



Deliverable D6.20 – Activity 6.09, CNR-IPSP-NA

Tools for novel approaches to identify phenotyping pipelines to assess crop properties across scales from lab to field combined with central access to coordinated information systems for data

Authors: André P.M. Fabbri, Sabrina Mazzoni, Giulia Atzori, Mauro Centritto, Michelina Ruocco, Josè Saporita

30 June 2025



Deliverable number:	D6.20
Work package:	WP6 – Terrestrial Biosphere
Intermediate Objective:	IO1.1
Deliverable type:	<input checked="" type="checkbox"/> Document, report
	<input type="checkbox"/> Websites, patent filings, videos, etc.
	<input type="checkbox"/> Other: please specify
Dissemination level:	<input checked="" type="checkbox"/> Public
	<input type="checkbox"/> Restricted
Estimated delivery (bimester):	28/02/2025 (B14)
Actual delivery date:	30/06/2025 (B16)
Author(s) (Partner-OU):	André P.M. Fabbri, Sabrina Mazzoni, Giulia Atzori, Mauro Centritto, Michelina Ruocco, Josè Saporita
Reviewed by:	ITINERIS Executive Board
Note:	

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)

Funded by EU - Next Generation EU

Mission 4 “Education and Research” - Component 2: “From research to business” -

Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Table of contents

<i>1</i>	<i>INTRODUCTION</i>	<i>4</i>
1.1	Purpose of the document	4
1.2	Proposal	4
1.3	Structure of the computational infrastructure	5
1.4	Definitions, acronyms and abbreviations	6
<i>2</i>	<i>DATA ACQUISITION INSTRUMENTS</i>	<i>7</i>
<i>3</i>	<i>ON-PREMISES INFRASTRUCTURE</i>	<i>7</i>
3.1	Hardware resources	7
3.2	Software resources	9
3.2.1	Field-work tools	9
3.2.2	Post processing.....	9
3.2.3	Statistical Analysis.....	10
3.2.4	Remote Access Control.....	10
<i>4</i>	<i>CLOUD INFRASTRUCTURE</i>	<i>10</i>
4.1	Virtual Research Environment	10
4.2	Data Portal	11
<i>5</i>	<i>CONCLUSION</i>	<i>11</i>

1 INTRODUCTION

1.1 Purpose of the document

The current document describes the main characteristics of what will be the final deliverable of the Activity 6.09 of the ITINERIS project (to be delivered at the bimester 16) and processes to create pipelines from the lab to the field and vice versa.

This document has been developed in the context of the collaboration with the Activities 6.12, 6.18 and 6.10 of the ITINERIS project, which are also linked to the deployment and enhancement of PhenItaly, the novel Research Infrastructure (RI) serving the Italian Phenotype Research Community. PhenItaly is itself part of EMPHASIS, the European initiative aimed at fostering collaboration within the European plant phenotyping community and stakeholders. One of the key objectives pursued by this initiative is the analysis of crop performance with respect to structure, function, quality, and interaction with the environment and exploitation of the crop genetic diversity required for enhancing plant productivity and plant breeding innovations. It provides 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.

Researchers from the Italian National Research Council's Institute for Sustainable Plant Protection (IPSP-CNR) as part of the Italian node of EMPHASIS, have been developing a novel and comprehensive phenotyping Research Infrastructure (RI): the Digital Ecosystems for MEtabolomics, plant Trait Research, and Imaging Systems Hub (DEMETRIS-Hub). This includes a range of state-of-the-art, environmentally controlled platforms as well as advanced field-based systems, offering an integrated approach to plant phenotyping.

Such high-throughput plant phenotyping infrastructure requires a suitable computational infrastructure to support data acquisition, processing and extraction pipelines. We present hereby the computational solution adopted for the DEMETRIS-Hub ranging from the field instrument software up to the data diffusion on cloud platform. In this Deliverable the overall computational architecture is detailed presenting the different types of computational resources and their interaction whereas Deliverable 6.12 focuses more specifically on cloud platform resource as a crucial element to support in the future the Research Infrastructure.

As the number of phenotyping platforms increases across Italy, this deliverable outlines a possible computational infrastructure to develop such pipelines that may be potentially extended to the Italian Plant Phenotyping community.

1.2 Proposal

High-throughput plant phenotyping enables non-invasive, fast evaluations of a large number of plants for size, development, and physiological variables. These state-of-the-art infrastructures capture plant and environmental data using digital sensors and automatable carriers and, therefore, produce so-called big data: frequently, high volume of data that are also diverse according to the sensors or their combination with the carrier.

As any automated process, high-through plant phenotyping requires standardized pipelines to cope with data production. Following that perspective, pipelines range from the data acquisition itself of the sensor, the post processing of the raw data to produce exploitable data and, finally, the analysis or data extraction. In this context, pipelines are a determinant aspect of data validation and should be continuously updated, incrementally enhanced and shared to the research community along with data.

Nonetheless, the computational infrastructure should be versatile enough to cope with the variety of usage, data and lifecycle requirements for such pipelines. Indeed, the development and implementation of pipelines requires tools and skills from Computer Science and Information Technologies that will be progressively incorporated into Plant Phenotyping practices along with the current usage of computational resources. Furthermore, the diversity of data captured by cutting-edge sensors requires adaptive resources to support different use cases and, similarly, the computational resources vary according to the maturity of the pipeline: from prototype to long term solutions.

1.3 Structure of the computational infrastructure

Computational architecture involves various typologies of resources that will be split into three main typologies:

- Data acquisition instruments are computers embedded in field instruments. They are specific to each product.
- On premises resources are general purpose computer or specific software provided locally within the research infrastructure. The choice of software and hardware characteristic depends on the processing requirements.
- Cloud resources are services provided online and are operated on hardware resources provided by cloud providers. The configuration and implementation of services depends on the usage and functional requirements.

The typologies of resources have been combined to support a standard pipeline workflow as presented in Figure 1. Data Acquisition computational resources are usually the first step of any pipeline by providing raw data to be further processed. On premises resources will facilitate pipelines' prototyping and their quick deploy on site: closest from the field. The first steps of pipelines such data acquisition or immediate post-processing may remain on these types of resources in the long run. These resources are limited and require circumscribed maintenance. On the other hand, a cloud infrastructure aims to create a dedicated community to develop and share open-source tools. The last steps of pipelines such as data analysis and extraction are more suited to be developed on such platform but, in the long run, post-processing may expect to run directly on the cloud. These resources are potentially unlimited, on-demand and upscale with a reasonable maintenance effort.

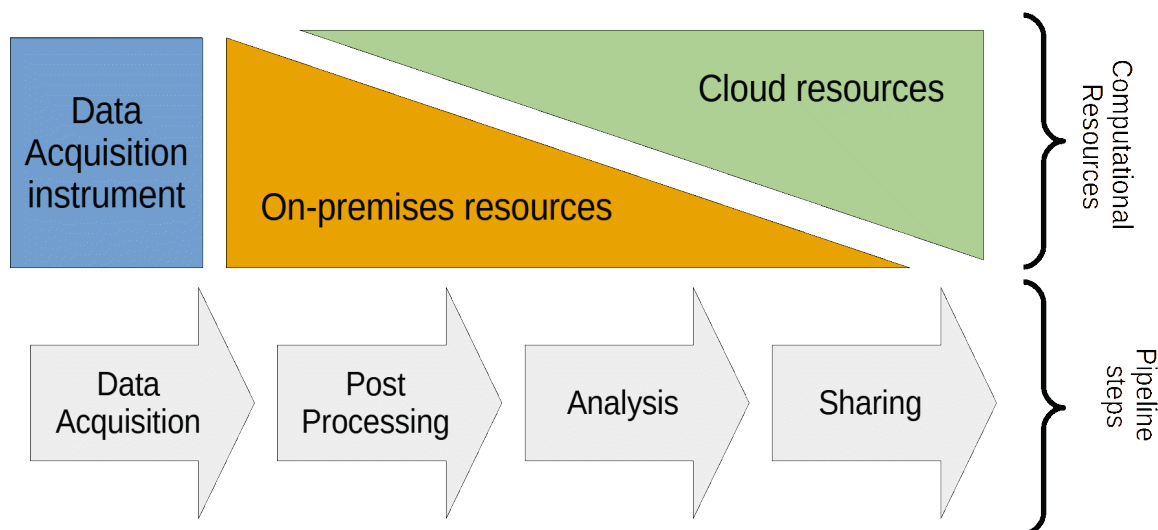


Figure 1. Computational resources to support pipelines development.

1.4 Definitions, acronyms and abbreviations

- API: Application Programming Interface
- CSV: Comma Separated Value
- EMPHASIS: European infrastructure for Multi-scale plant PHenomics And Simulation for food Security in a changing climate
- FAIR: Findable, Accessible, Interoperable, Reusable
- GB: GigaByte
- HTPP: High Throughput Plant Phenotyping
- ITINERIS: ITalian INtegrated Environmental Research Infrastructure System
- LIDAR: LIght Detection And Ranging
- PhenItaly: Italian Plant Phenotyping Network
- RAM: Random Access Memory
- RGB: Red-Green-Blue
- RI: Research Infrastructure
- SaaS: Service as a Software
- SLAM: Simultaneous Localization And Mapping
- TB: TeraByte
- UAV: Unmanned Aerial Vehicle
- VRE: Virtual Research Environment

2 DATA ACQUISITION INSTRUMENTS

Data acquisition instruments generate raw data that represent the input for any subsequent pipeline. Digital instrumentations are provided with an embedded computer with dedicated software to operate them and retrieve the data generated, leading to the development of a specific post-process for each instrument. Indeed, each manufacturer provides a different data model based on the features and capability of the instrument, software is tailored for the instrument and partially adaptable to other contexts. The common baseline of these instruments is the data format that usually refers to open standards, when possible.

The sensors purchased for the experimental platform (deliverable D6.7) are reviewed to highlight their main differences from a workflow perspective.

- Field handheld instruments: the data produced are simple CSV files while more advanced save in spreadsheet format directly. These instruments are manually operated and, therefore, the main concern is to automate the association between the dataset and the experimental condition of the plot.
- Field UAV/Rover/Overhead platform: usually these instruments carry various types of optic sensors. The data produced are usually multilayered raster for the camera (RGB, Thermal, multispectral hyperspectral, Fluorescence) or cloud point for radar (LIDAR, SLAM). These carriers are not fully autonomous and therefore an operator is still required. The main concern is to automate the consecutive post-process that may require more advanced computational resources and technical expertise.
- Growth capsule: with respect to field instruments, these environments control the growing condition. The data produced depends on the embedded optic sensors (RGB, Fluorescence) and all the environmental sensors (temperature, humidity, etc.). These ambient can be fully automated. This instrument is almost fully automated, the main concern is to configure it and integrate it with the overall architecture.
- Metabolomics analyzer: The data produced are important tabular data, CSV files or similar proprietary file format. These complex instruments are manually operated, and the process can be time-consuming, generating a very large amount of data that needs further processing. The main challenge is identifying relevant features within the data and integrating them effectively.

3 ON-PREMISES INFRASTRUCTURE

The on-premises infrastructure is intended to support the first steps of the pipelines and provide quick prototyping for the last steps of the pipelines. In this section we detailed the hardware and software resources purchased for such purposes.

3.1 Hardware resources

In the IPSP laboratory of Sesto Fiorentino, several computers have been purchased for temporary storage and data processing. What follows is a list of the main elements of the in-house hardware infrastructure and how they take part in the workflow from the sensors to the data repository.

- 1 heavy load workstation with a 256 GB RAM, 1TB fast Solid State Disk, 24TB Hard Disk and a A4000 GPU with 18GB dedicated.

- 2 light load workstations with 64GB RAM, 2TB fast Solid State Disk and NVIDIA 4090 with 24GB dedicated.
- 1 Network Accessible Storage (NAS) station with 48TB of hard disk.
- 2 different rugged tablets of 8” and 12” size.
- 10 individual laptop terminals

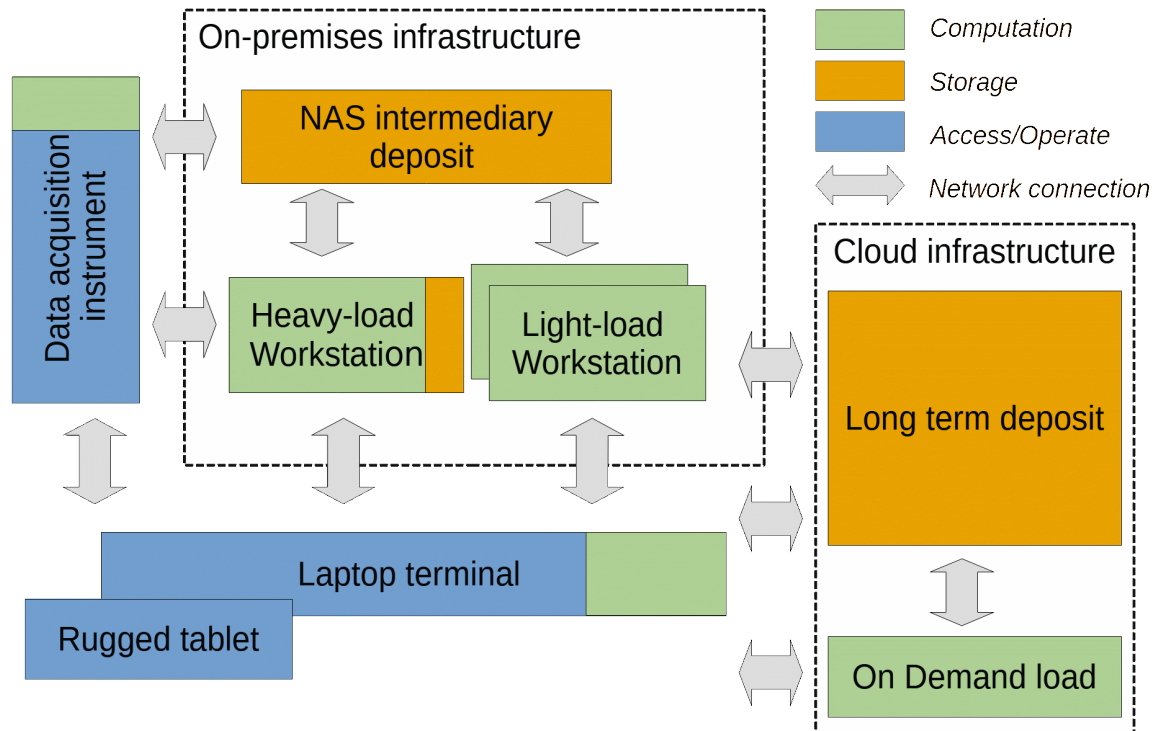


Figure 2: Computational infrastructure overview

Starting from the field, the rugged tablets are meant to collect data directly from the field, associate GPS coordinate or, when available, associate immediately the unique identifier of the plot to the measurement (through QR code scanning). These tablets are also necessary to operate field instruments, such as the sensors on the rover or handheld spectroradiometer.

All the data gathered by the various sensors need to be processed further, checking both the validity of the measurement and additional information regarding the experiment. For all devices, the NAS will serve as a first deposit for raw material. The NAS may store up to an estimated one-year research data (about 48 TB of space available). The temporary data is stored and retrieved by any computer of the local infrastructure: an individual laptop, a rugged tablet, a workstation or the computer operating automated instruments.

The workstations have been chosen for their Central Processing Unit, Random Access Memory and Graphic Processing Unit performances. The light load workstations serve for image post-processing and to train straightforward machine learning models while the heavy load workstation will perform the long metabolomic matching procedure or to train deeper machine learning model. More

specifically, powerful Graphic Processing Unit are indicated to train neural network model which are suitable model for data fusion algorithm¹.

Finally, all the collaborators will interact remotely with the various components through their individual laptop. The laptop acts in this context as a terminal to access the local infrastructure and bridge the gap with the cloud infrastructure services until fully automated pipelines have been implemented.

3.2 Software resources

As for the hardware, software represents a resource supporting the workflow. They range from the sampling step with on-filed utilities to the analysis step with statistical software. Our efforts will be directed to integrate open-source software or scripts wherever possible, making sure they follow the FAIR principles². The proprietary software that is listed have been considered to fasten the treatment with respect to open-source alternatives as a first step or if no mature open-source alternative is currently available.

3.2.1 Field-work tools

Improvements in field sampling procedures have been introduced to reduce potential sources of uncertainty and errors. During the field sampling campaigns, some repetitive manual processes, shown to be sources of errors, have been replaced by introducing digital processes. For example, the generation of unique QR codes³ for each plot, allowing rapid and accurate identification through a simple digital scanning procedure, and subsequent storage of the corresponding values sensed for each plot. Some lean phenotyping instruments, such as porometer provide such features integrated to the instrument, whereas general purpose rugged tablet may offer also a quick scan function following the sampling campaign.

3.2.2 Post processing

After sampling, the datasets are usually not ready to use for analysis. This depends on the experimental protocol or setup but also depends on the sensors themselves that may be designed for multiple purposes beyond the plant phenotyping or prototypes that the research is working out.

The required post-process to make these datasets ready to use may rely on dedicated software that perform the missing step between the raw data and the ready-to-analyze. In instruments such as Metabolomics analyzer and Overhead crane, a dedicated computer (as part of the instrument itself) contains the necessary post-processing software whereas the UAV and Rover require a photogrammetry program to combine all the images sensed to ease the procedure.

For the metabolomic part, the proprietary Agilent software has been provided with the instrument itself (Agilent Mass Hunter suite) and for photogrammetry the open-source software Spectron for the hyperspectral data has been installed, the proprietary software Riegel RiUNITE for the LIDAR data associated to the sensor and the proprietary AgiSoft Metashape software for other types of images. On one light-load workstation a Metashape license server has been installed in order to allow up to 3 concurrent usages of the software. On a long-term basis, open-source software that is

¹Meng, T., Jing, X., Yan, Z. (2020). A survey on machine learning for data fusion. *Information Fusion*, 57, 115-129. <https://doi.org/10.1016/j.inffus.2019.12.001>

²Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 1-9. <http://doi.org/10.1038/sdata.2016.18>

³The code is available at <https://gitlab.com/adrianoconte/barcoder>

commonly shared by the community such as QGIS will be preferred to document and share the processes performed.

3.2.3 *Statistical Analysis*

Once the datasets are ready to analyze, the usual next step in plant phenotyping or agronomic studies is to perform statistics to identify trends or validate the experimental hypothesis. Inside the infrastructure several perpetual statistics licenses have been purchased from user-friendly system such as SigmaPlot or SigmaStat to more expert users such as MATLAB or even specific to certain type of analysis as for Metabolomics such Agilent MassProfilerProfessional. On one lightload workstation a SigmaPlot and a SigmaStat license server have been installed in order to allow up to 10 and 5 concurrent usages of the respective software.

As mentioned previously, open source statistical and scripting tools that are commonly shared by the community such as R or Python will be preferred on a long-term basis to document and share the processes performed. On top of that scripting tools such as R or Python will be already predisposed to communicate with the data repository services presented in the deliverable 6.12.

3.2.4 *Remote Access Control*

Most of the local infrastructures are accessible remotely, allowing concurrent access and integrity by keeping the processing computer and storage together for efficient network communication. The only resource that will be accessible only from local network is the NAS for safety reasons.

To guarantee a safe remote access to desktop computers, an exclusive Remote Desktop Protocol access has been configured through to an on-site gateway using the Apache Guacamole (<https://guacamole.apache.org/>) given the limited number of users.

4 CLOUD INFRASTRUCTURE

The cloud platform combines virtual research environments for online computation and data portal for storage in an integrated solution for plant phenotyping. This computational resource aims to provide an online platform to develop and implement mainly the last steps of the pipelines, i.e. analysis and sharing of the pipelines. In this section, the services provided, and their usage are detailed within the computational architecture whereas the specific implementation for data portal is described with more detail in Deliverable 6.12.

4.1 *Virtual Research Environment*

The virtual research environment provides an online collaborative platform dedicated to HTPP with a state-of-the-art data science integrated development environment (Jupyter, Rstudio) to punctually perform statistical analysis or machine learning training on demand. This online platform opens to provide the newly developed relevant pipelines as standalone services and readily reuse them for the community or to support processes in the workflow.

Indeed, among the most urgent needs identified by the community are improvements in data analysis, the expansion of modelling capabilities, and the seamless integration of novel technologies into user workflows that will be facilitated by such a collaborative virtual online environment.

4.2 Data Portal

The data portal will be a deployment of OpenSILEX⁴, an open-source information system developed by the plant phenotyping community to store and expose phenotyping data. The adherence to FAIR principles ensures that datasets are properly documented, searchable, and available for future research, contributing to the collective efforts of advancing the plant sciences and addressing global agricultural challenges. OpenSILEX is an ontology driven system, whereby metadata is generated using community driven reference vocabularies. The stored metadata complies with the MIAPPE protocol⁵, shared by the plant phenotyping community and the system provides the open standard programming interfaces BrAPI⁶ which facilitates automated interoperability with other existing systems.

5 CONCLUSION

A high throughput plant phenotyping platform requires cutting-edge sensors that can produce large amounts of data but also requires that the entire infrastructure and associated pipelines avoid bottlenecks. The main objective of Activity 6.20 is to provide the infrastructure to develop the pipelines necessary to support the high throughput plant phenotyping data for the DEMETRIS Hub over the years. The overall architecture has been designed to foster data harmonization and reuse. We describe a two-fold infrastructure an on-premises prototyping infrastructure and an on-line community-oriented cloud infrastructure. The approach detailed could be applied therefore to other contexts or enhanced further.

⁴ Neveu P, Tureau A, Hilgert N, et al. (2019) Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. *New Phytologist* 221: <https://doi.org/10.1111/nph.15385>.

⁵ Papoutsoglou EA, Faria D, Arend D, et al. (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. *New Phytologist* 227: <https://doi.org/10.1111/nph.16544>.

⁶ Selby P, Abbeloos R, Backlund JE, et al. (2019) BrAPI - An application programming interface for plant breeding applications. *Bioinformatics* 35: <https://doi.org/10.1093/bioinformatics/btz190>.