively easy in small datasets but needs automated methods if hundreds of plants are characterised every day. A methodological study (Alvarez Prado *et al.,* 2019) showed that

genetic analyses are profoundly affected by the presence or absence of outliers. Unexpectedly, outliers generated false-positive allelic effects, in addition to decreasing the overall power of the analysis. Using traceable, reversible and common methods for outlier identification is an essential condition for joint analyses of datasets collected in different platforms. Our strategy and methods are presented in deliverable D2.1.

Level 1 consists in (semi) automatic identification of outlier points in curves corresponding to each studied plant (Table 1). After a survey of local practices, a software was designed for automatic labelling of outlier points, and case studies were distributed to the consortium. In this way, partners could test their methods together with the outputs of the software. This was followed by bilateral and consortium-level discussions

Level 2 consists in identifying plants that are likely to be too specific for inclusion into genetic analyses, for instance likely to be of the wrong genotype or which suffered serious physiological disorders. This is a risky exercise, which needed discussion at consortium level, because it interacts with the spatial analysis of phenotypes as presented in 3.3, but also because it is intrinsically subjective, even with the help of statistical tools. Software and case studies were distributed to the consortium, followed by discussions. No absolute protocol was distributed to the consortium, except the (absolute) rule that labelled results should not be deleted (reversible process), and the name of the person who annotated (plus criteria) is a necessary metadata in the information system, allowing one to trace the whole process.

3.2. Common tools for Experimental designs

Reliable experimental designs are necessary in experiments in greenhouses and growth chambers, somewhat incorrectly labelled as 'controlled conditions'. As stated above, light and temperature have a higher spatial variability in indoor conditions than in the field (Cabrera-Bosquet *et al.*, 2016), but the phenomic community has a limited experience to account for such variability. The first action of the JRA was to perform in situ visits of most platforms, suggest protocols for identification of environmental gradients, including measurements as presented in §2.1. A training session was held before an annual meeting, aimed at increasing the awareness of the consequences of spatial heterogeneity for data analysis, and at presenting tools that were recommended at levels 1 and 2 (Deliverables D2.2 and D2.5).

Level 1 for project partners consisted in the clear description of the layout of the platform, together with obvious environmental gradients. After interaction with the JRA, partners adapted the experimental designs they currently use to take these gradients into account, in particular for TNA experiments.

Level 2 consisted in designing an automatic generator of experimental design, which takes into account the platform layout and environmental gradients. This software was first tested in three platforms and fine-tuned, before being distributed to the whole consortium. It is now widely used, in particular for TNA, and will be finally tested in the trans-platform trans JRA experiment. It will be published and distributed, whereas a "packaged" version will be made commercially available by an SME partner of the project.

3.3. Data analysis, methods for trans-platform analysis

Data analysis cannot be standardized, and probably should not be because it intrinsically depends on the objectives of every study. It is now a common practice that the same dataset is analysed by several groups with different viewpoints and objectives. However, each specific analysis may or may not involve advanced tools that can, and should, be shared in the consortium and, beyond it, to the whole phenomic community. Hence, there is no compulsory requirement corresponding to level 1, all items presented below correspond to level 2.

3.3.1 An automatic tool for 2D characterization of spatial variability of phenotype.

A tool was designed (SpATS) for interpolating, visualizing the spatial variability of phenotypes, together with tools for correcting phenotypes for spatial variation encountered at the platform. The software was designed, distributed and tested by consortium members. It is used for the multi-platform multi-JRA experiment.

3.3.2 Modelling of time-series at platforms

Phenotypic traits at platforms are typically recorded as time series, or repeated measurements. The parameters that describe the growth and development dynamics of the traits in relation to time and environmental conditions contain important information about the potential of genotypes to adapt to changing conditions. Flexible statistical methods were developed to model plant growth at platforms and to extract dynamical parameters. Teaching material on how to fit flexible functional forms to plant growth data was tested in various courses on the analysis of phenomic data.

3.3.3 Mixed models for multi-site analyses

Plant growth parameters as estimated at multiple phenotyping platforms can be analyses in a joint platform analysis to study how plant growth dynamics change between platforms. This type of analysis provides insight in the environmental dependency of growth parameters and helps to understand and predict adaptation. A further modelling step can be made by integrating field data and platform data, where platform and field data can include both dynamic traits (growth parameters) as well as static traits (yield at the end of the season, biomass at a particular time). The central objective is to model the genetic and environment dependency of a complex trait like yield under field conditions, with genotype by management and genotype by environment interactions, as a function of information obtained in the field and at the platforms. An adequate class of statistical models to performs such an analysis is the class of mixed models. An integrated analysis pipeline for data coming from phenotyping platforms is presented in van Eeuwijk *et al.* (2019).

4. BUILDING A CONSISTENT INFORMATION SYSTEM IN THE DIFFERENT NODES AND DEFINING STANDARDIZATION STRATEGIES (JRA3)

4.1. Data interoperability

Each platform and each project has its own way of naming measured variables. At first sight, this causes technical problems for computer analysis of datasets originating from different platforms. For instance, no package nor computer language can identify 'Leaf area', 'LA' and 'Deployed L area' as being the same variable. Another difficulty is that the person who performed the analysis may have left the group, so it is not possible any more to retrieve how and on which precise organ the measurement was taken. Beyond these technical problems, a theoretical difficulty deals with the representation of knowledge. Frequently, a common name can refer to different objects, characterized with different methods. For example, 'water use efficiency' has a meaning that differs in different scientific communities, from 'the ratio of yield to irrigation + rainfall during the crop cycle', 'the ratio of biomass to transpired water by a whole plant, at the end of a season', or 'the ratio of photosynthesis to stomatal conductance in a given leaf at a given time'. Similarly, plant height has profoundly different meanings in the field (mean height of the canopy, taking into account leaf bending) and in an indoor platform (position of the highest pixel of a plant). Finally, data differ whether they represent ta genotypic mean after correction for spatial variations and outliers, or raw values corresponding to individual plants. Our experience shows that it is not possible to fully reconcile these views that are deeply rooted in each community's practice. It is more efficient to track what every platform actually did, to provide automated ways to name the variables accordingly and, whenever possible, to map the measured variables onto published ontologies, following rules provided by other initiatives such as MIAPPE or Crop Ontology. The strategy is presented in deliverable D3.4.

Level 1. The very first rule set by EPPN2020 is that every platform carefully stores this information as 'metadata', with a 'quadruple definition' that states the measured object, the measured variable, the method and the unit (e.g. 'meristem temperature measured with a thermocouple, in °C', or 'canopy height calculated from two images of a drone, in m'. Reaching this level first required a survey of local practices, with online questionnaire plus bilateral interactions, to understand both the current status in each platform but also the necessary information to be collected for the relevant traits measured by the platform. A training session followed to improve the awareness of partners for traceability. Finally, rules for variables naming and collection of metadata were agreed and distributed in the consortium.

Level 2. A method for reproducibly naming variables was designed and diffused in the consortium, together with a software that generates names with the 'quadruple definition'. Partners are currently testing the software and correcting it for unexpected difficulties. The software also provides tools that facilitate the access to published ontologies. This effort is done in close interaction with the MIAPPE working group that joins efforts of the infrastructures EMPHASIS, ELIXIR and of EPPN2020.

4.2. Local data management

The same problem of traceability is posed for all objects in an experiment. At first sight it seems useless to trace individually a cart that carries a plant, the pot in which the plant is grown and the plant itself. In the same way, sensors could be traced according to their position in the greenhouse. However, a plant with its pot may change position in the greenhouse, some pots may cause problems of toxicity and a given plant can travel across different installations, for instance from a high throughput platform in which it is imaged every day to a specialized platform for detailed physiological experiments. In the same way, if samples are taken for omic measurements, labelling them with a number and keeping track of the genotype makes it virtually impossible to track, for instance, the seed lot the plant originated from or its position in the greenhouse, thereby impeding spatial corrections or analysis of outliers. Finally, if events corresponding to a plant (e.g. physiological disorders or accident) are kept in a lab book, the amount of work for retrieving this information for thousands of plants every day makes this task impractical. It is therefore essential that all the information associated with an experiment is traced, and organised in an information system (deliverables D3.1 and D3.4).

Level 1. As above, the first step was a survey of local practices and constraints, followed by a training session. Rules for naming variables and for collecting metadata were diffused and used in the consortium. Notably, the syntax for names largely differed between partners at this level, the only requirement for TNA was that the information was collected and stored.

Level 2. Information systems with common principles were designed by the JRA and by local infrastructures. Three of them were considered as acceptable, and are currently deployed in different sites. A key feature is an automatic identification of Uniform Resource Identifiers, which allow any of the objects mentioned above to be identified in a unique way, with an automatic link to the metadata collected for the considered plant, sensor, vector or event. A software for automatic generation of URIs was designed and diffused in the consortium. Further, training sessions were organised for the implementation of an information system, which gathered both biologists in charge of platforms and data persons. The JRA then helped local personnel to install a local information system. This process is now engaged or finished in nine installations, and is rapidly increasing.

Noteworthy, this information system differs from those which store genotypic means for every variable and genotype, also essential for genetic analyses. Discussions are under way with MIAPPE and ELIXIR for building tools that link both categories of information systems.

4.3. Distributed information system

A final goal of the project, together with the future infrastructure EMPHASIS, is to link local information systems via webservices that allow a user to collect information from several platforms. For example, the traits corresponding to a given genotype, collected in different platforms, can be accessed together with environmental information and the necessary metadata. This is still at the state of proof of concept and is tested in three installations.

5. ENCOURAGING INTEGRATIVE PROJECTS FOR TRANSNATIONAL ACCESSES (TNA AND NA)

Transnational accesses follow a review process described in deliverable NA1.4. It is specified that the review process takes into account the existence of previous datasets and the strategy to use the EPPN2020 dataset to perform a multi-scale, multi-platform analysis. As a result, most of the selected projects combine datasets, and an appreciable proportion of them combine field and platform data.

Workshops and training sessions organised by the networking work package also diffuse success stories of combined analyses.

6. A TRANS PLATFORM TRANS JRA EXPERIMENT FOR TESTING METHODS AND ADDRESSING METHODOLOGICAL ISSUES (JRA1)

Integration and standardization are more efficient when involved researchers work on common objects and datasets. This was the rationale for a joint experiment, carried out by 13 platforms and covering domains of the three JRAs. The objectives were three-fold, (i) testing all the software elements and methods recommended in the three JRAs on a common dataset, (ii) using this dataset as a way to interact with other communities (iii) collecting a dataset which allows addressing methodological issues for phenotyping.

A challenge, and an opportunity for this exercise, was that involved platforms are of different types, for instance measuring root systems in 3D in soil (X-ray tomography), in 2-D in soil or artificial medium or in aeroponics. Other platforms measure integrated responses of shoots with different methods and throughputs, in controlled conditions or field. A common panel of genotypes (15 inbred lines, one common tester and the corresponding 15 hybrids) is used in all platforms, with consistent protocols designed by JRAs. Experiments are either finished or ongoing.

The objective of testing JRA methods in such a diversity of platforms is a 'stress test', in terms of environmental characterization, spatial analysis, design and data organization. As such, it was considered as an interesting case study for different operations as use cases for the distributed information system (§4.3), or for variable naming and ontologies, in a work common with MIAPPE (§3.1).

The dataset will also allow addressing methodological issues, on the common topic of how information collected in a given platform helps interpreting datasets in another platform or in the field. This objective is foreseen to include several issues, for example (i) how 3D information on root systems, considered as 'ground truth', translate into 2D information in rhizotrons or 3D information in aeroponics, (ii) how information about roots and shoot can be collectively used for an analysis of field datasets. Addressing these questions will be a test for methods presented in §2.3.

7. CONCLUSION

This document presents the strategy of EPPN2020 for integration and standardization of data throughout the whole range of activities of the project. Indeed, this was an overall objective of the project from its beginning. Whereas integration of multi-scale multi-platform datasets is a main objective per se, standardization is a way to facilitate it. It proved to be relatively straightforward in some cases, such as environmental characterization or camera calibration, and could be dealt with via a top down approach following a long process of discussion and co-construction. In other cases, such as outlier detection or variable naming, it was considered as essential to trace the process followed by partners rather than imposing top-down practices that proved inefficient for these topics. Finally, we believe that TNA experiments and the trans-platform trans JRA experiment will generate dataset that will be used for several years in the improvement of methods for data integration. All these methods are either published or will be so, and software elements will be made available to the whole community, both as R codes freely accessible but sometimes difficult for beginners, and as packages made commercially available by the SMEs associated to the project.

Alvarez Prado S, Sanchez I, Cabrera-Bosquet L, Grau A, Welcker C, Tardieu F, Hilgert N. 2019. To clean or not to clean phenotypic datasets for outlier plants in genetic analyses? *Journal of Experimental Botany* 70, 3693-3698.

Cabrera-Bosquet L, Fournier C, Brichet N, Welcker C, Suard B, Tardieu F. 2016. High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist* 212, 269-281.

Caldeira C, Bosio M, Parent B, Jeanguenin L, Chaumont F, Tardieu F. 2014. Rapid changes in leaf elongation rate are compatible with a hydraulic control under soil water deficit or high evaporative demand. *Plant Physiology* 164, 1718–1730.

Millet EJ, Welcker C, Kruijer W, Negro S, Coupel-Ledru A, Nicolas SD, Laborde J, Bauland C, Praud S, Ranc N, others. 2016. Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios. *Plant Physiology* 172, 749-764.

Van Eeuwijk FA, Bustos-Korts D, Millet EJ, Boer MP, Kruijer W, Thompson A, Malosetti M et al. (2019) Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Science* 282, 23-39.