Meeting EMPHASIS – ELIXIR 15 May 2018 :

Data standards and Information Systems: strategies of the European infrastructures EMPHASIS and ELIXIR

Participants : See appendix

Main authors:
C. Pommier and F. Coppens (ELIXIR)
P. Neveu and F. Tardieu (EMPHASIS)

I. Background and objectives

EMPHASIS and ELIXIR are two ESFRI infrastructures dedicated to Phenomics and Data sharing and integration respectively (Fig. 1-2). The meeting reported here involved the Plant Community of ELIXIR (EXCELERATE WP7) and (essentially) the data WP (WP4) of EMPHASIS and EPPN[2020].

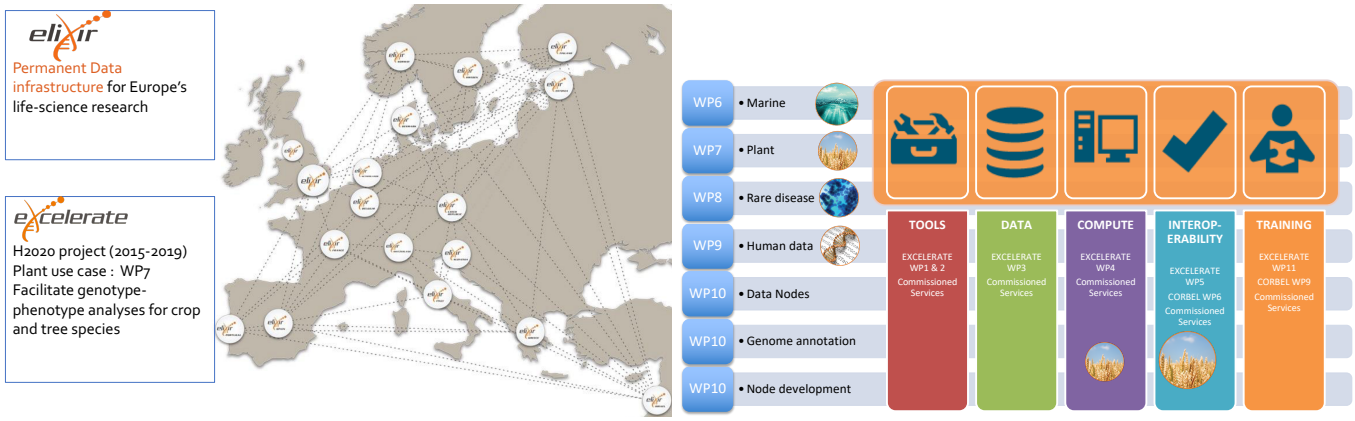

Fig. 1. Summary presentation of ELIXIR and the EXCELERATE grant

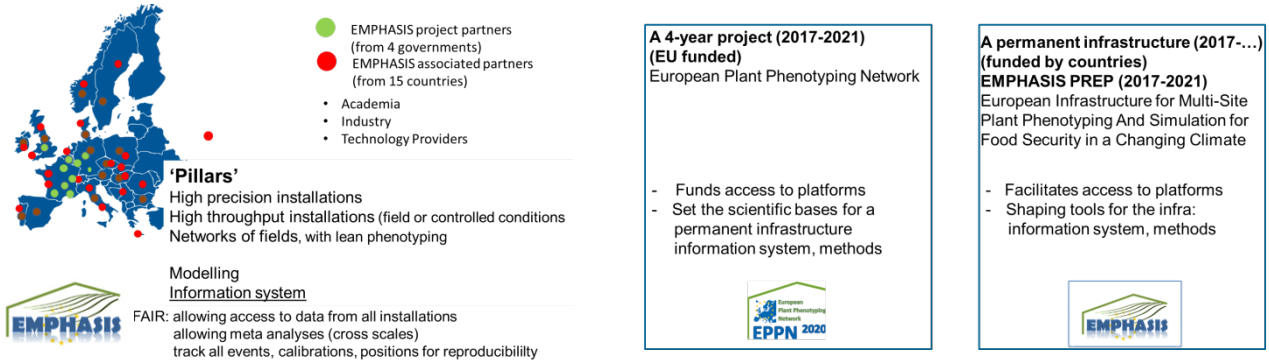

Fig. 2 Summary presentation of EMPHASIS, EMPHASIS PREP and EPPN[2020]

The main objectives of the meeting were to (i) clarify the strategies and respective roles of both infrastructure in data management, from data production to analyses and publication; (ii) look for complementarities, common tasks, in particular the joint work for MIAPPE (Minimum Information about a Plant Phenotyping Experiment) and strategies regarding calls; (iii) set the bases for the establishment of a Memorandum of Understanding between the two infrastructures and (iv)

potentially write a paper published in an academic journal, similar to that presenting complementarities between the ESFRI infrastructures AnaEE and EMPHASIS[1].

Beyond these objectives, discussions also focused on the nature of data handled by both infrastructures and the strategies in information systems. The meeting began with brief presentations of information systems in ELIXIR Plant and EMPHASIS

II. Data flow and respective domains of each infrastructure

The main domain of the EMPHASIS community concerning data (Fig. 3) is to:

- Produce datasets that jointly include phenotypic and environmental information, most often as time courses of variables[2]. An essential feature of these datasets is that they allow 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 or variable/ratios with biological meaning such as radiation use efficiency or stomatal conductance[2].
- 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 (e.g. stating the x-y positions of plants or plots and sensors over time, events during experiments etc) (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. Pipelines and models (e.g. for reconstructing 3-D shoot or root systems or to simulate fluxes) are not a service of EMPHASIS but the information system allows embedding them, so the workflow is traceable and reproducible.

The main domain of the ELIXIR community concerning data (Fig. 3) is to:

- Enable FAIR publication of datasets (Findable, Accessible, Interoperable, Reusable) to allow their use for phenotyping, genetic and genomic analyses. ELIXIR as an infrastructure handles all types of data produced in life sciences: the whole spectrum from raw (e.g. images, time courses) to processed data (e.g. scalars representing traits).
- 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, hosted by ELIXIR databases or not. Therefore, a dataset found through ELIXIR services could be accessed through a link to an EMPHASIS database hosting multidimensional phenotype and environment time courses and spatial distributions.
- 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, possibly computed from EMPHASIS data, i.e. data matrices handling values for traits (i.e. mean/minimum/maximum rates, conductance, yield, biomass or size) observed on each repetition (plant, micro plot) or genotype. Those traits are generally computed for reuse in a given scientific question.
- 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 above and in Fig. 3 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. However, it seems essential to define here the 'domains of excellence' of each community in order to better identify common tasks.

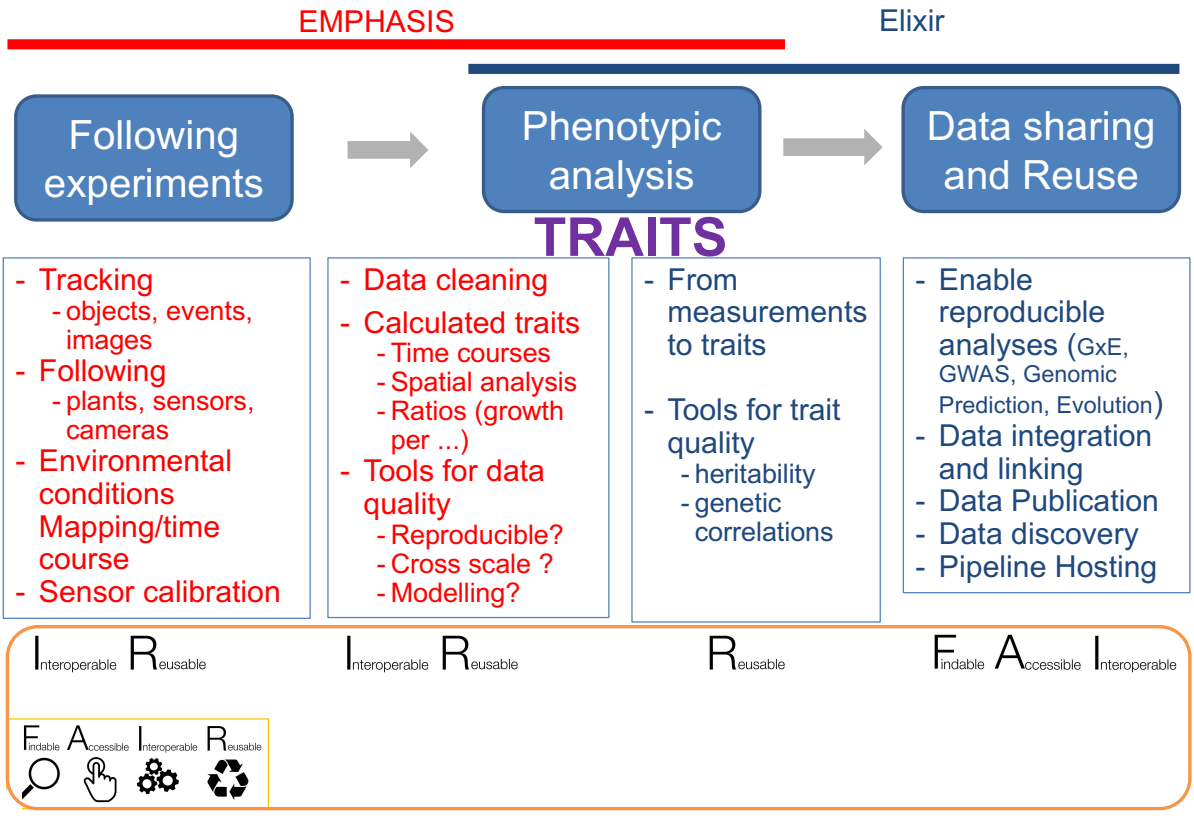

Fig. 3 Schematic representation of the flow of data, from experiments to genetic analyses.
Noteworthy, another continuity between data production/analysis and modelling exists between the communities of phenomics (EMPHASIS) and of crop modelling (AGMIP).

III. Information systems in each infrastructure

EMPHASIS is organized in such a way that full datasets (phenotypic, environmental and meta data) will eventually be accessible with FAIR requirements for a wide community for genetic analyses and modelling. This is an on-going process, currently under development in the EMPHASIS PREP and

EPPN^{2020} projects. Briefly, the information system will involve (i) local information systems that collect, organize and store datasets from local nodes of EMPHASIS (e.g. PHIS in Montpellier or PIPPA in Ghent), (ii) an 'EMPHASIS' layer that connects local information systems, provided that those have the necessary properties for being connected (identification, ontologies). Both layers are connected to existing ontologies and ontology repositories (e.g. Crop Ontology, Agroportal) and external APIs (e.g. Breeding API).

ELIXIR is organized as a distributed information system enabling FAIR principles across a federation of data repositories. These data repositories layer covers several datatypes, i.e. genomic, genetic, phenotyping, publications. The second layer is the ELIXIR Data Lookup Service that provides Findability and Accessibility, i.e. a single point for accessing all the ELIXIR federation data. It relies on the use of Interoperability means (ontologies like the crop ontology, dataset and biological material identification) and standard API (Breeding API) and general data standards (MIAPPE, MCPD, …). This Service is highly customizable with the possibility to instantiate not only the ELIXIR portal, but also community portals like the WheatIS or portals dedicated to other infrastructures such as EMPHASIS.

IV. Common tasks identified during the 15 May meeting

4.1 MIAPPE

In the MIAPPE meeting held at Gatersleben (DE) in May 2017, it was decided that ELIXIR and EMPHASIS will both be full participants to MIAPPE, alongside the CGIAR. This is one of the points that could be stated in a MoU. Further, tasks have been defined.

- *Improving environmental characterization (EMPHASIS→ELIXIR).* The list of environmental variables in MIAPPE has been the object of debates, and too specific variables have been removed from the MIAPPE list. It is proposed here that a group in EMPHASIS/EPPN^{2020} proposes an approach and a list of requirements with two levels (essential/desirable), which could be based on the levels 1 and 2 defined in EPPN^{2020} as requirements for installations to provide accesses funded by the project. This will be provided to the MIAPPE group as a draft that will be discussed for feasibility / consistency with the MIAPPE policy. Case studies (often already existing) will then be worked jointly.

- *Defining "abstract datasets" in EMPHASIS databases (ELIXIR→EMPHASIS).* The difference in the nature and structure of datasets identified in section II highlights the necessity for EMPHASIS databases to define lists of single point scalars extracted from time series and spatial data from full datasets, which can then be used in other analyses including genetics and climate change studies (e.g. total fluxes, rates, biomass, LAI at anthesis, conductances etc.). 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. It is proposed that ELIXIR/MIAPPE propose such lists, which will be discussed afterwards in EMPHASIS and EPPN^{2020} consortia for feasibility. Case studies will be worked jointly afterwards.

4.2 Interoperability between information systems

The two above mentioned tasks will facilitate interoperability, but a large effort is still needed in such a way that ontologies and formats are built in common. This will be largely facilitated by the common use of ontologies (e.g. crop ontology) and APIs (e.g. BrAPI), but a common work is needed to ensure full usability by both infrastructures. This work could be done in the contexts of MIAPPE and EPPN^{2020}.

4.3 Publication policy

A joint effort may be useful for defining a strategy to convince editors of scientific journals and funding agencies for minimum requirements for phenotyping data publications. This needs to be handled carefully, based on user's experience and straightforward principles, probably the 'minimum-minimum requirements' in such a way that this is acceptable, and that users of both communities find them feasible and useful.

Reciprocally, it is probably necessary to ensure that the information systems in EMPHASIS and ELIXIR are recognised as acceptable repositories for datasets. A risk exists otherwise that non-European repositories are mandatory for publishing results and datasets in international journals.

V. Next steps

1. The current document will be discussed in the executive committees of EMPHASIS-PREP, ELIXIR, MIAPPE and EPPN2020.
2. After amendments, this text will serve as the basis of a Collaboration Strategy , and possibly a Memorandum of Understanding, between ELIXIR and EMPHASIS-PREP signed by respective coordinators.
3. This document might also be the base for a scientific paper.
4. The tasks identified here will be carried out first by first responsible groups identified in §4, then by working groups for final presentation to executive committees.

Appendix Participants to the meeting

Abbeloos Rafael	VIB	raabb@psb.vib-ugent.be
Alary Pierre Etienne	INRA/EMPHASIS	pierre-etienne.alary@supagro.fr
Cabrera-Bosquet Llorenç	INRA/EMPHASIS/EPPN2020	llorenc.cabrera-bosquet@inra.fr
Coppens Frederik	VIB/ELIXIR	frederik.coppens@ugent.vib.be
Cwiek-Kupczynska Hanna	IPG PAS/ MIAPPE/EPPN2020	hcwi@igr.poznan.pl
Dhont Stijn	VIB/EMPHASIS	stdho@psb.vib-ugent.be
Fahrner Sven	FZJ/EMPHASIS	s.fahrner@fz-juelich.de
Neveu Pascal	INRA/EMPHASIS/EPPN2020	pascal.neveu@inra.fr
Pieruschka Roland	FZJ/EMPHASIS/EPPN2020	r.pieruschka@fz-juelich.de
Pommier Cyril	INRA/ELIXIR	cyril.pommier@inra.fr
Pompe-Novac Marusa	NIB	marusa.pompe.novac@nib.si
Rossello Marc B	EBI-EMBL/ELIXIR	mrossello@ebi.ac.uk
Tardieu Francois	INRA/EMPHASIS/EPPN2020	francois.tardieu@inra.fr
Turdukulov Ulan	WUR	ulan.turdukulon@wur.nl
Vadez Vincent	IRD/ICRISAT	vincent.vadez@ird.fr
Yemadje Lammoglia Karen	INRA	sabine-karen.yemadje@inra.fr

1 Roy, J., Tardieu, F., Tixier-Boichard, M. & Schurr, U. European infrastructures for sustainable agriculture. *Nature Plants* **3**, 756-758, doi:10.1038/s41477-017-0027-3 (2017).

2 Tardieu, F., Cabrera-Bosquet, L., Pridmore, T. & Bennett, M. Plant Phenomics, From Sensors to Knowledge. *Current Biology* **27**, R770-R783, doi:http://dx.doi.org/10.1016/j.cub.2017.05.055 (2017).