Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

# InfraPhenoGrid: A scientific workflow infrastructure for plant phenomics on the Grid

Christophe Pradal [a,b], Simon Artzet [c,b], Jérôme Chopard [d,b], Dimitri Dupuis [e], Christian Fournier [c,b], Michael Mielewczik [c,f], Vincent Nègre [c], Pascal Neveu [d], Didier Parigot [e], Patrick Valduriez [e], Sarah Cohen-Boulakia [b,e,g,*]

[a] CIRAD, UMR AGAP, Montpellier, France
[b] Inria, VirtualPlants, Montpellier, France
[c] INRA, UMR459, LEPSE, F-34060 Montpellier, France
[d] INRA, UMR729, MISTEA, F-34060 Montpellier, France
[e] Inria, Zenith, Montpellier, France
[f] ICCH, NHLI, Imperial College London, UK
[g] Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623, Université Paris-Saclay, Orsay, France

## HIGHLIGHTS

- An infrastructure to manage huge datasets produced by plant phenomics platforms.
- Modular and highly expressive scientific workflows are designed to analyze datasets.
- Scientific workflows are distributed over the Grid using an extensible middleware.
- Provenance is managed to allow users understand results and ensure reproducibility.

## ARTICLE INFO

*Article history:*
Received 31 July 2015
Received in revised form
1 June 2016
Accepted 2 June 2016
Available online xxxx

*Keywords:*
Phenomics
Scientific workflows
Provenance
Grid computing

## ABSTRACT

Plant phenotyping consists in the observation of physical and biochemical traits of plant genotypes in response to environmental conditions. Challenges, in particular in context of climate change and food security, are numerous. High-throughput platforms have been introduced to observe the dynamic growth of a large number of plants in different environmental conditions. Instead of considering a few genotypes at a time (as it is the case when phenomic traits are measured manually), such platforms make it possible to use completely new kinds of approaches. However, the datasets produced by such widely instrumented platforms are huge, constantly augmenting and produced by increasingly complex experiments, reaching a point where distributed computation is mandatory to extract knowledge from data.

In this paper, we introduce InfraPhenoGrid, the infrastructure we designed and deploy to efficiently manage datasets produced by the PhenoArch plant phenomics platform in the context of the French Phenome Project. Our solution consists in deploying scientific workflows on a Grid using a middleware to pilot workflow executions. Our approach is user-friendly in the sense that despite the intrinsic complexity of the infrastructure, running scientific workflows and understanding results obtained (using provenance information) is kept as simple as possible for end-users.



## 1. Introduction

Biological research derives its findings from the proper analysis of experiments. However, over the last three decades, both throughput of experiments (from single observations to terabytes of sequences of images produced during a single day) and the breadth of questions studied (from single molecules to entire genomes) have increased tremendously. One of the main challenges remains to efficiently analyze, simulate and model such big datasets while keeping scientist users in the loop.

In this paper we introduce InfraPhenoGrid, the infrastructure we designed and deployed to efficiently manage and analyze

* Corresponding author at: Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623, Université Paris-Saclay, Orsay, France.
*E-mail address:* cohen@lri.fr (S. Cohen-Boulakia).

http://dx.doi.org/10.1016/j.future.2016.06.002
0167-739X/

datasets produced by the PhenoArch plant phenomics platform. In this context, one difficulty remains to enable users to analyze, simulate and model increasingly huge datasets on a more frequent base. More precisely, the design of InfraPhenoGrid is driven by three needs, described here-after.

First, management of large-scale experiments involving possibly large numbers of interlinked tools has to be supported. Users should be able to analyze and simulate complex structural-functional relationships of plant architectures, integrating multi-disciplinary models developed by different teams. Experiments should be easy to design by users and it is important that over time they can be changed, adapted to new needs (new analysis algorithms are constantly available), and then shared. As a result, the first brick of our infrastructure is a *Scientific Workflow System*.

Second, each experiment can be replayed several times, varying datasets and/or parameter settings. Keeping track of the exact datasets and parameter settings used to produce a given result (provenance) is of paramount importance for scientists to ensure the results' reproducibility and allow to properly interpret and understand them. The possibility of comparing results, obtained on several experiments when varying datasets and/or parameter settings are used, is another need directly associated with provenance. Consequently, the second brick of our infrastructure is a *Provenance Layer*.

Last but not least, our infrastructure has to efficiently deal with the analysis of huge datasets, possibly acquired on multiple sites. Analysis may involve combining data produced by platforms with completely different kinds of data, including data obtained from public data sources. Data acquisition is fast compared to the time needed to analyze them. The size of datasets has reached a turning point at which local infrastructures are no longer sufficient to provide adequate computational power and storage facilities. Hence, distributed computation has become a major requirement. However, deploying jobs on a parallel environment might be complex for end users. Therefore the third brick of our infrastructure introduces a *Middleware* able to pilot the execution of jobs on parallel (Grid) environments.

This paper is organized as follows: Section 2 introduces the precise context of this work, that is, the Phenome Project, PhenoArch platform and one use case of interest. Section 3 describes in detail the architecture of InfraPhenoGrid. Section 4 demonstrates the benefit of using our solution for managing plant phenomic datasets. Section 5 provides related work while Section 6 concludes the paper and draws perspectives.

## 2. Use case

### 2.1. The Phenome project and the PhenoArch platform

Selecting genotypes that maintain and increase crop performance is a particularly challenging and important topic in the context of societal challenges such as climate change adaptation, food security and preserving natural resources.

A large variety of tasks have to be performed to collect information on plant traits (called phenotyping), including measuring the size of the leaves, counting the number of tails.… Performing such tasks manually makes it impossible to consider more than a few plants at a time and it thus cruelly confines the kind of analyses that can be conducted.

In the meantime, massive plant phenotyping in the field, that is, the evaluation of crop performance (yield) of millions of plants in a large range of environmental and climatic scenarios, has been very efficient for driving plant breeding. However plant breeding is now facing a stagnation of genetic progress in several species. New strategies, such as genomic selection, are now evolving to directly link the allelic composition of a genome, available at much higher throughput and lower cost than field phenotyping, to crop performance. The existence of large marker–environment interactions, *i.e.* the fact that a given combination of markers has very different genetic values depending on the climatic scenario, lead concomitantly to a revolution in phenotyping strategies. Such strategies aim to capture under controlled conditions, the genetic variability of plant responses to environmental factors for thousands of plants (reference panels), hence identifying more heritable traits for genomic selection.

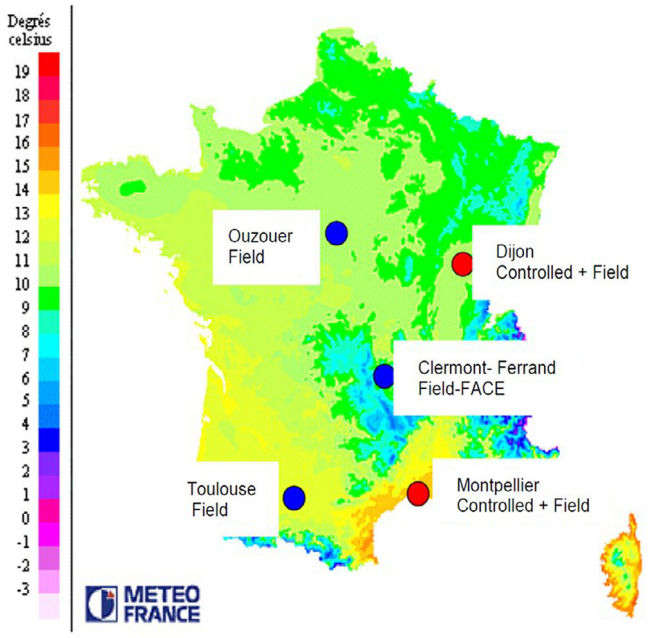


**Fig. 1.** Location of Phenome platforms, superimposed on a map of mean temperature in France. Phenome platforms are representative of the variability of temperature. They also represent different risks of water deficit.

This first implies the necessity to automate quantification of a large number of traits, to characterize plant growth, plant development and plant functioning. Second, it requires a tight control or at least accurate measurement of environmental conditions as sensed by plants. It finally requires fluent and versatile interactions between data and continuously evolving plant response models. Such interactions are essential to be considered in the analysis of a given marker–environment interaction and in the integration of processes to predict genetic values of allelic combinations in different environment scenarios.

High-throughput phenotyping platforms have thus been designed to allow growing and observing traits of a large number of plants. These platforms provide many measurements and imaging functionalities for different plant species grown in various environmental conditions. They potentially allow to assess the genetic variability of plant responses to environmental conditions using novel genetic approaches requiring a large number of genotypes.

Nine of such platforms, distributed over various regions of France, are gathered in the Phenome project (Fig. 1). More precisely, Phenome consists of two controlled condition platforms (greenhouses with automated irrigation, $CO_2$ control and temperature control) for 1900 plants, two field platforms (800 plots) equipped with environment control ($CO_2$ enrichment, automated rain shelters) and three larger field platforms (2000 plots) that use natural gradients of water availability or soil contents. All these platforms are equipped with environmental sensors and permit automated imaging of plants in one or multiple wavelengths (thus allowing functional analysis) using robots to convey plants (for green houses) or to carry instruments to automatically acquire data in the field (Phenomobile, drones). Finally two supporting omic

Please cite this article in press as: C. Pradal, et al., InfraPhenoGrid: A scientific workflow infrastructure for plant phenomics on the Grid, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.06.002