EMPHASIS

D4.3: Integration of EMPHASIS into general e-infrastructures

Francois Tardieu, Pascal Neveu, Cyril Pommier Roland Pieruschka, Clément Saint Cast, Xavier Draye, Llorenç Cabrera Bosquet, Björn Usadel. | 20 December 2019

EUROPEAN INFRASTRUCTURE FOR PLANT PHENOTYPING

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	emphasis.plant-phenotyping.eu		

Deliverable	**N°**	D4.3	**Title**	Link with other information systems
Work Package	**N°**	4	**Title**	Information system and imaging workflows

Date of delivery	**Contractual**	31/12/2019	**Actual**	31/12/2019 (Month 36)
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)				
Responsible author	**Name**	Francois Tardieu	**Email**	francois.tardieu@inrae.fr

Version log

Issue Date	Revision N°	Author	Change
13/11/2019	1	Francois Tardieu	First version
20/12/2019	2	Francois Tardieu	Final version

This project has received funding from the European Union's Horizon 2020 Coordination and support action programme under grant agreement No 739514. 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.

Contents

Executive Summary

Objectives

Phenomic datasets are at the crossroad of several disciplines, in particular plant physiology, genetics, ecology and crop modelling. All these communities have developed their own solutions to organise and store resulting datasets in dedicated repositories or information systems, so it is crucial that the effort of EMPHASIS to organize phenomic datasets results in information systems that are interoperable with those of other infrastructures such as ELIXIR (genomics and genetics), ANaEE (Ecology) and AgMIP (crop modelling). This is in a context in which data sciences are receiving an increasing interest, in particular with the EOSC initiative of the European Commission.

Main results

The information systems developed by different communities for phenotypic data differ in their organization and content. Phenomic information systems need to track all elements for re-analysis of individual experiments, in particular the time courses and spatial variability of environmental variables, the time course of traits and the link between samples, organs, genotypes, plants,events, x-y positions of each plant or plot. This requires the use of semantic web to automatically generate complex metadata from a few indications. This level of detail is not necessary in Genetic-Genomic or in Ecology, so corresponding information systems relate integrative plant traits with integrative indicators of environmental conditions, and/or with genomic information. The crop modelling community developed its own standards for traits, environmental variables and managing practices. EMPHASIS has launched common working groups with each of these communities, resulting in a clarification of the role of each information system, in common tasks for mapping ontologies and in a list of tasks that will facilitate interoperability between information systems.

1. Introduction

1.1. Different conceptions of phenotypic information systems in the phenomic, genomic-genetic, ecology and crop modelling communities

The term 'information system for phenotypic data' involves different objects and aims in different scientific communities. The objective of this deliverable was to clarify the objectives of our own community in relation to those in other communities/infrastructures, and to facilitate the organisation of phenomic information systems to make them compatible with those used by other communities/infrastructures.

- The Phenomic community needs specific information systems. Indeed, phenomic experiments in field or indoor conditions have a value that tremendously increases if combined with other experiments, at different scales or in different climatic conditions (1). Gathering phenomic data is expensive and time consuming, while phenomic experiments cannot be reproduced because the same combination of environmental conditions that occurred in one experiment will never occur in another experiment (2). Furthermore, phenomic data may involve measurements with a temporal definition of minutes, such as stomatal conductance or growth rate which can vary by one order of magnitude over minutes, to integrative measurements over months, such as yield or cumulated transpiration. These features require complex information systems allowing a researcher who did not participate to an experiment to reanalyse the same dataset, for instance with other hypotheses or to include it in a larger dataset. In particular, reanalyses require that time courses of environmental conditions are stored with high spatial and temporal definitions, that the exact spatial position of every plot or plant in the experiment is traced, and that the relation between organs, samples, plants, locations, events, experimenters and projects can be automatically extracted from the information system. Such an information system was developed (2) with an open source software and is deployed in several installations in Europe.
- The genetic-genomic community uses other information systems involving mean values for each trait and studied genotype. These values are used, for example, to perform genome-wide association studies relating a given trait to genomic information at thousands to millions of positions on the genome, or to compare gene expression to measured traits (3). Because traits are environment-sensitive, this analysis requires the taking into account of environmental information, usually via mean values of environmental variables during key phases of the crop cycle. The resulting information systems do not need to directly store the complexity of information for each experiment as that in the former paragraph. Indeed, here traits and environmental conditions are scalars, vs time courses above, and the spatial variability of traits and environment are dealt with, previously to inclusion in the database, via the statistical models that extract genotypic values. The MIAPPE working group developed standards for such datasets (4, 5), and specific information systems such as GnpIS are currently available in the frame of the ELIXIR European infrastructure (6).

- The Ecology community essentially uses the same type of information as in the above paragraph, for instance to derive large meta analyses relating traits to environmental indices such as cumulated light or average temperature (7). The MIAPPE working group also takes into account the requirements of the Ecology community (4, 5).
- Finally, the crop modelling community uses phenomic data to calibrate models. The latter relate iteratively environmental conditions every day or hour to transpiration, growth and development on the same day, resulting in cumulative traits such as yield or total soil water/nitrogen uptake. These models need to be calibrated for every genotype, thereby requiring high throughput data collected in the field or in phenotyping platforms (8). This requires precise time course of environmental conditions at several experimental fields, together with (i) traits such as leaf area index, biomass or transpiration in the same fields, (ii) response curves of, for instance, transpiration or growth rate to environmental conditions, often obtained in indoor phenotyping platforms. The AGMIP community organised terms using the ICASA Master Variables List (9) and datasets following the AgMIP data standards and data translator tools (10).

A main question addressed by this deliverable is the extent to which the information systems presented above are compatible. In particular, we addressed the question of the feasibility of a common information system for all communities, or of a pragmatic approach consisting in developing web-services that facilitate communication between distinct information systems.

1.2 The EOSC initiative and the FAIR Guiding Principles

The FAIR Guiding Principles (Findability, Accessibility, Interoperability, and Reusability) now make a large consensus in the scientific community for organizing datasets and analysing scientific results (11). All information systems presented above follow these principles. They need 'Interoperability' in such a way that phenomic data can be used by all communities. Conversely, the other terms 'Findable', 'Accessible' and 'Reusable' have contrasting consequences in each category of information system presented above, because the involved communities do not require the same information for reanalyzes.

The EOSC initiative was launched by the European commission to foster the FAIR principles in all disciplines involved in the H2020 and Horizon Europe frameworks. EMPHASIS is involved in the EOSC-Life project that brings together the 13 Biological and Medical ESFRI research infrastructures (BMS RIs) to create an open collaborative space for digital biology in Europe. Through EOSC scientists will gain direct access to FAIR data connected to workflows and tools in a cloud environment available throughout the European Research Area. EMPHASIS objective to make plant phenotyping data FAIR is therefore developed and implemented in a wider and collaborative context within EOSC as an essential part to make plant science data reusable. Specifically, EMPHASIS is involved in a demonstrator within the EOSC-Life project called A+ (see § 3.4).

2. Methods

EMPHASIS took the initiative to develop common working groups with each of the above mentioned communities. We used a pragmatic method consisting in specific meetings and working groups with each individual community, whereas the general approach is aimed to be synthesised in a review paper in an academic journal. Meetings were kept to a minimum, whereas we dedicated much of our effort to writing common white papers, documents and scientific papers, and to participating to existing working groups such as MIAPPE or AgMIP.

We also worked in common projects. This was the case for the above-mentioned EOSC call, but also of other projects such as EU DROPS/ANR AMAIZING for genotype-to-phenotype analyses (12), and UE DROPS and SOLACE for the relation between phenotype and modelling.

3. Relation with other infrastructures

3.1 ELIXIR (Genetic-genomic community)

The first meeting between EMPHASIS and ELIXIR was held on 22-23 May 2017 at IPK Gatersleben, in a strategic meeting of the MIAPPE working group for re-defining its objectives and working method (https://www.miappe.org/). This was the first occasion for Genomic-genetic, Ecology and Phenomic communities to clarify their objectives and relations. A second meeting was held in Montpellier on 18 May 2018 and a third meeting was a videoconference on 26 October 2018. A white paper followed each meeting, each of them was formally approved by the executive committees of EMPHASIS PREP, MIAPPE and ELIXIR. Several ad-hoc meetings occurred since then for facilitating the writing of white papers, in particular during the 15th Integrative Bioinformatic Symposium (September 18-20, 2019, Paris). Three papers in academic journals were published in the MIAPPE context, coauthored by members of EMPHASIS prep, Elixir and of the Ecology community.

The main domain of the EMPHASIS community concerning data (Fig. 1) is to:

- Produce datasets that jointly include phenotypic and environmental information, most often as time courses of variables in a way that allows traceability of objects, images or events during an experiment, thereby facilitating future meta analyses (‘reusable’ in “FAIR”).
- Analyse data and produce new objects, e.g. 3-D representations of plants or canopies, 2-D maps of environmental conditions, response curves of a given phenotypic variable to one or several environmental conditions.
- Run models that allow dissection/simulation of time courses or spatial variations. Models, either statistical or process-based are a tool for testing hypotheses, but also to cross scales, e.g. between controlled and field conditions. They are also a way for checking data quality.
- Facilitate the access to full datasets, phenotypic environmental and metadata (Findable, Accessible and Interoperable in FAIR).

EMPHASIS, as an infrastructure, does not run these analyses but provides the information system and data quality policy that facilitate access to datasets and their (meta)analyses.

The main domain of the ELIXIR community concerning data (Fig. 1) is to:

- Enable FAIR publication of datasets to allow their use for genetic and genomic analyses.
- Enable data, tools and repositories interoperability by seeking collaboration with relevant communities to build and recommend standards, metadata and repositories.
- Allow findability and accessibility of any scientific data type, including multidimensional phenotype.
- Help the building of integrative datasets that links phenotype to genotype or other data types. This integration is possible with elaborated data, like genetic variation inferred from resequencing experiments or phenotyping two-dimensional data matrices inferred from time series or direct measurement.
- Enable publication and integration of elaborated phenotyping datasets, i.e. data matrices handling values for traits.
- Provide the infrastructure for the quality check of datasets, mainly at the syntactic level and to ensure the presence of minimal metadata like biological material description and complete measurement methodology traceability and provenance.

It is noteworthy that the domains represented in Fig. 1 deal with the specificities of the communities and infrastructure in EMPHASIS and ELIXIR, whereas individual scientists in each community can cover the whole range of activities. Fig. 1 defines the 'domains of excellence' of each community in order to better identify common tasks.

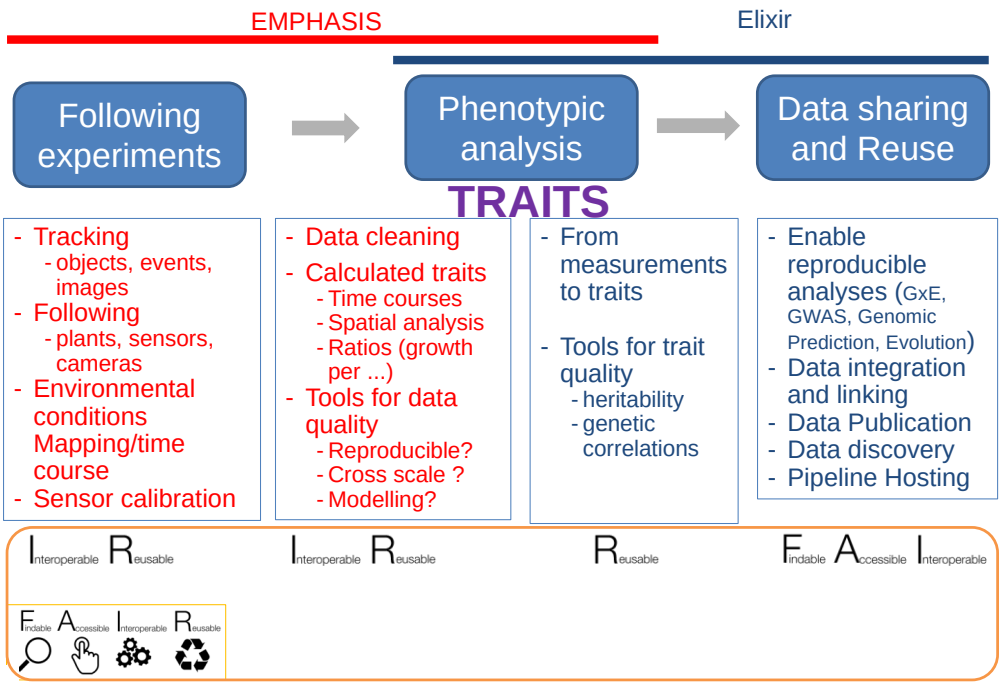

Fig. 1. Schematic representation of the flow of data, from experiments to genetic analyses, with the respective domains of excellence in EMPHASIS and ELIXIR

The common tasks that were identified are

- *Improving environmental characterization*. The list of environmental variables in MIAPPE has been the object of debates, and too specific variables have been removed from the MIAPPE list. A group in EMPHASIS/EPPN^{2020} proposed an approach to better identify the requirements of each community, with two levels (essential/desirable as requirements for installations to provide accesses funded by the project. A document was accepted by the EMPHASIS PREP and EPPN^{2020} executive committees and provided to the MIAPPE group for discussion.

- *Defining "abstract datasets" in EMPHASIS databases*. The difference in the nature and structure of datasets identified above highlights the necessity for EMPHASIS databases to define lists of single point scalars extracted from time series and spatial data from full datasets. Most scientists in EMPHASIS do this exercise but there is currently no room in databases to keep and trace it. A common task is to identify the nature of such lists, their requirements and formats in such a way that they can be queried by ELIXIR information systems. ELIXIR/MIAPPE is writing such lists based on common case studies, for discussion in EMPHASIS and EPPN^{2020} consortia for feasibility.

3.2 ANaEE (Ecology community)

An essentially similar approach was adopted with ANaEE, with meetings within the MIAPPE group which includes an Ecology component (22-23 May 2017 at IPK Gatersleben (see §3.1) and September 18-20, 2019, Paris. Specific meetings with the coordinators of ANaEE resulted in the writing of a common paper published in Nature Plants (13)

We fine- tuned our respective domains for data management. Briefly, AnaEE focuses on ecosystem-wide interactions between plants, soil biota and fauna, for the assessment and forecasting of ecosystems functional trajectories in response to global changes. AnaEE platforms involve the measurement of more variables than those in EMPHASIS, but over a limited amount of genotypes and with a lower throughput of plants. The ANaEE information system therefore follows MIAPPE standards, and have a structure largely similar to those in ELIXIR but with precise time courses of environmental variables on air and soil environment, including variables that characterize the biotic interactions, largely absent from EMPHASIS information systems. Measuring yield quantity and quality obtained in EMPHASIS for many genotypes allows to extend the results present in ANaEE information systems to assess simultaneously agricultural productivity and its environmental footprint, as well as their trends in response to global changes, in particular with biodiversity and climatic changes.

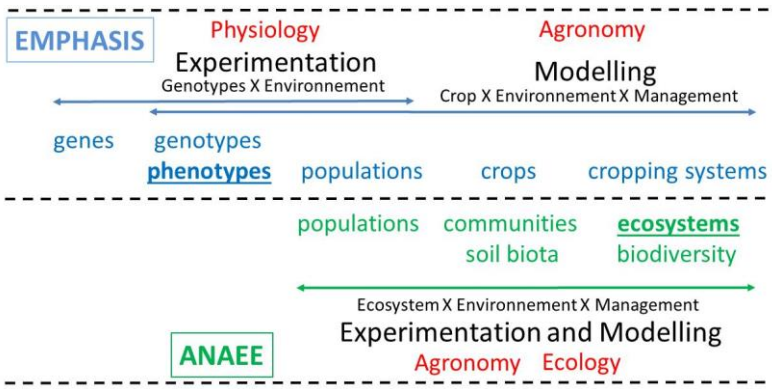

Fig.2 EMPHASIS and ANaEE domains of excellence for data collection and organization. Redrawn from Roy et al 2017, Nature Plants

3.3 AgMIP (Crop modelling community)

A similar approach was also used with the crop modelling community, in particular with meetings in Bruxelles (24-26 September 2018) with members of the AgMIP community and in Montpellier (17 April 2019) with AgMIP and CGIAR centres (CIMMYT and Biodiversity International). Another meeting will be held at the iCROPM conference of Montpellier in February 2020. The International Consortium for Agricultural Systems Applications (ICASA (9)) developed comprehensive standards for documenting and describing field experiments, providing detailed descriptions of management practices and traits of soils and plants for crop experiments. The ICASA standards emphasize common vocabularies, relations among variables, and the ability to implement the specifications in various formats. This ICASA dictionary was extended and modified for modelling use within AgMIP (e.g. the definition of new model variables that were not previously described) (10).

This repository is very similar to ontologies in EMPHASIS, so the main work consists in mapping terms between the two sources. A clear benefit for EMPHASIS is the access to adequate ontologies for management practices, developed by AgMIP and Biodiversity international. A common group was created for performing these operations.

Another crucial element was to perform a comparative mapping of existing models and phenomic information systems. In this context, an online portal referencing plant and crop simulation models was developed (UCLouvain, https://quantitative-plant.org/). This website maintains a catalogue of plant models that can be used to access or simulate new variables from phenomics data. It will raise the interest of the phenomics community in model-assisted phenotyping. Beyond the mapping activities, a task is currently under work to facilitate the exchanges of data between the phenomic information system PHIS and the platform of models developed by AgMIP.

3.4 Research Infrastructure projects and EOSC

In the EOSC framework, EMPHASIS is one of the 13 Research Infrastructures involved within the EOSC-Life cluster project coordinated by ELIXIR to develop digital space for European Life Science, Biological and Medical research addressing key societal challenges (https://www.eosc-life.eu). EOSC-Life deals with biomedical data that cross all scales from data collection, data organization and analyses with FAIR specificities with the overarching goal to enable reusability of these data and tools and by doing that fuel new scientific disciplines. This includes close interaction with many research infrastructures, development of common data management policy, user management, training etc. Specifically, EMPHASIS is involved in a demonstrator within the EOSC-Life project called A+ that links different life science infrastructures (EMPHASIS, ELIXIR, ISBE) by enabling the use and analysis of genomics, proteomics, metabolomics, and phenotyping datasets of tomato, potato and maize. The goal is to demonstrate phenotype to genomic data integration for analysis in a cloud environment.

Plant A+ will improve the toolset needed to document and give access to the datasets using FAIRDOM (https://fair-dom.org) and allow their findability using FAIDARE (https://urgi.versailles.inra.fr/faidare/), both tools being connected using the Breeding API (www.brapi.org).

Additionally, EMPHASIS closely connects with the Life Science Research infrastructure community within the CORBEL cluster project (13 Research Infrastructures, https://www.corbel-project.eu/) and the environmental community within the ENVRIPlus Research Infrastructure cluster project (26 Research Infrastructures, https://www.envriplus.eu/) that aim among other at identification of synergies in data management and development of shared services.

3.5 Data science community

The data science community is directly involved in EMPHASIS via national infrastructures as well as in the companion project EPPN[2020] , whose work package leaders for information systems belong to this community (in particular B Usadel, P Neveu and C Pommier). The MIAPPE group also involves both biologists and computer scientists, thereby allowing rich and efficient work to be done. This facilitated the emergence of common projects and common papers with the mathematics-computer science community, in particular with INRIA (French institute for computer sciences and robotics) (14, 15) and Excellence Cluster PhenoRob (http://www.phenorob.de/). These common projects aims at technology development for targeted monitoring and management particularly by machine learning and cloud computing techniques to analyse huge amounts of data. They formalize and improve the efficiency of workflows for data analysis(14) and optimize temporary (cache) storage and re-computing in clouds, thereby decreasing the duration of data analysis by e.g. half (15).

Pipelines and models are not a service of EMPHASIS but the information system allows embedding them, so the workflow is traceable and reproducible in cloud environment such as EOSC.

Conclusion

It would probably not be realistic to aim at a common information system for all communities in view of the tremendous complexity that the diversity of requirements would generate, but it is still essential that the different information systems can communicate via web-services. For example, it is thinkable that standardized workflows automatically extract genotypic means and environmental indicators used by the genomic community from the complex phenomic information systems used by the phenomic community.

We believe that the common activities with other infrastructures, the resulting position papers and academic papers and the tools for interoperability will soon result in interoperable information systems and common workflows of data analysis.

References

1. Tardieu F, Cabrera-Bosquet L, Pridmore T, & Bennett M (2017) Plant Phenomics, From Sensors to Knowledge. *Current Biology* 27(15):R770-R783.
2. Neveu P, *et al.* (2019) Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. *New Phytol.* 221(1):588-601.
3. Dhondt S, Wuyts N, & Inze D (2013) Cell to whole-plant phenotyping: the best is yet to come. *Trends in Plant Science* 18(8):433-444.
4. Krajewski P, *et al.* (2018) Towards recommendations for metadata and data handling in plant phenotyping. *J. Exp. Bot.* 69(7):1819-1819.
5. Cwiek-Kupczynska H, *et al.* (2016) Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods* 12.
6. Adam-Blondon AF, *et al.* (2017) Mining Plant Genomic and Genetic Data Using the GnpIS Information System. *Plant Genomics Databases: Methods and Protocols*, Methods in Molecular Biology, ed VanDijk ADJ), Vol 1533, pp 103-117.
7. Poorter H, Niinemets U, Walter A, Fiorani F, & Schurr U (2010) A method to construct dose-response curves for a wide range of environmental factors and plant traits by means of a meta-analysis of phenotypic data. *J. Exp. Bot.* 61(8):2043-2055.
8. Parent B & Tardieu F (2014) Can current crop models be used in the phenotyping era for predicting the genetic variability of yield of plants subjected to drought or high temperature? *J. Exp. Bot.* 65(21):6179-6189.
9. White JW, *et al.* (2013) Integrated description of agricultural field experiments and production: a The ICASA Version 2.0 data standards. *Computers and Electronics in Agriculture* 96:1-12.
10. Porter CH, *et al.* (2014) Harmonization and translation of crop modeling data to ensure interoperability. *Environmental Modelling & Software* 62:495-508.
11. Wilkinson MD, *et al.* (2016) Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3.
12. Millet EJ, *et al.* (2019) Genomic prediction of maize yield across European environmental conditions. *Nature Genet.* 51(6):952-+.
13. Roy J, Tardieu F, Tixier-Boichard M, & Schurr U (2017) European infrastructures for sustainable agriculture. *Nature Plants* 3(10):756-758.
14. Pradal C, *et al.* (2017) InfraPhenoGrid: A scientific workflow infrastructure for Plant Phenomics on the Grid. *Future Generation Computer Systems* 67:341-353.
15. Pradal C, *et al.* (2018) Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping. *Ercim News* (113):36-37.