European Plant Phenotyping Network EPPN 2020

Standardized Phenotyping storage solutions ranging from small to very large systems

Vincent Nègre, Patrick Moreau, Llorenç Cabrera-Bosquet and Pascal Neveu

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731013. 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.

Document information

EU Project N°	731013	**Acronym**	EPPN^{2020}
Full Title	European Plant Phenotyping Network 2020		
Project website	www.eppn2020.plant-phenotyping.eu		

Deliverable	**N°**	D3.2	**Title**	Standardized Phenotyping storage solutions ranging from small to very large systems
Work Package	**N°**	JRA3	**Title**	Building a consistent Information System in the different nodes and defining standardisation strategies

Date of delivery	**Contractual**	30/04/2019 (Month 24)	**Actual**	18/11/2019 (Month 29)
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)	INRA			
Responsible Author	**Name**	Vincent.negre@inra.fr	**Email**	Vincent.negre@inra.fr

Version log			
Issue Date	**Revision N°**	**Author**	**Change**
03/07/2019	1	WP leader	review by WP leader
13/11/2019	2	WP Leader	Second review
18/11/2019	3	Coordinator	Final review

Executive Summary

The objective of this work was to provide a data storage solution that fits the complex and large datasets produced by EPPN^{2020} partners. We propose a reliable and scalable solution based on a distributed e-infrastructure supported by the European Grid Infrastructure (EGI), partner of EOSC. This solution takes into account the scientific requirement of installations located in different countries. By using EGI, partners can take advantage of a set of sustainable IT services. A system based on a European e-infrastructure also makes easier data interoperability.

Objectives: Assessment of distributed storage for phenomics.

Rationale: Storage solutions are central for data reuse, they need to be flexible enough, efficient, independent but compatible in relation to infrastructures in other countries/continents and with other ESFRI infrastructures.

Main Results: A state of the art was made. The functionalities of OneData and iRODS systems was studied. A case study based on French nodes was implemented and assessed.

Authors/Teams involved: *Vincent Nègre (INRA), Patrick Moreau (INRA), Llorenç Cabrera-Bosquet (INRA), François Tardieu (INRA), Jérome Pansanel (France Grille), Baptise Grenier (EGI), Enol Fernandez (EGI), Björn Usadel (FZ-Juelich), Pascal Neveu (INRA)*

Table of contents

1. INTRODUCTION

In recent years, plant phenomics produced massive datasets involving the outputs of hundreds of sensors for tens of variables that characterize plants, soil and air (Furbank and Tester, 2011) in experiments performed in the field and in controlled conditions. Taken together, these datasets are unprecedented resources for identifying and testing novel mechanisms and models (Tardieu et al., 2017). However, making them available to a range of users, allowing re-analyses and combination with other datasets, requires reliable and flexible infrastructures able to store and organize such massive multi-source and multi-scale datasets.

This document presents a data storage solution for the datasets produced by EPPN^{2020} partners. The solution we propose is based on a distributed e-infrastructure supported by the European Grid Infrastructure (EGI), which takes into account the scientific context and the availability of computer system resources in local infrastructures. EGI, partner of EOSC, provides sustainable IT services that are close to the users and take into account contextual elements of each country (technical, funding, cultural, etc.). A system based on a European e-infrastructure makes easier data exchanges among the phenomic community, it allows sharing and reusing software environments and tools, and allows compatibility with other infrastructures, e.g. in genomics and modelling.

1.1. Distribution of data in local infrastructures

Storing and organizing datasets produced by phenomic installations is challenging in view of (i) the geographical distribution of local infrastructures across Europe (Fig. 1), (ii) the specific characteristics of installations dedicated to particular species or scientific topics and (iii) the evolving nature of phenotyping platforms. Experience has shown that a centralized data management would be ineffective in handling such multi-source and multi-scale datasets with different and heterogeneous sources.

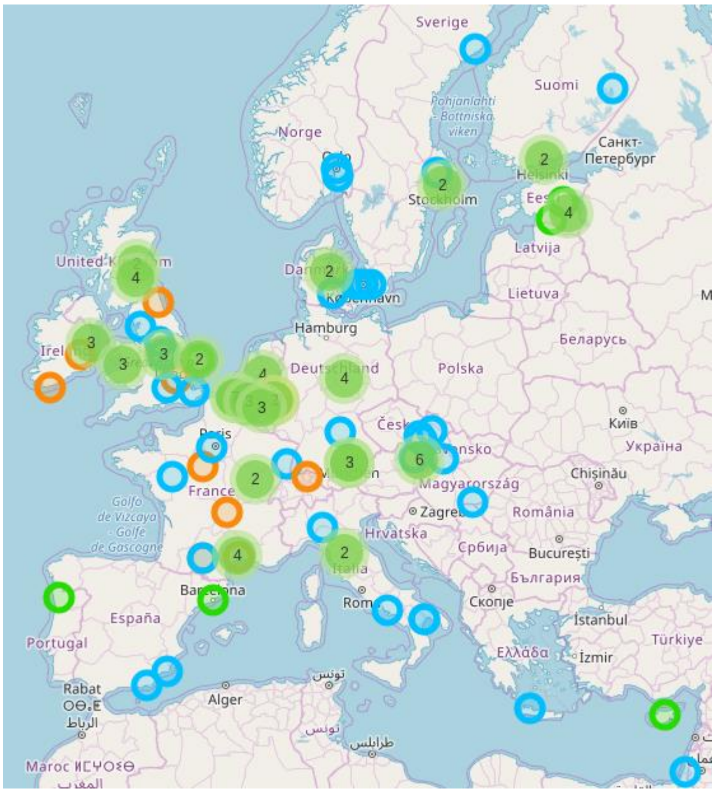

***Figure 1:** Local infrastructures in Europe (partners of the EMPHASIS future infrastructure)*

1.2. Data Volume

The data volume is very large. It depends on the type of installation, of experiment and on the number and type of sensors. For instance, in 2018 the cumulative data volume for EPPN^{2020} partners (35 installations) exceeded 1Pbyte. Elements to analyze this volume are exemplified in the French infrastructure that produced over 100 Tbytes in 2018 (Box 1).

20 experiments in field/greenhouse per year
1 experiment generates from **2** to **10 Tbytes**
(several tens of thousands of images per day)
Total data volume is over **100 Tbytes / year**

Box 1. Data volume produced in the French infrastructure (Phenome-Emphasis)

Because of the high volumes, the dispersion of phenotyping installations and the heterogeneity of datasets, the storage system needs to be distributed and scalable. It can be based on a 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 centres. 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. Selecting either OneData and iRODS selection may depend on national nodes of the European Grid Infrastructure (EGI). Local storage in servers is another option, which cannot be recommended as a flexible and scalable solution.

The international iRODS consortium supports ongoing development and evolution of the iRODS distributed open source data management software, 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 reliable backups. It also improves the speed of data transfers and its 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 better integration of new hardware.

- Workflow automation: the iRODS technology automates data administration tasks such as replication, backup or archiving of data. Furthermore, it is based on specific sequencing rules and mechanisms. It allows creating powerful and customized workflows that save time and avoid human errors.

- Data discovery: A metadata catalogue contains standardized information about the data. Metadata can include user-defined metadata in addition to traditional management system metadata, such as filename, file size, and creation date. Custom descriptors can also be applied to platform data. Implementing metadata-based functionalities is extremely useful to discover and locate data or to retrieve description on the datasets produced by software agents.

Solutions such as OneData or iRODS provide a logical representation of stored information allowing users to access this information without worrying about their physical location. Different types of resources can be integrated into these solutions such as a traditional file system, a cloud storage system or data sets dynamically provided via web services. This allows the volume of data to be increased dynamically and considerably. In order to facilitate data management, a metadata catalogue is contained in a dedicated database. Finally, EGI develops and provides a transparent gateway allowing EPPN^{2020} teams to access any of the two OneData or iRODS systems.

Key user features	**Key technical features**
1. Data virtualization provides a unified namespace for digital objects. 2. Data management tasks such as data replication, backup, archiving, or control quality can be defined according to the user needs. 3. A metadata catalogue manages user-defined metadata and track data provenance. 4. Data access is secured through authentication and authorization operation.	1. The technology supports multiple of data storage resources including databases, storage systems or tape archive systems. 2. The technology is extensible via plugins and micro-services. 3. The technology is adaptable for a variety of users and applications. 4. The technology is scalable up to petabyte-sized data sets. 5. The technology is open source, with ongoing sustainability ensured via the EGI consortium.

Box 2. Key user and technical features of iRODS storage distributed system.

1.3. Case study: The France Grilles infrastructure (French EGI node)

France Grilles is a partner of EGI (http://www.france-grilles.fr/home/). It offers a catalogue of complementary and interoperable services. These services involve Big Data analytics, computing, storage and training areas. Most of these services are compatible and/or operated on the EGI infrastructures. The storage service is based on the iRODS technology.

Regarding the amount of data produced within the EPPN^{2020} project, partners need to integrate (and cover the cost) of specific servers in the grid infrastructure. For instance, the French infrastructure has integrated physical resources to the France Grilles infrastructure, consisting in existing servers (3 Dell PowerEgde R730XD servers with 6 direct-attached storage enclosures MD1400) to increase the storage capacity to 900Tb (450Tb in replication mode). These servers were integrated to the node of the e-infrastructure in early 2017 and can be managed by PHENOME-EMPHASIS members (Figure 2).

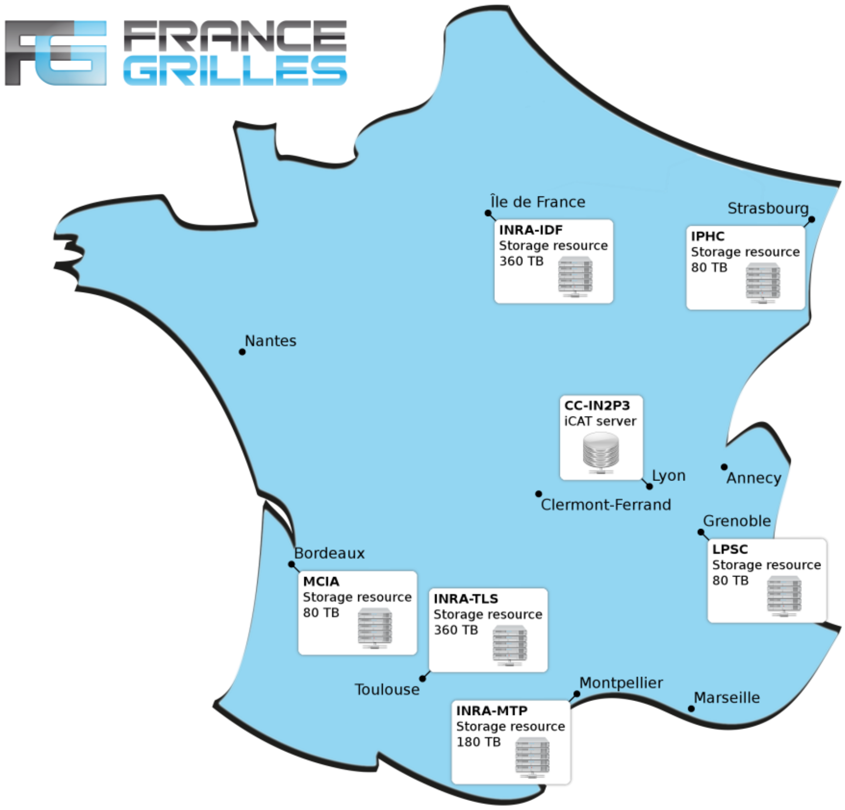

Figure 2. FranceGrilles iRODS platforms.

We tested different procedures. For large files (>32Mb), a high speed parallel connection between the client and the server was an appropriate, provided that a storage resource was set allowing to speed up the data transfers. Data transfer speeds measured (currently network nodes are 1Gb/s or 10Gb/s) on several infrastructures are indicated in the **Table 2.**

iRODS platform	**Data transfer speed**
INRA-MTP	10-12 Mb/sec
INRA-TLS	60-80 Mb/sec

Table 2. Data transfer on FranceGrilles iRODS infrastructure.

Some figures regarding the use of the iRODS infrastructure by the PHENOME-EMPHASIS project are indicated below in Box 3.

Number of files stored on the FG-iRODS infrastructure : 38 million
Number of user account : 15
Data volume: 100 To
Phenotyping facilities using the FG-iRODS infrastructure: 4
Transversal projects using the FG-iRODS infrastructure: 2

Box 3. Usage of the FranceGrilles iRODS infrastructure by the PHENOME-EMPHASIS users.

2. TOWARDS A EUROPEAN E-INFRASTRUCTURE FOR PLANT PHENOTYPING

2.1. The EGI foundation

The EGI foundation (https://www.egi.eu/about/egi-foundation/) coordinates an e-infrastructure set up to provide advanced computing services for European research and innovation. The EGI e-infrastructure is publicly-funded and comprises hundreds of data centres and cloud providers spread across Europe and worldwide.

2.2. The EGI services

EGI delivers advanced computing services to support scientists, international projects and research infrastructures including computation, storage, security and training services.

- ✓ Computing
  - **Cloud computation** service: allows to run virtual machines on demand with complete control over computing resources. It provides flexibility to partner's needs. Indeed, each partner can compose suites of software (image analysis pipelines, data cleaning, calibration, etc.). Finally, groups can easily exchange and reproduce analyses.
  - **Cloud Container Compute** (beta service): Run Docker containers in a lightweight virtualized environment provide a way to deploy software tools for the project users. Partners can take advantage of compatible version software sets. This also makes it easier to share tools and results.
  - **High-Throughput Compute**: allows performing thousands of computational tasks to analyse large datasets. For instance, it provides scalable and flexible computing resources for plant trait extractions.

- ✓ Storage services:
  - **Online Storage:** store, share and access files and associated metadata on a global scale. EGI supports the iRODS and OneData technologies. In collaboration with EGI we can provide a gateway between both systems (a prototype is available). This is an important result allowing each of the EPPN^{2020} partners (depending on the country) to choose either OneData or iRODS.
  - **Data Transfer:** An efficient transfer of large sets of data from one place to another is provided, based on multi-sources.
  - **Archive Storage:** Back-up institution data for the long term and future in a secure and low energy environment.

- ✓ Security services:
  - **Check-in (beta service):** login with credentials.

- ✓ Applications services:
  - **Applications on Demand (beta service):** use online applications for data and compute intensive research
  - Notebooks (beta service): create interactive documents with live code, visualisations and text.

- ✓ Training services:
  - **ISO 27001 Training:** managing and securing datasets
  - **Training Infrastructure**: dedicated computing and storage for training and education.
  - **FitSM Training**: managing IT services with a pragmatic and lightweight standard

2.3. The 'Design your e-Infrastructure' workshop

The EGI foundation organized the workshop "*Design your e-Infrastructure*" on Thursday, 9 May 2019 in Amsterdam. The workshop focused on selected community use cases and co-designed e-infrastructure setups for them. Three pilots were designed to support the e-infrastructure for plant phenotyping, which rely on EGI services for storage, computing and authentication. In the first pilot (France Grilles-based, Fig. 3A), the *e*xisting information system PHIS *for plant phenomics information system* (Neveu et al., 2019) was deployed on EGI. Existing Galaxy environment was also deployed on an EGI virtual machine. Large datasets (images) were stored on the existing France Grilles iRODS infrastructure previously presented. The pilot contained an authentication layer based on the EGI check-in service and a computing layer provided with the EGI Notebooks service (Figure 3*A*). Early adopters will be responsible for deploying, testing and evaluating these pilots with the support of the EGI team. In a first step, two main partners were identified: INRA Montpellier and Wageningen University and Research. Most appropriate solutions will then be proposed to the other EPPN^{2020} project partners.

CONCLUSION

EPPN2020 local infrastructures can take advantage of the standardization and the compatibility of EGI e-services. Tests performed using iRODS and OneData distributed storage systems demonstrated their efficiency for storing large data sets and for managing data produced by different phenomic installations. The coordination of hardware and services by each of the European countries through the EGI provides an affordable and sustainable choice for the phenomics community. In addition, the collaboration with national and European infrastructure teams allows taking advantage of local skills, which is crucial for deploying and maintaining a robust and durable e-infrastructure.

REFERENCES

Furbank RT, Tester M. 2011. Phenomics - technologies to relieve the phenotyping bottleneck. Trends in Plant Science 16, 635-644.

Neveu P, Tireau A, Hilgert N, Nègre V, Mineau-Cesari J, Brichet N, Chapuis R, Sanchez I, Pommier C, Charnomordic B, Tardieu F, Cabrera-Bosquet L. 2019. Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. New Phytologist 221, 588-601.

Rajasekar A, Moore R, Hou C-y, Lee CA, Marciano R, de Torcy A, Wan M, Schroeder W, Chen S-Y, Gilbert L. 2010. iRODS Primer: integrated rule-oriented data system. Synthesis Lectures on Information Concepts, Retrieval, and Services 2, 1-143.

Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M. 2017. Plant Phenomics, From Sensors to Knowledge. Current Biology 27, R770-R783.

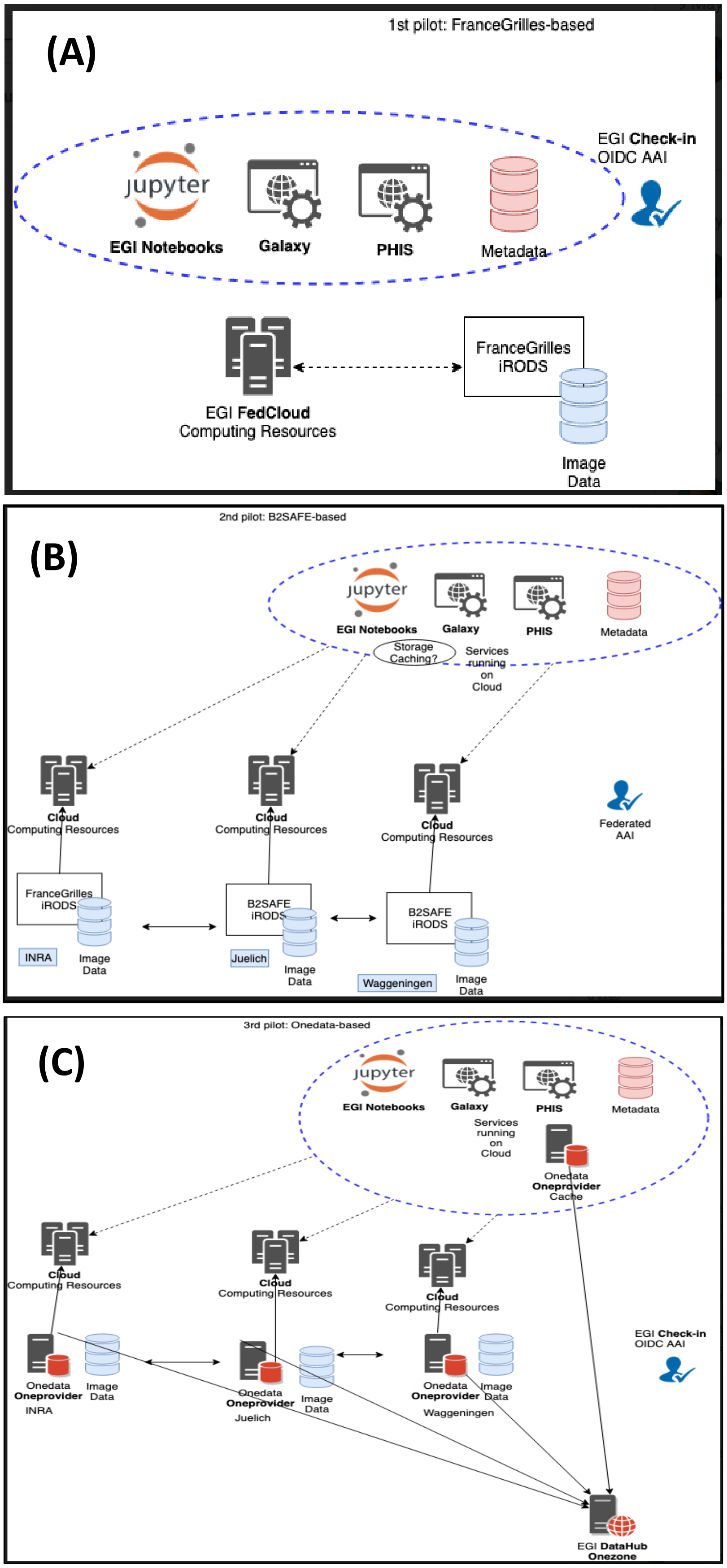

Figure 3: EGI pilots