D2.1 Criteria list for infrastructure

Merlijn Morisse, Darren Wells, Pierre-Etienne Alary, Roland Pieruschka, Stijn Dhondt

Document information

EU Project N°	739514	**Acronym**	EMPHASIS-PREP
Full Title	Preparation for EMPHASIS: European Infrastructure for multi-scale Plant Phenomics and Simulation for food security in a changing climate		
Project website	https://emphasis.plant-phenotyping.eu/		

Deliverable	**N°**	D2.1	**Title**	Criteria list for infrastructure
Work Package	**N°**	WP2	**Title**	Infrastructure setup, operation and funding

Date of delivery	**Contractual**	30/04/2018 (Month 16)	**Actual**	30/08/2018 (Month 20)
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)	Merlijn Morisse, Darren Wells, Pierre-Etienne Alary, Roland Pieruschka, Stijn Dhondt			
Responsible Author	**Name**	Merlijn Morisse	**Email**	**merlijn.morisse@ugent.vib.be**

Version log			
Issue Date	**Revision N°**	**Author**	**Change**
06/02/2018	1	Merlijn Morisse	First draft version – in collaboration with WP4 - the ontology document
25/06/2018	2	Merlijn Morisse	Shared version published
27/06/2018	3	Roland Pieruschka	Edit and comments
28/06/2018	4	Darren Wells	Edit on the part of Controlled conditions
29/06/2018	5	Merlijn Morisse	Edit on the part of Lean Field
20/07/2018	6	Pierre-Etienne Alary	Edit on the part of intensive field.
20/07/2018	7	Merlijn Morisse	Edit all
22/07/2018	8	Stijn Dhondt and Roland Pieruschka	Edit all
06/08/2018	9	Darren Wells and Merlijn Morisse	Edit and finish

Executive Summary

Objectives

The objective of EMPHASIS-prep is to develop a long term, distributed, pan-European infrastructure for state-of-the-art plant phenotyping experimental platforms which aims to improve crop performance needed to cope with climate changes and to keep pace with population growth. To tackle these global challenges, novel approaches to identify improved plant phenotypes and explain the genetic basis of agriculturally important traits are required. The new and existing plant phenotyping platforms uses non-destructive, image-analysis based determination of the phenotype of plants and allow for a characterization of plant traits.

The criteria list provides a detailed overview of plant phenotyping infrastructure within the scope of EMPHASIS, all which can be divided into five major pillars of plant phenotyping infrastructure:

1. Plant phenotyping in (semi-)controlled conditions.

2. Intensive field experiments in highly equipped field sites or semi-controlled field sites.

3. Field sites with minimal equipment, which could be combined in a network of fields with different environmental conditions.

4. Modelling platforms to support plant phenotyping data analysis.

5. Information systems for plant phenotyping data management supporting open science.

The criteria list was developed in discussion with the plant phenotyping community and by mapping the existing and upcoming infrastructures in pan-Europe. Four regional workshops have been organized to that end in different regions of Europe. Plant scientist of these regions were asked to present their plant phenotyping infrastructures and activities. Starting with this information these pillars of EMPHASIS could be confirmed. Furthermore, work package 2 (WP2), together with WP3 and WP4, has developed a survey to the community to extract more details of these infrastructures and to be able to translate this into a community driven criteria list for EMPHASIS.

The current criteria list presents a result up-to-date status of the ongoing participation and discussion with the plant phenotyping community in the development of EMPHASIS. It is thus still open for modifications in the course of ongoing stakeholder engagement.

Main Results

By mapping the pan-Europe state-of-the-art infrastructures, the criteria list of EMPHASIS could be defined and subdivided in five pillars. Below these five pillars and their criteria are described as main results of the criteria list:

Pillar 1: Plant phenotyping in (semi-)controlled conditions can be defined as automated glasshouse and growth chamber infrastructures for plant phenotyping to characterize the architecture, growth and physiology of individual plants and monitor environmental conditions. Throughputs are typically 100s-1000s plants, thereby allowing the analysis of genetic variability and calculation of allelic effects in a large range of climatic scenarios.

Pillar 2: Intensive field experiments are field sites fully equipped for detailed environmental characterization at plot level, and vectors (typically ground or airborne) able to perform measurements of phenotypic variables such as leaf area, plant number, plant architecture or physiological status via, for example, hyperspectral or fluorimetric cameras. Plants are subjected to natural conditions of soil and climate, but some intensive field installations could have for example, modified rainfall or atmospheric conditions, via special equipment such as Free Air CO2 Enrichment (FACE), mobile rain-out shelters or heat shields. With throughputs of typically 100s – 1000s microplots, this allows assessing plant performance in climatic scenarios that mimics possible climate conditions of the future and novel management schemes.

Pillar 3: Field trials with precise environmental monitoring equipment, and equipment for precise measurement of yield components and grain moisture in thousands of micro plots. This is potentially combined with field-based imaging or remote sensing (airborne) systems. Individual fields are usually combined into a network of field trials over multiple geographical regions with different environmental conditions, thereby allowing estimation of yield of hundreds of genotypes across diverse climate scenarios.

Pillar 4: Modelling platforms are virtual platforms of different nature that have the intention to support plant phenotyping. It has been proven very difficult to form criteria for modelling phenotyping platforms, therefore, a number of different models that could be used to support plant phenotyping are described in this criteria list.

Pillar 5: plant phenotyping information systems, which aim to provide methods and interfaces to manage, share, reuse and visualize heterogeneous plant phenotyping data stemming from different sources, often functioning in an interdisciplinary context.

Table of contents

Criteria for plant phenotyping infrastructure

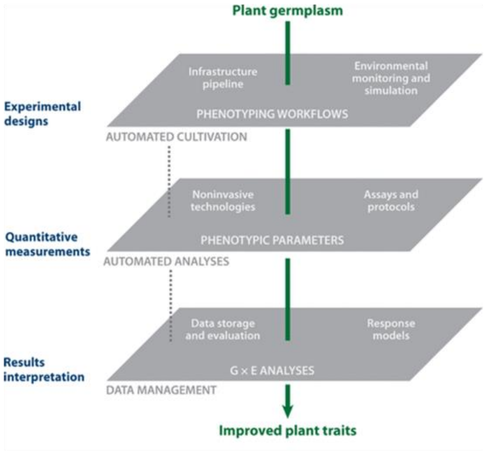
Infrastructure pillars

Plant phenotyping requires an integrated concept to fully explore its potential and achieve high-throughput assessment of plants under a variety of environmental scenarios. Within EMPHASIS we follow a concept that addresses different levels of phenotyping pipelines and allows the establishment of routine measurements as well as the development and implementation of approaches for new research questions. The five infrastructure pillars in EMPHASIS complement that concept to address relevant questions in basic and applied plant science across different scales.

The general plant phenotyping concept can be characterized by four levels:

- **Experimental design** with plant cultivation under controlled conditions in pots or small containers often working in an automated mode or under field conditions with plants or crops in dedicated plots under natural conditions. The plant cultivation requires a dedicated environmental control and monitoring of, that allows the analysis of the plant environment interaction. This is usually integrated into the infrastructure pipeline that enables dedicated treatments.
- **Quantitative measurements** refer to non- or minimally invasive sensors in different measurement modes to measure relevant traits and to analyze plant structural and functional properties during development, based on defined assays and protocols. Often the trait assessment is attached to automated analysis of phenotyping parameters and molecular analysis can be linked through organ or tissue sampling. Destructive measurements are also performed, mostly at the end of the plant lifecycle during harvest e.g. direct estimation of biomass or yield components.
- **Data analysis** includes the calculation of traits/elaborated variables from images and sensor outputs, the analysis of the plants or genotypes in relation to environmental conditions and the genetic analysis including heritability's and, most often, identification of quantitative trait loci of raw or elaborated variables.
- **Storage and provision** of the datasets involve their organization, annotation and short/long term storage in accordance with the Open Science principles enabling re-use data as well as validation and development of models.

Within EMPHASIS we define the common characteristics and criteria of different infrastructure pillars based on the phenotyping concept to additionally allow a synergistic operation of these infrastructures and provide services within and between the pillars. In this context, we aim at the development of phenotyping pipelines to assess plant properties across scales with different levels of precision from lab to field (Pillar 1-3), combined with central access to coordinated information systems for data management and storage (Pillar 5) and with platforms for plant/crop modelling (Pillar 4).



1. Phenotyping platforms in (semi)-controlled conditions
2. Intensive field experimental sites
3. (network of) field sites with minimal equipment

4. Modelling platforms to support plant phenotyping data analysis
5. Information systems for plant phenotyping data management

Modified after Fiorani & Schurr 2013

The five infrastructure pillars within EMPHASIS

1. Phenotyping installations in (semi-)controlled conditions for high-resolution and high-throughput phenomics.

Objectives

Phenotyping installations under controlled conditions allow the investigation of the variability of measured plant traits as a response to well-defined and monitored environmental conditions (i.e. abiotic and biotic conditions) with a capacity of several hundreds to thousands of plants. Facilities may also be linked to high precision platforms for *deep phenotyping* with lower throughput (tens to hundreds of plants) and measurements over shorter timescales (weeks) and time steps (minutes to hours).

Experimental design

Plant growth

Plants are typically grown as single plants in pots containing soil or artificial media; small stands of several plants in larger containers; in rhizotrons; on plates filled with artificial media; hydroponically in defined media; aeroponically; or on germination paper.

Environmental control

Controlled environments are characterized by the ability to monitor, and in some cases manipulate, environmental conditions allowing the simulation of varied environmental scenarios. *Semi-controlled environments* include glasshouses and poly-tunnels that allow manipulation of one or more of the following: water and nutrient supply, temperature, humidity, light (supplementary to daylight), and biotic stresses. *Well-controlled environments* include growth cabinets, chambers, and rooms and typically allow control of two or more of the following: water/nutrient supply, light (intensity, spectral range, photoperiod), humidity, temperature, atmospheric gas composition, and biotic stresses.

In both semi- and well-controlled installations, environmental conditions are continuously monitored throughout each phenotyping experiment using calibrated sensors deployed in several places of the installations, with measurements at suitable temporal resolutions. Creating environmental models of the installations will allow spatial and temporal interpolation of measured variables. Information from the sensor array is used to design the experiment in terms of initial plant positioning (and movement during the experiment if applicable) to minimize, or take advantage of, the effects of environmental variability. Samples are phenotyped by sensors in fixed positions or using automated systems that either move the samples to the sensor (e.g., conveyors, carousels) or move the sensors (e.g., gantries, actuators).

Quantitative measurements

Facilities are usually equipped with non-invasive or minimally-invasive sensors for structural and functional characterization of shoot and root architecture, biomass, photosynthesis, and nutrient relations and can functionally analyze large numbers of plants subjected to different environmental conditions with varying water, nutrient or atmospheric conditions (see Experimental design).

Examples of non-invasive sensors:

- Cameras for imaging shoots in multiple modalities: visible light (RGB), near infrared, multispectral, hyperspectral, thermal. Most often the imaging is done from different direction for 3D reconstructions, and may involve moving cameras for detailed images.
- Fluorescence imaging platforms for characterizing photosynthetic parameters.
- Imaging of root structure and function either via rhizotrons, aero-/ hydroponics, or in soil columns for imaging in tomographic systems (X-ray computed tomography, magnetic resonance imaging, positron emission tomography).
- Time-series imaging (4D imaging) allowing measurement of changes with time of phenological stages of plant leaf area, angles and growth rate of individual leaves, root growth rates and tropisms.

Key Protocols

Plants are distributed in installations based on statistically relevant designs (e.g. alpha lattice or blocks), with position in installations defined based on the maps of environmental variables established based on distributed sensors and models. The plants should be labeled according to a standardized system to make them trackable and to avoid mismatches between plants. For example each plant could be identified by an unambiguous Unique Reference Identifier, and its position is recorded via x-y coordinates, or possibly changes in coordinates with time when protocols require changing positions.

Environmental sensors are calibrated to suitable standards with sensor and calibration data linked to experimental datasets. Phenotyping sensors also require regular calibration with calibration data linked to experimental datasets. Current minimal requirements for calibrations and designs are in line with the Level 1 requirements within the Horizon2020 research and innovation program EPPN^{2020} (project Grand Agreement: 731013; https://cordis.europa.eu/project/rcn/210944_en.html), a program that provides transnational access to (semi)controlled condition phenotyping platforms in Europe. The future goal is that all installations linked to EMPHASIS would meet the EPPN2020 Level 2 requirements.

Data management

Installations are connected to an interoperable information system (see Pillar 5: *Data management and information systems*). The future goal of EMPHASIS is that all data linked to and produced from experiments in phenotyping installations in (semi-) controlled conditions for high-resolution and high-throughput phenomics should be properly managed by an information system that integrates with the general EMPHASIS layer for centralized data access and enables interoperability with datasets from other local infrastructures.

Facility management and access

Installations are typically managed as part of a *local infrastructure* (see *EMPHASIS Ontology of objects involved in phenotyping*) which shares governance, and follows common principles for costing and user access.

Examples

WIWAM XY: a plant-to-sensor system for canopy imaging of small plants in a controlled environment room with automated high-precision irrigation[1].

GrowScreen-Rhizo-1: a sensor-to-plant system for rhizotrons in a glasshouse allowing phenotyping of root and shoot growth and architecture[2].

PhenoArch: a plant-to-sensor system for 3D imaging of plants, together with pipelines for high throughput calculations of architecture, light interception or stomatal conductance[3].

[1] https://www.wiwam.be/phenotyping-system/wiwam-xy
[2] http://www.fz-juelich.de/ibg/ibg-2/EN/Research/ResearchGroups/JPPC/technologies/GROWSCREEN_rhizo/growscreen_rhizo_node.html
[3] https://www6.montpellier.inra.fr/lepse/M3P/PHENOARCH

2. Intensive field experimental sites for high throughput phenomics

Objectives

Intensive fields are installations set up to allow the detailed study in the field (natural light condition) of hundreds of plant micro-plots through frequent measures of several plant traits. Intensive fields are highly equipped in order to monitor both the plants phenotype and the environment during the plant growth cycle. The quasi-continuous data acquisition paired with the storage of time courses enables the study of the plant growth dynamics, which can be analyzed with respect to the dynamics of environmental variables. Intensive fields can therefore host experiments aimed at deciphering complex plant traits through the study of genotypic variation and genotype-by-environment interaction.

The combination of high frequency measurements and the high number of monitored variables (both phenotypic and environmental) makes intensive fields a framework for high throughput plant phenotyping under natural conditions. Additionally, some intensive field installations, referred to as semi-controlled intensive field sites, are equipped to alter specific key environmental conditions in order to simulate future climatic scenarios, for example through rainout shelters which can be used to generate drought or a FACE (Free-Air CO2 Enrichment) system to simulate elevated CO_2 condition. There is usually a trade-off between the level of control and the throughput, intensive field installations typically involve some hundreds of micro-plots whereas high throughput field installations typically involve thousands of micro-plots.

Experimental design

Plant growth

Plants are grown in micro-plots usually in natural soil. The number of micro-plots and replicates in each experiment should be high enough in order to perform relevant statistical analyses and to allow genetic analyses of target traits as defined by the users (e.g. screening of mapping populations). Additionally to being highly equipped for environmental characterization and plant phenotyping, intensive field installations should prove to display thorough descriptions of both the experimental design and the material (plant material, sensors, conveyors, algorithms) used in each experiment. Keeping track of when and how sensors used during an experiment have been calibrated is important for data provenance matters.

Environmental control

A qualitative and quantitative description of the environment is associated to the intensive field sites. Soil characterization includes spatialized data on its physical properties, chemical composition and water content (typically maps with a resolution of one meter). Irrigation and nutrient input linked to each experiment is quantitatively recorded (e.g. water volume for irrigation, measured with rain gauges). *Semi-controlled intensive field installations* allow manipulation of some of these conditions. The latter includes for example (i) Free Air CO2 Enrichment (FACE) that increase CO2 concentrations to levels forecasted in the future, (ii) rainout shelters that control rainfall and soil water content, (iii) defined nutrient application, and (iv) manipulation of soil or air temperature applications or management schemes. These platforms are well equipped with distributed sensors for light, rainfall, temperature, evaporative demand and air humidity, allowing measurements of gradients of the conditions that are manipulated (typically gradients of temperature around heaters or distribution of CO2 concentration in FACE systems. Soil and air dynamic characteristics are mapped via sensors (e.g. water and nutrient contents, biotic factors).

The environmental sensors gathering data on soil (e.g. lysimeters) and air characteristics (e.g. temperature, radiation, rainfall and air humidity) have their (X, Y, Z) position known. If there are different depths of soil or of canopies, the data points are accompanied with metadata stating the exact z-position of sensors. Time definition is typically one hour or less for most sensors, if variables measured by soil sensors do not fluctuate excessively, measures can be made less frequently (e.g. twice a day). Conversely, specific installations such as FACE require more precise temporal definition, e.g. every 15 minutes.

The events associated to every experiment can be planned (sowing date and other crop management interventions) as well as unanticipated (abiotic stress, e.g. caused by frost or a storm, or biotic stress). In the same way, experimental layouts may change compared with the initial plan because of practical reasons. It is essential that metadata associated with the experiments are stored in the database for planned and real dates, events or layouts in order to allow subsequent data analysis.

Key Protocols:

Experimental micro-plots are organized with statistically relevant designs (e.g. alpha lattice or blocks). Their positions in the field are defined based on maps of environmental variables. Each micro-plot is referred to via GPS coordinates, with metadata such as area, distribution and position in relation to paths.

Environmental sensors are calibrated to suitable standards with sensor and calibration data linked to experimental datasets. Phenotyping sensors also require regular calibration with calibration data linked to experimental datasets.

Quantitative measurements

As for environmental data, intensive field installations display frequent measures of the plant growth during its life cycle, through non-destructive techniques such proximal or airborne sensing. This high-throughput field phenotyping enables the study of the dynamics of plant growth through time courses made from several data points.

Intensive field installations record dynamic imaging of canopy architecture and function via proximal sensing systems, often carried in gantries, phenomobiles, accompanied with images collected with UAVs for variables with rapid temporal changes such as canopy temperature. This involves, for example, (i) stereo imaging, LiDAR (Light Detection and Ranging) and structured light measurements giving access to the geometry of the canopy and light interception, (ii) cameras with different wavelengths (multi, hyperspectral, or thermal imaging), allowing evaluation of leaf area, light interception and chemical composition in different elements (e.g. pigments, nutrients, quality parameters), (iii) photosynthesis proxies via active and passive chlorophyll fluorescence measurement.

Data management

Installations are connected to an interoperable information system (see Pillar 5: *Data management and information systems*), The goal of EMPHASIS is that all data linked to and produced from experiments in phenotyping installations in (semi-) controlled conditions for high-resolution and high-throughput phenomics should be properly managed by an information system that integrates with the general EMPHASIS layer for centralized data access and enables interoperability with datasets from other local infrastructures.

Facility management and access

Installations are typically managed as part of a “local infrastructure” as defined in EMPHASIS Ontology of objects involved in phenotyping. They are generally running by an infrastructure manager, scientists and technical staff.

Access to the platform can be achieved by contacting its manager or through a national network or a transnational access as the one set up by EPPN^{2020}.

Examples

1. Semi-controlled intensive field installations equipped to alter major environmental conditions:

Breed-FACE (Germany)[4]

Pheno3C (France)[5]

2. Intensive field installation:

Field Scanalyzer Rothamsted (UK)[6]

[4] http://www.fz-juelich.de/ibg/ibg-2/EN/Research/ResearchGroups/shoot_dynamics/technologies/Breed-Face/Breed-Face_node.html
[5] https://www6.inra.fr/pheno3c_eng
[6] https://www.rothamsted.ac.uk/field-scanalyzer

3. Field experiments using minimal equipment, linked as a network of fields

Objectives

Phenotyping in field experiments using minimal plant phenotyping equipment are performed in agriculturally-relevant and breeding-like conditions, which allows investigation of the (genetic) variability of measured plant traits in a range of environmental scenarios as a response to differences in field management, soil composition, or environmental conditions. Monitoring of the environmental conditions (i.e. abiotic and biotic conditions) are essentially the same as those in intensive fields, with sensor networks for measurement (e.g. temperature, radiation, rainfall, air humidity and soil water potential), at a temporal definition of typically one hour and should allow to understand differences in plant or crop performance and could explain statistical outliers which are not due to genetic variation of the species. The amount of plants, plots or fields per year is dependent on the experiments and should be statistically relevant, with the relevant number of repetitions according to the data analysis protocol. Throughputs are typically hundreds to thousands of (micro)plots.

A key objective for EMPHASIS will be to increase the capability of networks of field trials, number and geographical/ climatic coverage, and the coordination of existing infrastructures and their integration to facilitate analyses of plant performance across climatic gradients, the development of appropriate phenotyping methods and statistics. The scientific communities involved here are; biologists, geneticists, statisticians, agronomists, breeders, modelers and specialists of information systems and technology development.

Experimental design

Plant growth

Plant and crops are mostly grown following standard agricultural practices of the considered region and crop. Field experiments could be in distributed sites following environmental gradients (e.g. north-south, oceanic–continental) which will allow prediction of plant performance in current and future climatic scenarios, and establishment of the link between their performance and underlying traits at the stand level when imaging is performed, typically with UAVs or remote sensing. Plants are grown as micro canopies in micro-plots that can vary in size according to the experiment. In rare cases minimal equipped field experiments could be done by analysis of individual plants e.g. phenotyping of trees or annual plants without branching, sunflower or maize, for looking into specific traits that are better measured at plant level.

The number of micro-plots and of their repetitions in each experiment should be high enough in order to perform relevant statistical analyses and to allow genetic analyses of target traits as defined by the users (e.g. screening of mapping populations). To cope with light variation on the sides of the experiments boarder-plots should be planted and not be included in the experiment analysis.

Key Protocols

Experimental micro-plots are organized with statistically relevant designs (e.g. alpha lattice or blocks) similar to those of intensive fields. Their positions in the field are stored via experimental lay out that allow performing statistical spatial analysis (e.g. row-column analyses). Environmental sensors are calibrated to suitable standards with sensor and calibration data linked to experimental datasets.

Environmental sensors are calibrated to suitable standards with sensor and calibration data linked to experimental datasets. Phenotyping sensors also require regular calibration with calibration data linked to experimental datasets.

Environmental control

The plants are grown in natural conditions, following the standards of crop/plant during agriculture practices. Field facilities for phenotyping requires recorded meteorological data and soil characteristics with a definition of one hour in close proximity to the field. Some meteorological variables can be collected in weather stations located at some distance (e.g. light or dew-point temperature), whereas other variables are necessarily collected at close distance (100s m) from the experiment (e.g. rainfall or air temperature).

Quantitative measurements

Europe has a wealth of field experimental stations run by public research institutes and private (seed) companies.

Field phenotyping data is mostly produced by quantitative measurements of variables such as; seedling emergence, flowering time, plant height, biomass and yield components. Simple imaging techniques involving UAVs and, probably, satellite imaging in a near future, become increasingly available, thereby allowing more detailed analyses in particular:

Networks of environmental sensors beyond classical meteorological data of simple use.

Measurement of basic plant structural and functional properties via imaging with cameras carried by airborne systems like drones, measuring in visible range, multi and hyperspectral range (e.g. surface coverage, biomass and canopy height, chemical composition) thermal cameras to estimate water relations via e.g. canopy temperature.

The use of such sensors will improve the quality of data analyses and make field phonemics data more re-usable and interoperable.

Data management

Resulting datasets are stored in an interoperable information system (see Pillar 5: *data management and information systems*), that includes (i) data on yield components and other phenotypic variables, (ii) environmental data following the same phenotyping ontologies as in other categories of installations, (iii) all the necessary meta data, in particular statistical design, position of micro-plots, events and methods used for data analysis.

Facility management and access

Experiments are run by a group, installations refer here to the use of environmental equipment, of the database and ontologies for traits, objects, and events and are typically managed as part of a “local infrastructure” as defined in *EMPHASIS Ontology of objects involved in phenotyping*. They are generally running by an infrastructure manager, scientists and technical staff.

Examples

1. field installations with minimal equipment:

Flanders Institute for agriculture, fishery and food (ILVO): has multiple hectares of agriculture field trials with mostly food and biomass crops. The institute uses remote sensing technology for quantitative measurements, sometimes combined with field phenotyping models[7].

Rothamsted research: has multiple hectares of agriculture field trials with mostly food and biomass crops. Is specialized in multi-year field trials[8].

2. Projects with networks of field trials

DROPS - Drought-tolerant yielding plants

[7] https://www.ilvo.vlaanderen.be/

[8] https://www.rothamsted.ac.uk/

From 2010-07-01 to 2015-12-31, closed project. Funded under: FP7-KBBE. Developed novel methods and strategies for analyzing the genetic control of yield components of a panel of 250 maize hybrids across 30 experiments in Europe, grouped into environmental scenarios[9].

European Consortium for Open Field Experimentation (ECOFE)
Network of existing field stations, in different climatic regions, across Europe. Mainly focusing on the standardized multi-location field trials on Maize and wheat[10].

[9] https://cordis.europa.eu/project/rcn/95052_en.html
[10] https://www.ecofe.eu

4. Modelling

Objectives

Models constitute important tools integrated in, or interfacing with, phenotyping pipelines. Unlike experimental platforms, models can be duplicated, have no accessibility limit, and may be run independently by users. All these advantages alter substantially their modalities of access and use.

EMPHASIS, as an infrastructure, provides the data/model quality policy that facilitates the articulation between its experimental platforms and models. Furthermore, the EMPHASIS information system (see Pillar 5. Data management and information systems), which organizes and stores datasets, embeds models within phenotyping pipelines by keeping track of standardized simulation inputs, outputs, and metadata in addition to experimental datasets, so the combined modelling and experimental workflow is traceable and reproducible. Finally, trainings constitute a cornerstone of the modelling pillar, allowing users to run models independently, while the EMPHASIS administration will dispatch extra modelling support requests towards the appropriate modelers.

Categories of models supporting phenotyping activities:

- PROCESS-BASED CROP MODELS use mathematical equations to simulate growth, development and yield as a function of weather, soil conditions and crop management. Such models integrate scientific knowledge from diverse agronomic disciplines, ranging from plant breeding to soil physics but do not include multidimensional descriptions of plant architecture;
- FUNCTIONAL-STRUCTURAL PLANT MODELS (FSPM) combine the representation of three-dimensional plant architecture (structures, including topological and geometrical aspects) with selected processes (functions, including physical and physiological aspects that affect growth and developments). A variant of the FSPM is the STRUCTURAL PLANT MODEL that does not include functional processes on top of the structural layer;
- STATISTICAL MODELS, such as multivariate regressions, are trained to predict specific variables like historical yields based on simplified measurements of weather and soil without being constrained by state-of-the-art descriptions of physical or physiological processes. In the same category, artificial neural networks may be trained to estimate complex features such as plant organ type based on large datasets of images with a method called deep learning.

Important tasks in phenotyping pipelines relying on models:

- MODEL-ENABLED TRAIT QUANTIFICATION:
  Various models are used to compute sophisticated information from raw sensor data captured in phenotyping installations. An example of this is the computation of individual plant light interception using radiation transfer models combined with virtual plant mockups (structural plant models) parameterized with images of the plant . Similar problems may also be addressed with statistical models. For instance, variables such as total root surface were recently quantified by parameterizing an image analysis tool by deep learning after being trained on large amounts of root system images simulated with a root structural model. Here, models are used or developed with a specific purpose of extracting hidden data.
- MODEL-ENABLED TRAIT DISSECTION:
  The exploitation of dense and high-resolution observations of phenotypes and environments (from field and/or controlled conditions phenotyping installations) with models helps dissect complex traits into their physiological/genetic components (e.g. patterns of responses to water deficit across environmental gradients in mapping populations). This process is very close to plant (eco-)physiology and very often leads to (re-)formulation of component trait crop models and FSPM. Such models are cornerstones in the analysis and dissection of field traits using phenotypic data obtained in EMPHASIS platforms.
- MODEL-ENABLED PREDICTION:
  Statistical and process-based crop models can also be used to predict the fitness of crops to diverse soils and climates, possibly accounting for genotypic features. This approach allows in silico testing of thousands of interesting allele combinations in the diversity of current and future European climates. Crop models able to reconstruct complex traits from component traits and to predict the behavior of novel genotypes in various management by climate scenarios are the cornerstone of this type of activity.
- MODEL-ASSISTED EXPERIMENT DESIGN: Complex traits such as drought tolerance highly depend on the type of drought (e.g. frequency, duration, etc.) and soil. While testing combinations of genotypes and environments in the field is particularly expensive and time consuming, modelling studies offer the possibility to isolate few high-potential genotypes among candidates prior to field characterization. Modelling may also be used to identify experimental conditions that best discriminate between genotypes, or maximize the information content of the experiment when characterizing plant traits.

5. Data management and information systems

Objectives

EMPHASIS aims to create centralized access to phenotyping data by building and integrating compatible, consistent information systems that will provide methods and interfaces for interoperability of datasets to manage, share, reuse and visualize heterogeneous, high-throughput plant phenotyping data stemming from different sources often in an interdisciplinary context. The main objective is that datasets are findable, accessible, interoperable and reusable (FAIR standard) in such a way that the datasets can be analyzed by several groups inside and outside EMPHASIS. This requires software interfaces providing different levels of access to different users to allow a data analysis in relation to environmental conditions. Data organization and storage needs to be done in a secure way over long term periods, the data can be interpreted in a biological context and used for meta-analyses of experiments. The basic ontology of objects defined in EMPHASIS is described in the "Ontology of objects involved in Phenotyping" [11]

Data description

Collected data can be numerical data, images, documents, texts, and manual measurements. The data volume can be of several Pbytes at infrastructure level.

Data categories and sources that need to be managed are:

- Resource description and management: species, genotype, seed origin, accession.
- Facilities: installations, sensors, cameras, vectors (e.g. conveyors) and specific devices.
- Trait recovery workflows: the sensor and image analysis methods and software tools used to extract traits from raw image and other data.
- 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).
- Date and description of management events (e.g. irrigation, pruning, sampling) and of observations (e.g. plant disease, unintended events).
- Characteristics of experiments (e.g. design, protocol and organization).

[11] https://emphasis.plant-phenotyping.eu/Infrastructure_ontology

Data organization, storage and archiving and sharing

An integrated information system aims at hosting datasets produced in one or several categories of platforms in EMPHASIS. Its main features are:

- The references and names of all objects involved in experiments are standardized and unambiguous, which is essential to trace objects and events in further analysis.
- Data need to be stored in such a way that they are not lost in case of server failure.
- Sharing datasets in a publicly accessible repository requires the use of descriptive metadata such as described by MIAPPE[12].
- Using ontologies for an unambiguous description of traits.
- In the future we aim at integrating data for central access web based access to multiple connected databases in EMPHASIS facilities.

File formats

Proposed formats

Data files may have 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 (standardized XML-Schemes) or RDF

Management and access

Data management may require a dedicated management and operation team to ensure data quality and support the development of structures that enable reuse. Access to data is open

[12] http://www.miappe.org/

for any data acquired from publicly funded research projects based on data policies that will be developed to ensure clear rules of data use.

Next steps

This criteria list is established by engaging with the plant phenotyping community in pan-Europe to list criteria based on the input of users and owners of plant phenotyping platforms. EMPHASIS will continuously discuss and update the criteria with the community at large. Therefore, EMPHASIS-prep will establish a process for revision of the criteria list, fitting in the governance of EMPHASIS and with fixed time lines and responsibilities. The process of revision will be established in close collaboration with WP5 Legal framework / governance and will depend on the formation of the EMPHASIS governance.

Annex 1: Check list

Deliverable Check list (to be checked by the "Deliverable leader")

	Check list		Comments
BEFORE	I have checked the due date and have planned completion in due time		*Please inform Management Team of any foreseen delays*
	The title corresponds to the title in the DOW		*If not please inform the Management Team with justification*
	The dissemination level corresponds to that indicated in the DOW		
	The contributors (authors) correspond to those indicated in the DOW		
	The Table of Contents has been validated with the Activity Leader		*Please validate the Table of Content with your Activity Leader before drafting the deliverable*
	I am using the EMPHASIS deliverable template (title page, styles etc)		*Available in "Useful Documents" on the collaborative workspace*
The draft is ready			
AFTER	I have written a good summary at the beginning of the Deliverable		*A 1-2 pages maximum summary is mandatory (not formal but really informative on the content of the Deliverable)*
	The deliverable has been reviewed by all contributors (authors)		*Make sure all contributors have reviewed and approved the final version of the deliverable. You should leave sufficient time for this validation.*
	I have done a spell check and had the English verified		
	I have sent the final version to the WP Leader and to the Project coordinator (cc to the project manager) for approval		*Send the final draft to your WPLeader and the coordinator with cc to the project manager on the 1st day of the due month and leave 2 weeks for feedback. Inform the reviewer of the changes (if any) you have made to address their comments. Once validated by the 2 reviewers and the coordinator, send the final version to the Project Manager who will then submit it to the EC.*