Ontology of objects involved in Phenotyping

January 2018

EMPHASIS, (F. Tardieu, X. Draye, R. Pieruschka, C Pommier, B Usadel and P Neveu main authors, with participation of PE Alary, M. Bennett, M. Janni, M Morrisse, T Pridore, and D Wells).

The ESFRI-listed project EMPHASIS aims at a synergistic development and long-term operation of plant phenotyping infrastructure in Europe (https://emphasis.plant-phenotyping.eu/) by developing tools and methods for multi-scale phenotyping, allowing one to analyze genotype performance and trait diversity in current and future European environmental conditions. EMPHASIS aims at:

- Developing an integrated pan-European network of instrumented phenotyping platforms able to test genotypes in a diversity of agro-climatic scenarios, in controlled and field conditions.
- Linking data acquisition to a European-level data management system, with local information systems linked by web-services, and to state-of-the art crop models to simulate plants and crops in current and future climates.
- Developing, evaluating and disseminating novel technologies, thereby providing new opportunities to European SMEs involved in phenotyping and precision agriculture.

We propose here an ontology of objects involved in phenotyping activities, the relationship between them and the rationale for proposed choices. This ontology is used in the EMPHASIS information system but also allows better communication between EMPHASIS members and beyond in internal or external documents. It follows discussions within the consortium and will be periodically re-assessed. Hereafter, all italicized terms correspond to controlled vocabulary, i.e. have a precise meaning that is defined at their first use (capitalized). It is common with the I3 project EPPN^{2020}[1], and contributes to the MIAPPE[2] initiative.

1 MAIN DEFINITIONS

Distributed research infrastructures have different levels of organisation. Here we follow the general nomenclature provided by the EC[3,4] , adapted to plant phenotyping. Furthermore, the organization of information systems requires precise definitions of words frequently used as synonyms in common language, but that in fact correspond to different concepts and practical uses. We attempt here to define these terms and their relationships (Fig. 1)

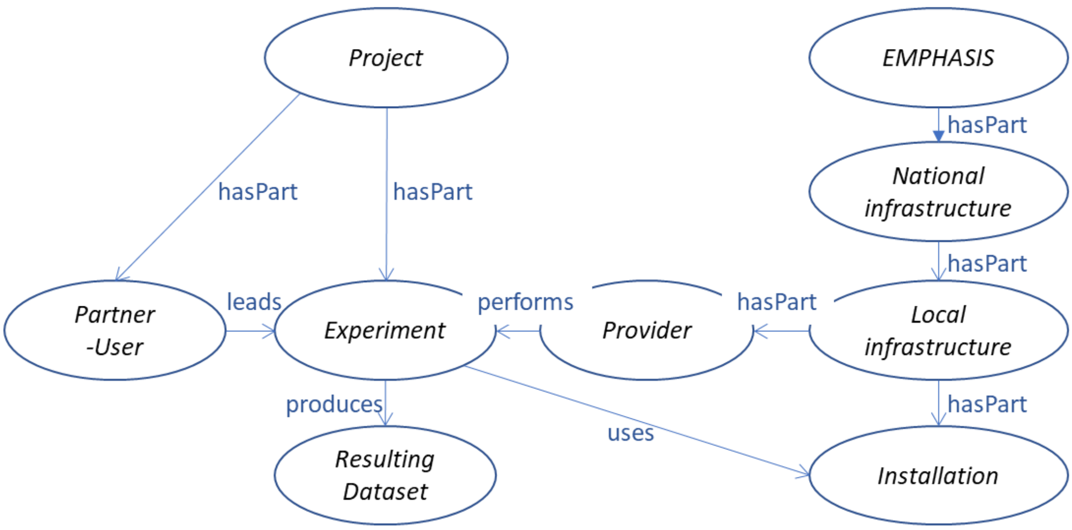

Fig. 1. Main terms used in this document and their functional relationships. Each instance has its own attibutes, not represented here for better legibility. *Provider*: group of people involved in a *local infrastructure*, usually involving subgroups at *installation* level. *Project:* any project, (e.g. UE or national) that involves *partners* that perform experiments, themselves part of the *project*. The terms *Partner* (in a project) and *User* (in a phenotyping infrastructure) map to each other. All these terms are implemented in the information system.

1.1. National infrastructure. A *NATIONAL INFRASTRUCTURE* is an organized group of *local infrastructures* (see §1.2) with governance recognised by national authorities (e.g. respective ministry or national structures). Examples: DPPN (http://www.dppn.de/dppn/EN/Home/), or Phenome-FPPN https://www.phenome-fppn.fr/phenome_eng/. The component *local infrastructures* (see §1.2) share rules or guidelines for *user access* (see §1.3.1) and, potentially, cost calculation, common scientific and industry advisory boards and often common funding. They usually share a common e-infrastructure connecting their information systems with common tools (e.g. ontologies) and web services. *National infrastructures* are currently in development in several countries, with different modes of organisation and governance.

1.2. Local infrastructure. A *LOCAL INFRASTRUCTURE* is a group of *installations* (see §1.3) located in one site depending on one institution (or more), which share governance committees, a common (or at least highly interoperable) information system, common principles for cost calculation and pricing and a common tool for *user access*. Examples: M3P at INRA Montpellier (https://www.phenome-fppn.fr/phenome_eng/), IBG-2: Plant Sciences at Foschungszentrum Jülich (http://www.fz-juelich.de/ibg/ibg-2/EN/Home/home_node.html). A *local infrastructure* may have its own committees such as a Scientific Advisory Board or an Industrial Advisory Board. The latter can also be at *national infrastructure* level.

1.3. Experiments and projects.

A *PROJECT* (e.g. EU or national *projects*) involves PARTNERS that are in charge of *experiments (*see below) performed in one or more *local infrastructures*. A *project* often involves specific ontologies or methods, which can be defined automatically in information systems for all experiments part of the *project*. Some *projects* provide access to *installations*, e.g. EPPN^{2020}.

An EXPERIMENT is a planned activity carried out by a given set of persons on a given set of observation unit, like *plants* or *microplots* (see §2.1), involving a protocol, methods and a resulting dataset. *Experiments* are carried out by partners of a *project*, who are *USERs*, i.e. a group of persons who uses one or more *installations* (see §1.5).

1.4. Activities. *Activities* involve (i) *DEEP PHENOTYPING*[5]*,* usually performed at organ level with a resolution of a few millimetres. It most often involves short *timescales* (weeks) and *time steps* (minutes to day) and medium *capacity* (tens to hundreds of *plants*) (see §2.1 for definitions). It can involve measurements such as local fluxes, organ anatomy or high definition expansion rate. (ii) *HIGH THROUGHPUT PHENOTYPING*[5] in field or controlled conditions involves measurements of integrative traits of *plants* or *micro-plots* such as plant height, biomass, transpiration, 3-D architecture or leaf area, possibly with functional imaging such as multi-spectral or fluorescence imaging. It involves a *capacity* of thousands of *plants* or *micro-plots*, allowing genetic analyses such as genome wide association studies or genomic prediction. (iii) *NETWORKS OF FIELD EXPERIMENTS*[5] involve series of fields in a region following environmental gradients (e.g. north-south, oceanic - continental), in which one performs *LEAN PHENOTYPING* (defined here as the collect of environmental conditions, yield components, simple traits, observations and/or images of outlier plants). Each of these activities can generate novel traits based on novel imaging techniques, on combinations of measurements or on reverse modelling (i.e. inferring an elementary trait from combination of integrated traits).

1.5 **Installation.** An *INSTALLATION* is the elementary level for data acquisition in a specific type of *experiments*. It stands for other frequently used terms such as 'platform', 'facility' or others. Examples: Phenoarch, Growscreenrhizo or others belonging to the EMPHASIS or EPPN^{2020} lists (https://www.plant-phenotyping.org/db_infrastructure#/).

A given *experiment* (see § 1.3.1) can involve one or several *installations* in controlled conditions (plants can move between installations) or in the field (see § 2.3). For the organization of information systems and calculation of costs, it is convenient to consider that each *experiment* and each location such as field or greenhouse can involve several *installations* which are frequently handled by different groups of people and involve different pieces of equipment. This allows identification of common operations (with common items in information systems and common procedures for evaluation of costs), for example in fields with high throughput phenotyping and in *networks of field experiments* with *lean* *phenotyping.* With this definition, an *installation* usually involves instances that produce and handle the data as a result of experiments and projects and activities:

- A set of non- or minimally invasive instruments operated in an automated or semi-automated mode such as (i) *SENSOR*s, i.e. device providing numerical data, e.g. environment sensors or cameras, (ii) *VECTOR*s i.e. devices that either transport plants to a given site for phenotyping, or carry *sensors* in a greenhouse (e.g. conveyor belt, gripping arm, robot) or in the field (e.g. UAV, gantry, field robot). By extension, a *vecto*r can also be a group of persons producing manual observations.
- A management and operating team, ACCESS *PROVIDER*, i.e. an organized group of people who perform *experiments*, usually together with *users*, handle equipment and uses a dedicated set of pipelines and procedures for the analysis of the data originating from the *installation.*
- A set of methods and designs available for *experiments*.
- A set of methods for data handling and analysis, including image and data analysis pipelines.
- A connection with an information system, usually but not necessarily situated at a higher level (typically the *local infrastructure*).

Installations are also considered as elementary units for which cost calculation can be defined (e.g. investment, personnel cost, maintenance, data storage, data analysis etc.). User access can be managed at the *installation*, *local infrastructure* or *national infrastructure* levels depending on the mode of organisation at the national level. It is managed on a *UNIT ACCESS* base (e.g. per day x duration of experiment, per flight or ground passage etc.).

2 FEATURES OF *INSTALLATION*S AND OF RESULTING DATASETS

2.1 Main attributes of installations in field and controlled conditions

SCALES include *spatial* and *temporal* domains that are often interlinked. The TEMPORAL SCALE covers processes that range from fraction of second e.g. electron transport, day e.g. developmental processes and season e.g. growth dynamics and yield. The *SPATIAL SCALES* of plant organization analysed in EMPHASIS are *ORGAN* (any part of the plant that presents defined boundaries and is defined by a specific ontology), *PLANT* (the continuous organism originating from a single seed, callus, rhizome or any other propagation mean) and *PLOT* (a community of plants located on a defined area of a field). The cellular scale is essentially carried out by other infrastructures. Field experiments usually refer to a *MICRO PLOT*, i.e. a portion of the experimental field with a common genotype and set of cultivation techniques. These *scales* correspond to the *OBSERVATION UNITS* defined in MIAPPE, i.e. the level at which traits are measured, calculated and stored in the information system.

The methods and techniques to capture *ENVIRONMENTAL CONDITIONS* are key attributes of any *installation*. The latter are defined as a matrix of values of *ENVIRONMENTAL VARIABLES* (outputs of a *sensor* with appropriate units, e.g. temperature, °C) collected at given *TIME STEPS* (the interval between two measurements, e.g. 10 minutes) over a *TIMESCALE* (the time during which an experiment is carried out, e.g. 2 months). The elementary attributes of an *environmental variable* are (i) the category of *sensor* used to measure it, including make, age, and calibration, (ii) the position of *sensors* (e.g. x-y-z position in a greenhouse, GPS coordinates in a field in which z stands for the altitude of atmospheric *sensors* and depth for soil *sensors*). *ENVIRONMENTAL METADATA* are the collection of attributes, *time steps* and *timescales* used in a particular experiment. Those metadata will be proposed for integration in a future version of MIAPPE.

Environmental conditions comprise calibrated data that depict the time-course of every variable, attached to each sensor (e.g. light or temperature every hour for a given sensor). They serve to calculate elaborate variables, for example the maximum /minimum /mean /median of all sensors outputs measuring a given variable, the number of hours with a temperature higher than a threshold or the light intensity per unit thermal time. Typically, time courses of variables are stored in information systems whereas elaborate variables are reported in the supplementary information of scientific papers. Indeed, the number of elaborate variable can be nearly infinite, depending on researcher's interest. They cannot be easily standardized, and only a limited number of them can be stored in information systems. If an installation or a project uses elaborate variables routinely, their definition, unit and mode of calculations (possibly values) are traced in the corresponding information systems.

Other attributes of a given *installation* are its *CAPACITY* (the number of *plants* or *micro-plots* that can be handled simultaneously in one experiment) and *THROUGHPUT* (the number of *plants* or *micro-plots* handled per unit time). The *annual throughput* indicates the maximum amount of plants handled in a given *installation* in one year. *Throughput*s over shorter periods (e.g. day), together with the *capacity* of the *installation*, are the main determinants of the frequency of phenotypic measurements (reciprocal of *time step* for phenotypic measurements).

2.2 Resulting datasets in controlled conditions

Controlled conditions installations involve both *HIGH PRECISION PLATFORMS*[5] *for deep phenotyping* and *controlled conditions platforms for high throughput phenotyping*[5]. They are hosted in climate chambers or greenhouses. Common characteristics of *controlled conditions installations* involve an *annual throughput* between hundreds and thousands of plants. This allows genetic analyses of measured traits based on one or several *experiments* and corresponds to a capacity of hundreds of plants in *high-precision platforms*, thousands in *high throughput controlled conditions platforms.*

Typical datasets of *experiments* in controlled conditions involve:

- Detailed measurement of *environmental conditions* with distributed *sensors* in different locations, and x-y maps of each measured variable in the greenhouse or growth chamber, for estimating the conditions sensed by each plant. These maps can originate from physical or statistical models. *Controlled conditions installations* can manipulate one or more environmental conditions such as varying light, temperature, water, nutrient or CO_2 availability.
- Images and *sensor* outputs. Imaging may involve active and passive methods with different wavebands for quantitative assessment of *organs* or *whole plants* (in some cases *canopies*) gathering structural and functional data on shoot and root architecture, biomass and growth rates, photosynthesis and nutrient relations. Meta-data include the type and calibration of cameras and sensors, the timing of imaging and environmental conditions during imaging (if different from those in the greenhouse or growth chamber). For the analysis of pathological (pests or pathogens) and beneficial interactions with other biota, novel *sensors* estimate temporal and spatial variations of number of spores and bacteria in air and soil and the progress of diseases can be tracked via imaging approaches.
- Specific experimental designs and protocols, in particular the spatial arrangement of *plants* (including its change with time if any), *time steps* of the environmental or phenotypic measurements, plant handling (e.g. planned and effective dates of irrigation or treatments)
- Precise recording of phenology, i.e. dates of main steps in the plant cycle (in time or thermal time) and the collection of *EVENT*s (e.g. accidents, visual observations on outlier plants etc.).
- The protocols used for loading data into an information system.

2.3 Installations and resulting datasets in field experiments

Field experiments can either be stand-alone (e.g. *intensive fields* with *high throughput phenotyping*) or be part of a *network of field experiments* with *lean phenotyping*[5]. It is convenient to consider that a field *experiment* involves several *installations*, which are frequently handled by different groups of people and involve different pieces of equipment.

2.3.1 Basic field installations.

The *basic field installation* is a field site carrying out *lean phenotyping.* It can be part of a *network of field experiments.* It can also be one of the many *activities* in an *intensive field*. In the latter case, identifying this *activity* as an *installation* helps to bridge datasets that were collected in different categories of fields with either lean or deeper phenotyping. A *basic field installation* is equipped with environmental *sensors,* material for cultivation techniques and for precision harvest at micro-plot level. It is often handled by a group of people with different skills from those handling other installations involved in field experiments (e.g. skills on crop management vs on sensors). Datasets associated with basic field installations/*lean phenotyping* involve the same data and meta-data in *intensive fields* and in *networks of field experiments*, namely:

- Knowledge of the spatial distribution of the physical properties of the considered field (e.g. field-level maps of soil type, soil depth of profile, water reserve etc.).
- *Environmental conditions* recorded at each time step of the experiment by each *sensor* in different locations characterizing the above- and below-ground environment during experiments, together with environmental metadata depicting the categories and makes of *sensors*, their spatial positions (GPS), their age and calibrations.
- Manual environmental measurements at beginning and end of experiments (e.g. vertical distribution of soil water or nutrient status.
- Experimental design and protocol, in particular a map of the spatial arrangement of *microplots.*
- *Cultivation techniques* e.g. dates of irrigation, pest control, weed control.
- Precise recording of phenology, i.e. dates of main steps in the plant cycle expressed in time or thermal time (sowing, emergence, flowering, maturity) and, potentially, of manual measurements such as SPAD.
- A record of *event*s (e.g. hail, pest attack, wind episode causing lodging etc.) and of obvious gradients or outliers.

2.3.2 Other specific field installations define the attributes of intense fields.
In the view presented here, a given field experiment involves different *installations* with many possible combinations of them (e.g. basic field installation + UAV in one field, and basic field installation + FACE + Phenomobile in another field). Identifying installations in this way greatly facilitates the organization of information systems and data analysis.

Installations allowing environmental manipulations contain equipment that allow comparison of plant behavior in either normal or manipulated conditions in the same field. This can involve rainout shelters which protect *plots* during rain episodes, FACE that enrich the air with CO_2 or other gases, local heaters for increasing canopy temperature or equipment for (fert-)irrigation with high precision. Many of these installations can be moved periodically from one field to another, so they need to be considered as *installations* per se, regardless of the field where they are installed. They usually involve specific personnel able to handle this equipment. Datasets associated with these installations are specific environmental data (e.g. high definition of maps of temperature or CO_2 in the installation every day or hour). Metadata involve the characteristics of equipment and *events* (e.g. dates and volume of irrigation or timing of rainout shelter opening/closing).

Plant imaging installations consists of a set of equipment, including imaging *sensors*, *vectors* (e.g. gantry, phenomobile or UAV), a management and operation team, and standardized methods and procedures for data handling, storage and analysis. They are characterized by the throughput of the considered *vector* (e.g. thousands of *microplots* day^{-1} for automated phenomobiles and UAVs) and the set of devices used for active and passive methods with different wavebands (e.g. cameras with RGB or thermal infrared wavebands, LiDars, fluorescence cameras). Except in the case of fields equipped with gantries, these *installations* are most often involved in several field experiments during the same crop season, so they need to be considered as *installations* per se. For both information system and calculation of costs, it is convenient to consider activities such as UAV and phenomobile imaging as different installations, even when they operate in the same field. Indeed, each of them correspond to specific equipment, operators and have different *unit access* bases (e.g. flight for UAV, day for a phenomobile). Datasets associated with these installations are images of one *microplot* at a given time characterized by GPS coordinates of the *microplot* and timing of images. Metadata are the characteristics of the device, the make, age and calibration of each sensor or camera and *events* such as defects or breakdown of a sensor.

Most often, elaborate protocols for imaging or environmental manipulations require a record of phenology more precise than in §2.3.1 (e.g. number of appeared leaves or phenology scale designed for a given species), which can be performed either on a reference genotype or on all genotypes.

Other installations can be defined as allowing any other operation involving specific equipment, group of people and methods. For example, the ability to apply a parasite or fungus can be considered as an *installation* involving equipment and methods for the application, sensors measuring the concentration of spores or bacteria in the air above the field every hour and the skills for the operation.

[1] https://eppn2020.plant-phenotyping.eu/
[2] https://github.com/MIAPPE
[3] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=access
[4] https://portal.meril.eu/meril/[4] https://ec.europa.eu/research/.../2016_charterforaccessto-ris.pdf
[5] F. Tardieu, L. Cabrera-Bosquet, T. Pridmore, M. Bennett, Plant Phenomics, From Sensors to Knowledge, Current Biology, 27 (2017) R770-R783.