EMPHASIS

# D4.5:Data policy rules

**Francois Tardieu, Pascal Neveu, Cyril Pommier Björn Usadel. | 25 May 2021**

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°** | D4.4 | **Title** | Strategy: Data Management Plan |
| **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** |
| 5/2/2021 | 1 | Francois Tardieu | First version |
| 25/2/2021 | 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.

# Table of contents

## Executive Summary

This document presents the plan for managing the datasets generated and processed during and after experiments carried out in the EMPHASIS infrastructure. The main objective is that datasets are findable, accessible, interoperable and reusable (FAIR standard), in such a way that the datasets can be analysed by several groups inside and outside of the EMPHASIS consortium. The data management plan presents the different data categories, data sources and how data are collected, structured, stored and made accessible for the purpose of analysis and reuse. It presents unambiguous identification of all objects involved in experiments, the rules for variable naming and data storage, together with standards for data identification, file formats, and workflows. The document also presents the ownership of the different categories of data collected in experiments, with the rationale of optimizing analyses by involved groups (users and providers), and by other groups of the scientific community.

The set of rules presented here describe the management of datasets generated and processed during and after experiments carried out in EMPHASIS and set up the general reuse rules, authorship traceability and legal aspects. It will help partners to manage data, meet funder requirements, and facilitate multiple use of data by the scientific community, hence applying the “findable, accessible, interoperable and reusable” (FAIR) principles.

## 1. Data description

**Definitions:**

*Digital Object Identifier* (DOI): A persistent identifier for an object or a document that can be handled by a resolution service to direct communications to the correct server. Developed by the International DOI Foundation (www.doi.org). Typically used for identifying whole datasets upon publication, or for identifying plant genetic resources.

*Uniform Resource Identifier* (URI). A persistent identifier for an experiment or any object (plant, pot, vector, sensor) involved in experiments.

*Metadata*: Information about data stored in a repository/database.

*Repository*: A digital repository is a mechanism for managing and storing digital content. Repositories can be subject or institutional in their focus

*Embargo period*: A period of time from the end of the experiment, during which the access to data collected in an EMPHASIS installation is limited to access users. For academic projects, it is defined in the Data Management Plan of the corresponding project and should not exceed five (5) years unless explicitly mentioned in the access convention. For industrial accesses, it is defined in the access convention and should not exceed ten (10) years.

*Open access*: data are not protected by copyright or patent, and can be used by any scientist, for example under the licence CC-BY https://creativecommons.org/licenses/by/4.0/

*Dataset*: Digital information created in the course of research but which is not a published research output. Research data excludes purely administrative records. Two categories of datasets originate from experiments in EMPHASIS

- *The collection of images, sensors, outputs and observations collected during an experiment, together with the metadata that accompany them*, in particular URIs and variable names considered with the "quadruplet" trait (entity, characteristic), method, unit (e.g. meristem_temperature_thermocouple_°C ; plant_height_image-analysis_m). Variables are most often time dependent (e.g. temperature every hour or intercepted light every day). These datasets are organized in information systems that relate all these elements and make them FAIR. This is, explicitly, in the domain of EMPHASIS, aiming at giving the possibility to any plant scientist to reanalyse datasets. These datasets are organized and stored by EMPHASIS local infrastructures, via the EMPHASIS information systems stored in a cloud (e.g. European Grid Infrastructure)
- *The collection of analysed data*, e.g. genotypic means after correction for spatial variability cumulated or averaged at a given time (e.g. mean leaf area at flowering time) or environmental indicators for a given period (e.g. number of hours with a temperature higher than 32°C during 10 days encompassing flowering time). These datasets typically link with a published paper. They are organized following the MIAPPE rules, and essentially stored in repositories compatible with ELIXIR-plant.

**Data categories and sources that need to be managed**:

- Genetic resources: species, genotype, seed origin, accession.

- Facilities: installations, sensors, cameras, vectors (e.g. conveyors or drones), specific devices.. See Figure 1
- Characteristics of experiments e.g. design, protocol and spatio-temporal organisation.
- Phenotypic data at plant or population level as collected by sensors (e.g. raw images or other sensor data), or manually (e.g. phenological stages).
- Environmental conditions as collected by sensors (e.g. soil water status, air temperature or evaporative demand).
- Date and description of management events
- Workflows: sensor and image analysis methods and software tools used to extract traits from raw image and other data.

Collected data can be numerical data, images, documents, texts or manual measurements. In each experiment, data are collected for typical periods of 20-100 days, including raw data (sensor or camera outputs) and curated and computed data (validated at further steps of analyses or resulting from analysis of raw data).

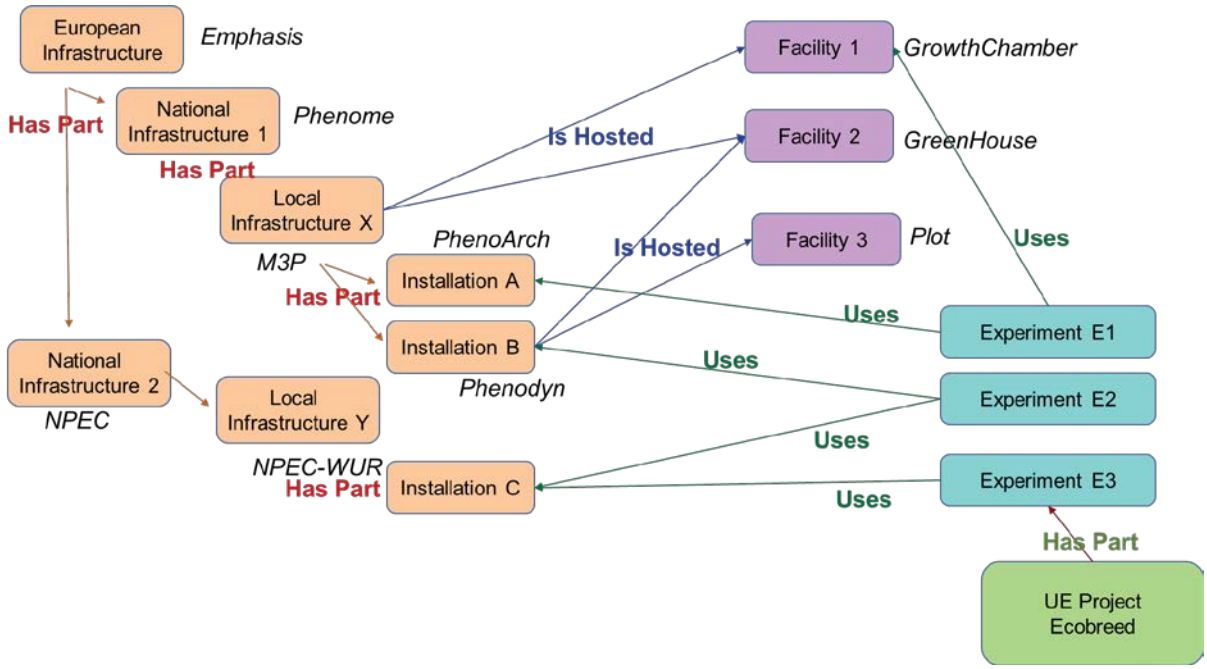



*Figure 1. Organization of infrastructures and experiments concepts. Emphasis is the European infrastructure, which involves national infrastructures, composed of local infrastructures. Each local infrastructure involves installations (e.g. phenotyping robot or equipped fields plus the relevant devices. They also involve facilities that may host the installation, e.g. greenhouse or growth chamber, or that can be used in experiments in addition to installations (e.g. vernalization chamber). Experiments are part of projects, and use facilities or installations.*

## 2. Data organization

An integrated information system is currently under development in the frame of EPPN[2020]. It aims at hosting datasets produced in all categories of platforms in EMPHASIS.

### *2.1. Its main features are that*

- The references and names of all objects involved in experiments (e.g. plants, sensors or images) are standardized and unambiguous. This is essential to trace these objects in further analyses, including those performed by groups not involved in experiments. Identification

systems are based on persistent unique identifiers for the objects mentioned above. The preferred technical solution is the use of URI which are progressively deployed in all local infrastructures of EMPHASIS. Files and folders are versioned and structured by using a name convention. They are all accessed from the Information System to enable efficient findability and grouping. Each Experiment being a consistent dataset, it also receives an identifier, preferably a DOI or an URI.

- Multiscale data integration is possible through events tracking and elements identification. For instance, tracking the x-y-z position of a given sensor in an installation (greenhouse or field) as well as all of its different calibrations allows one to correctly estimate the environmental conditions spatial distribution during experiments. The same applies to the x-y position of plants in platform experiments and of microplots in field experiments. Another important information is the movement of plants between facilities, especially in the case of perennial plants that spend a small part of their lives in a platform. In addition, provenance management makes it possible to know how and when the data was obtained (e.g. sensor used, agent in charge, data transformation software).

- Data security aims to avoid data loss: In all local and national infrastructures belonging to EMPHASIS, data will be stored with duplication in at least two locations (different building, or better different sites when possible). Data will be saved daily with backup on a remote location. Backup should be checked at intervals of two weeks.
- Data access is sustainable: for longer-term storage and data sharing, EMPHASIS progressively uses the European Grid e-infrastructure. We aim at preserving datasets for at least ten years in the dedicated EMPHASIS information systems that will provide both web access and documentation of the data as well as data download in standard format.

*2.2. A three-level standard was defined in the companion project EPPN$^{2020}$ and will be used in EMPHASIS*.

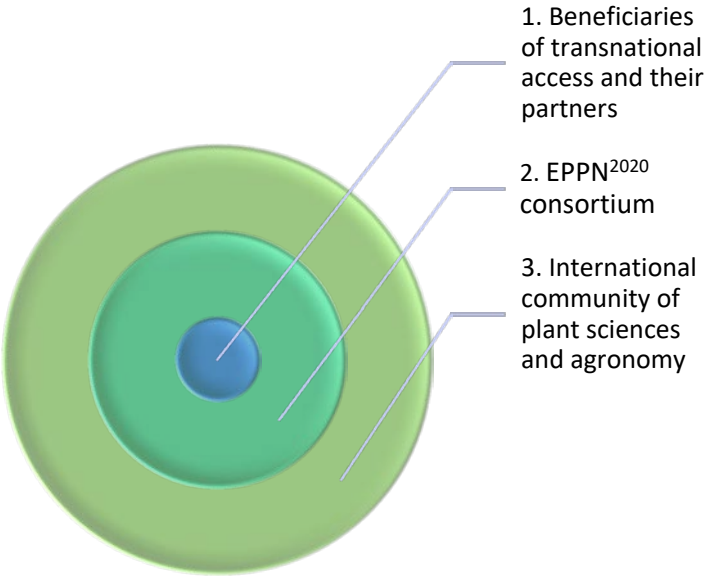
Level 1 deals with the F (findable) and R (reusable) of the FAIR principles and is necessary for a local infrastructure or an installation to be labelled by EMPHASIS. This requires that all objects (plants, sensors, vectors) are identified and that measured variables have unambiguous naming in each installation. Case studies and software tools were distributed in the consortia of EMPHASIS PREP, EPPN$^{2020}$ and beyond to help users to reach this step. It is now largely reached among members of the EPPN$^{2020}$ consortium. In addition, each experiment must be associated with metadata in a commonly agreed form so dataset are findable (Title, PUI, authors and roles, location, summary description) or reusable (license). Metadata follow the MIAPPE metadata standard for the Study and experimentation level.

Level 2 should be reached as rapidly as possible by all local infrastructures of EMPHASIS in order to guarantee that data are accessible and fully reusable (R). It consists in using local information systems to trace, organize, integrate and visualize phenotypic data together with the necessary metadata. Of particular importance is the identification of all sensors, vectors and plants allowing one to standardize metadata in an efficient way; and the generation of machine-readable variable names connected with public ontologies. This allows one

Level 3 enables global findability and data access across Europe through web data portals. The first step will be deployed at the end of the projects EMPHASIS PREP and EPPN$^{2020}$. It involves metadata indexing using the FAIDARE data portal. This allows any user to identify experiments, in the EMPHASIS local information systems, which involve particular features such as using specific genotypes, describing specific traits or comparing specific environmental conditions. A protocol for off-line data exchange is associated to this software. A further step will allow users to build, on line, an ad hoc dataset responding to specific objectives, based on datasets present in local information systems in EMPHASIS.

### 3. Data ownership and data sharing

Access to the data is organized using three circles of users, namely beneficiaries of access and their partners who require an experiment on a local infrastructure, the Emphasis consortium and the international community of plant sciences and agronomy. The objective is that the three categories of users potentially have access to all datasets, with the necessary metadata and information for the datasets to be reusable.




The scientific leaders of local infrastructures will be responsible for managing the data and ensuring that the data management plan is carried out.

The ownership and rules of release of the datasets generated during experiments primarily depend on the data management plan of projects that fund accesses. However, the following rules are strongly suggested, in which data ownership depends on categories of data. The rationale of these rules is to optimize data use and to facilitate meta-analyses.

- For scientific projects, phenotypic data (e.g. images, measurements, observations) belong to the infrastructure user. An advisable procedure is that the local infrastructure group is associated with data analysis and publication in order to obtain the best possible analyses. Resulting publication are published in open access, and associated datasets are made accessible via public repositories that benefit to open access and DOI. Datasets have themselves a DOI, and cite the DOI of the National Infrastructure(s) and of the installation(s) in which experiments were carried out. Publishing datasets as supplementary information in the journal website is strongly discouraged. All datasets should have a license of reuse, preferably an open one from the Creative commons list (CC-BY-SA suggested, CC0 to be used with caution and avoided as much as possible).

- For technological projects, data are published whenever feasible. The main results may also be made available to a large public via EMPHASIS services. SMEs are not obliged to diffuse their data to a larger circle but will be encouraged to.

- Four categories of data will remain the property of the installation providers and made available to the installation users for a given access.

(i) Environmental data collected during the experiment belong to the local infrastructure, in such a way that this group can perform meta-analyses of environmental data over seasons and years.

(ii) Calibration results and calibration procedures for sensors and cameras need to be analysed across experiments and years, so they belong to the local infrastructure.
(iii) Trait recovery workflows and procedures used to extract phenotypic measurements from raw sensor data that are developed by the local infrastructure
(iv) Innovations and packages developed to optimize experimental designs within the installation and the statistical tools to optimize them belong to local infrastructures. All this information is made available to the users and, once published, to the whole community.

**4. Publication**

When relevant, partners will share datasets in a publicly accessible disciplinary repository using descriptive metadata such as MIAPPE. Additional metadata will be stored and made available within a separate XML, JSON or RDF file in a standardized way by using machine readable schema or ontologies. Keywords will be added by using standardized controlled vocabularies.

Datasets will be made 'Findable' and 'Accessible' by using metadata and data repositories such as dataverse (data.inrae.fr) or eDale!. Most often, data repositories will make available synthetic data, such as genotypic means after correction for spatial variations, or averaged environmental conditions after detection of outliers. It is firmly discouraged to publish such datasets in journal's websites as supplementary information. Conversely, the high volume data will remain in the local infrastructure information systems, as well as most of the complex data including high density time series

Public groups will publish software codes along with datasets in a disciplinary repository. Whenever possible, analysis will be performed using freely available open source software tools.

Datasets corresponding to public projects are by default managed following the data management plan of the corresponding project. They should be publicly available in a disciplinary research data repository along with scholarly journal and open access publications after the primary publications, **in all cases five years after the end of the experiments**. All effort will be dedicated to make datasets understandable for other researchers by using standards and metadata.

Datasets originating from accesses granted to industry will be handled following the access convention agreed before the access. **The embargo period cannot exceed ten (10) years**.

**5. File formats**

We aim at producing data files with 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:

- Text files: txt, markdown, avoid binary format (ODF, PDF). No proprietary format (Word)
- Tabular data: CSV, ASCII (not Excel)
- Images files: PNG, TIFF, JPEG2000, GIF (classical JPG to be avoided because of a risk of obsolescence and of loss of information)

- Structured metadata: JSON, XML or RDF

## 6. Data storage

Storing and organizing datasets produced in EMPHASIS is a challenge in view of (i) the geographical distribution of local infrastructures across Europe, (ii) the specific characteristics of installations dedicated to particular species or scientific topics and (iii) the evolving nature of phenotyping platforms. In 2018 the cumulative data volume for the EPPN$^{2020}$ partners ( 35 installations) exceeded one Petabyte.

A distributed and scalable storage system for plant phenotyping experiments needs to be based on approved distributed architecture such as OneData or IRODS. The iRODS distributed open source data management software (Rajasekar *et al.*, 2010) is designed to enable policy-based distributed data management across the data lifecycle. The Onedata distributed system allows users to access, store, process and publish data using global data storage backed by computing centers. Onedata focuses on instant, transparent access to distributed data sets, without unnecessary staging and migration, allowing access to the data directly from local computers or work node. OneData or iRODS selection depends on national node of the European Grid Infrastructure (EGI). Local storage is also another solution, which cannot be recommended as a flexible and scalable solution. The international iRODS consortium supports ongoing development and evolution of iRODS thus guaranteeing long-term sustainability. It is currently used by many groups in a large spectrum of scientific domains. For instance, iRODS supports more than 20 petabytes at the Welcome Trust Institute, 6 petabytes of data at the French IN2P3, several thousands of users in the US iPlant collaborative project. The iRODS solution provides the following features:

- **data distribution:** Physical storage resources can be distributed on geographically separated locations. Data can be replicated on several locations for security or accessibility questions. Replication allows to have reliable backups. It also improves the speed of data transfers and the availability.
- **data virtualization**: Multiple resource servers and the metadata catalogue can be connected to a unified iRODS (or OneData) data Grid. For instance, this allows a better integration of new hardware.

***It is recommended, at this stage, that EMPHASIS local infrastructure use the*** distributed system supported by the European Grid Infrastructure (https://www.egi.eu/). EGI provides a sustainable set of IT services that will makes easier deployment and interoperability. This approach allows supporting the hardware technology evolutions.