Déploiement de la plateforme de traitement des données phénotypage haut débit 4P sur l'infrastructure France Grilles

Vincent Nègre
Laboratoire d'Écophysiologie des Plantes sous Stress Environnementaux
2, place Viala
34060 Montpellier Cedex 2

Eric David
Ephesia consult
7 rue Michel Royer
45100 Orléans

Philippe Burger
UMR1248 AGIR (AGroécologie, Innovations, teRritoires)
INRA Auzeville
24 chemin de Borde-Rouge - Auzeville CS 52627
31326 CASTANET-TOLOSAN CEDEX

Romain Chapuis
UE0398 MELGUEIL DIASCOPE - Domaine Expérimental de Melgueil
INRA Domaine de Melgueil
Chemin de Mezouls
34130 MAUGUIO

Boris Adam
UE1375 PHACC Phénotypage Au Champ des Céréales
INRA Site de Crouël
5 Chemin de Beaulieu
63000 CLERMONT-FERRAND

Anne Tireau
UMR0729 MISTEA Mathématiques, Informatique et Statistique pour l'Environnement et l'Agronomie
INRA - Campus Supagro Montpellier
2 place Viala
34060 MONTPELLIER CEDEX 2

Patrick Moreau
UMR1248 AGIR (AGroécologie, Innovations, teRritoires)
INRA Auzeville
24 chemin de Borde-Rouge - Auzeville CS 52627
31326 CASTANET-TOLOSAN CEDEX

Anthony Tong
Ephesia consult
7 rue Michel Royer
45100 Orléans

Gallian Colombeau
Hiphen
228 Route de l'Aérodrome
84140 Avignon

Samuel Thomas
ARVALIS - Institut du végétal

Pascal Neveu
UMR0729 MISTEA Mathématiques, Informatique et Statistique pour l'Environnement et l'Agronomie
INRA - Campus Supagro Montpellier
2 place Viala
34060 MONTPELLIER CEDEX 2

Jérôme Pansanel
Institut Pluridisciplinaire Hubert Curien
23, rue du Loess – BP28
67037 Strasbourg Cedex 2

Frédéric Baret
UMR1114 EMMAH Environnement Méditerranéen et Modélisation des Agro-Hydrosystèmes
INRA Domaine Saint-Paul - Site Agroparc
228 route de l'Aérodrome CS40509
84914 AVIGNON CEDEX 9

Marie Weiss
UMR1114 EMMAH Environnement Méditerranéen et Modélisation des Agro-Hydrosystèmes
INRA Domaine Saint-Paul - Site Agroparc
228 route de l'Aérodrome CS40509
84914 AVIGNON CEDEX 9

Résumé

Le projet PHENOME-EMPHASIS associant l'INRA, Arvalis et Terres-Inovia ambitionne de développer des infrastructures de phénotypage haut-débit au niveau national. Les systèmes d'acquisition aux champs (drone, phenomobile) embarquent différents capteurs (caméras haute résolution RGB, multispectrales et infra-rouge thermique, LIDARs) qui génèrent un volume important d'images qu'il convient de traiter, stocker et archiver.

Les modules de traitement prototypes créés par l'UMT CAPTE ont été industrialisés et intégrés dans la plateforme de traitement des données 4P (Plant Phenotyping

Processing Platform). Ces modules encapsulés dans des conteneurs Docker peuvent être enchaînés dans des workflows s'appuyant sur le moteur de traitements Cromwell. Docker Swarm permet de distribuer l'exécution des conteneurs sur un cluster.

Les données brutes et les traitements sont stockés sur une architecture distribuée basée sur la technologie iRODS.

La plateforme 4P est connectée au système d'information PHIS ayant pour objectif de stocker et d'organiser les données produites dans le cadre du projet PHENOME-EMPHASIS selon les principes FAIR.

La plateforme 4P repose sur l'utilisation de l'infrastructure France Grilles sur laquelle sont déployés différents services aux utilisateurs. Pour le déploiement de la plateforme 4P, nous nous sommes appuyés sur le service FG-CLOUD pour la partie applicative et sur le service FG-IRODS pour la partie stockage persistant des données.

Mots-clefs

phénotypage à haut débit, gestion de données, traitement de données scientifiques, France Grilles

1 Introduction

1.1 Le phénotypage des plantes à haut débit

Le phénotypage consiste à observer les traits qui caractérisent la structure ou le fonctionnement d'un individu (les plantes dans notre cas). Le suivi de ces traits permet aux agronomes d'améliorer la sélection variétale en identifiant par exemple des variétés plus adaptées à un environnement donné (climat, sol, pathogènes, ...). Le phénotypage permet également aux chercheurs de mieux comprendre les mécanismes de réponses des plantes à leur environnement, d'améliorer les modèles de fonctionnement des plantes, et par là même d'améliorer les outils d'aide à la décision pour le diagnostic et les prévisions de performance des cultures.

Afin d'établir des corrélations robustes, il est nécessaire de considérer du phénotypage haut-débit, c'est-à-dire de suivre un grand nombre d'individus (plusieurs centaines voir plusieurs milliers), parfois dans différentes modalités (introduction de pathogènes ou non, différentes conditions d'irrigation ou d'apport azoté) et de mesurer un grand nombre de traits (jusqu'à plusieurs dizaines). Ainsi pour chaque plateforme d'expérimentation de phénotypage haut-débit, on considère plusieurs milliers de micro-parcelles qui doivent être caractérisées et ce, tout au long du cycle de développement de la culture (voir **Figure 1**).

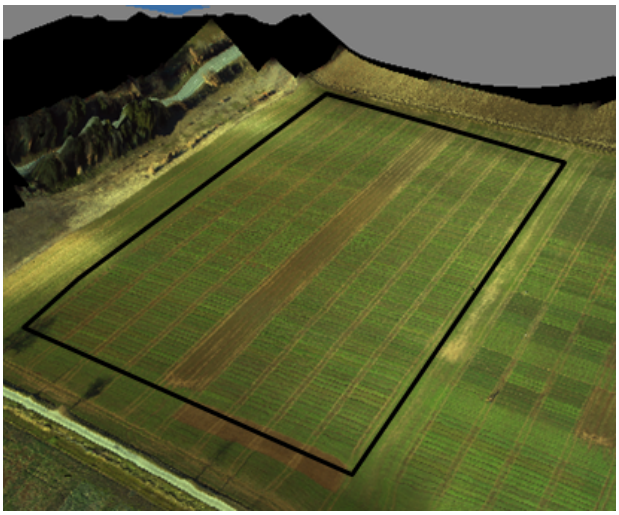


Figure 1 - Vue aérienne d'une expérimentation de phénotypage haut-débit.

1.2 Le projet PHENOME-EMPHASIS

Au cours de ces dernières années de nombreuses plateformes de phénotypage des plantes ont été développées. Le projet Phenome-Emphasis est une Infrastructure Nationale en Biologie Santé (INBS) composé de 11 sites[1] fortement instrumentés dont 6 sites en conditions contrôlées ou semi-contrôlées (plantes en pots sous serre ou abris); 3 plateformes omiques et 2 plateformes en condition au champ (voir **Figure 2**).

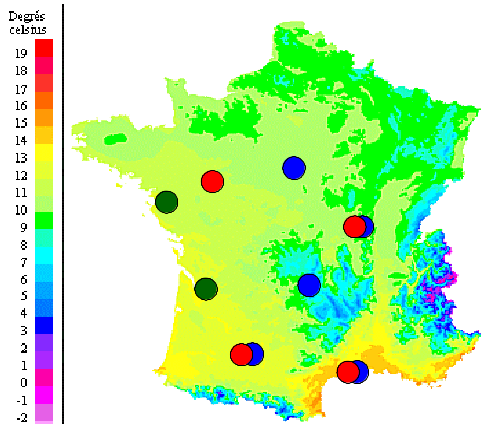


Figure 2 - Réseaux de plateformes Phenome-EMPHASIS
*(**bleu**: conditions contrôlées ou semi-contrôlées; **rouge**: serre; **vert**:omique)*

1. https://www.phenome-emphasis.fr/Installations

Le suivi tout au long de la croissance étant indispensable, l'acquisition d'images de toute nature [1] est très largement utilisée car elle est non destructive, automatisable et permet après analyse d'accéder à une grande variété de traits parfois complexes tels que l'architecture ou la biomasse des plantes (voir **Figure 3**).

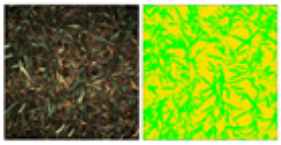
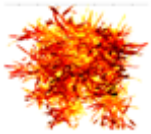
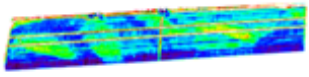
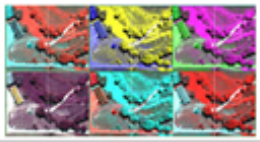
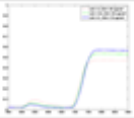
			Utilisations
	Camera haute résolution/RVB		• Identification adventices • Enherbement • Taux de couverture verte • Développement foliaire • Comptage de plante • Dégâts de nuisibles • Biomasse, risque de verse
	LIDAR Mesure de distance et profondeur		• Structure • Port du couvert • Interception lumineuse • Expérimentation
	Caméra IR thermique		• Température de surface • Stress hydrique • Pilotage de l'irrigation
	Camera multispectrale		• Développement • Statut azoté/chlorophylle • Stress • Pilotage de l'azote
	Spectromètre		• Chlorophylle/statut azoté • Stress (hydrique , ...) • Développement • Composés de la plante

Figure 3 - Types de capteurs utilisés en imagerie et utilisations.

Au champ les images sont acquises par un robot, appelé Phénomobile, conçu par l'INRA et Arvalis Institut du végétal pour le phénotypage au champ, ainsi qu'avec des drones (hexacoptères).

La Phénomobile (voir **Figure 4**) est équipée d'un bras télescopique de 12 mètres de long portant une tête de mesure orientable. Les quatre caméras (caméras visibles et caméras multispectrales) sont synchronisées avec des flashs, ce qui rend l'acquisition des images indépendante des conditions d'éclairement naturel permettant ainsi une qualité de données constante. L'appareil est également équipé de trois lasers à balayage (LIDAR).

Les images par drone sont acquises dans les spectres visibles et proche infrarouge pour des mesures de la hauteur des couverts ou le calcul d'indices comme le NDVI[2] et de variables caractérisant l'état du couvert végétal comme la surface foliaire. Des tests sont

2. Le NDVI (Normalized difference Vegetation Index) correspond au rapport entre la différence entre les canaux proche infrarouge et rouge sur leur somme.

en cours pour utiliser également des caméras infrarouge thermiques permettant d'accéder à la température des cultures, en lien avec les processus liés à l'évapotranspiration pour permettre d'évaluer les besoins en eau des cultures.

Figure 4 - Le robot de phénotypage au champ Phénomobile conçu par l'INRA et Arvalis-institut du végétal.

2 Problématiques

Cette production massive de données (en particulier grâce à l'imagerie) nécessite le développement des chaînes de traitements adaptées ([2] [3] [4] [5] [6] [7]). Ces chaînes de traitement sont développées par différents partenaires et s'appuient sur des langages variées (parfois des logiciels propriétaires). Le besoin d'une plateforme permettant d'assurer leur ré-utilisation, leur partage et leur traçabilité s'est donc rapidement posé.

Cette plateforme a dû répondre aux critères suivants :

- **Généricité** : la plateforme doit pouvoir intégrer des chaînes de traitement développées dans des langages variés (C++, MATLAB, Python), certaines utilisent un logiciel commercial sous licence (Agisoft Photoscan[3]) ;
- **Évolutivité**: la plateforme doit fournir la possibilité d'ajouter de nouveaux traitements ou type de données;
- **Traçabilité:** la plateforme doit offrir un système gestion et de versionnage des chaînes de traitement ;

3. https://www.agisoft.com/

- **Provenance**: elle doit également être compatible avec le système d'information scientifique PHIS [8] qui permet de stocker et gérer des données issues des expérimentations de phénotypage ;
- **Ouverture**: la plateforme doit être compatible avec les différents types de capteurs utilisés sur les installations expérimentales; mais aussi avec avec les modules développés par les partenaires (en particulier ceux développés par Hiphen[4]);
- **Facilité d'utilisation** : elle doit être utilisable par un public non informaticien.

3 La plateforme de traitement 4P

3.1 Les partenaires

La société EPHESIA consult[5] a conduit le développement logiciel sous la maîtrise d'ouvrage de l'INRA.

La liste des partenaires et leurs rôles respectifs sont les suivants :

- EPHESIA consult : maîtrise d'œuvre (développement logiciel) ;
- INRA : maîtrise d'ouvrage (spécifications, tests, suivi du projet) et développement des modules de traitements des données ;
- Hi-Phen ; ARVALIS : développement des modules de traitements des données ;
- France Grilles : mise à disposition des ressources de stockage et de calcul.

4 Architecture et outils utilisés

L'architecture de la plateforme 4P est décrite sur la **Figure 4**.

4. https://www.hiphen-plant.com/

5. https://www.ephesia-consult.com/

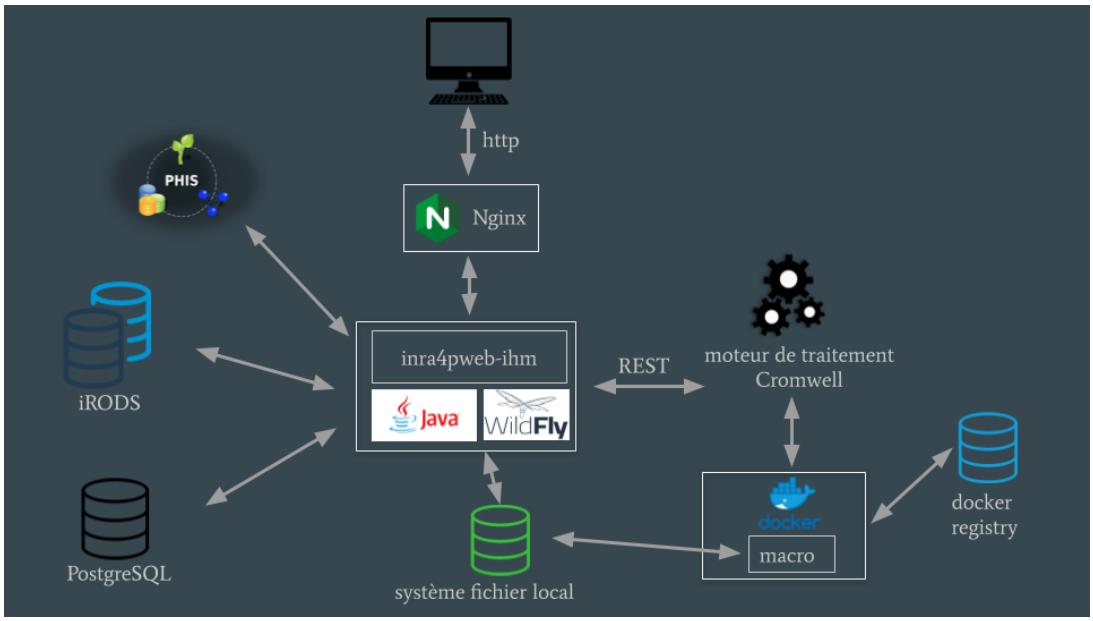


Figure 5 - Architecture logicielle de la plateforme 4P.

4.1 L'application Java

L'application Java est l'application métier qui permet de gérer la plateforme. Elle se décompose en 8 modules :

- le module *config* contient les configurations pour chaque environnement cible (production, qualification, ...) ;
- le module *common* regroupe les fonctionnalités transverses (classe utilitaires, constantes, etc.) ;
- le module *data* permet d'effectuer les opérations de sélection, sauvegarde et suppression en bases de données. Celui-ci s'appuie sur la librarie Hibernate qui implémente les spécifications JPA[6]. Le module est indépendant du choix du type de base de données relationnel. Il est donc possible de connecter n'importe quelle base de données relationnelle à l'application ;
- les modules *client* implémentent les appels aux web services (REST) de Cromwell (moteur de traitement) et PHIS (système d'information scientifique). Ceux-ci sont générés à partir de l'API Swagger[7], API aussi utilisée dans l'exposition des services Cromwell et PHIS ;
- le module service regroupe l'ensemble des services fonctionnels. C'est lui qui contient la logique métier de l'application. Il a donc accès aux autres modules ("data" et "client" en particulier) ;

6. Java Persistence API

7. https://swagger.io/

- le module *ihm* expose les services métiers à travers des interfaces graphiques en HTML5, CSS3 et Javascript. Ce module respecte le motif MVC grâce à l'utilisation du framework JSF. Par ailleurs l'utilisation de la librairie CSS et Javascript Bootstrap permet d'améliorer l'expérience utilisateur en rendant l'application compatible avec tous les formats d'écran.

4.2 Le moteur de traitement Cromwell

Cromwell[8] est le moteur de traitement en charge de la distribution des traitements (voir **Figure 6**). Nous l'avons retenu car il est simple d'utilisation (via la ligne de commande), « scalable » (il dispose de fonctions permettant de distribuer les traitements) et compatible avec le standard WDL[9]. Ce standard permet de décrire l'enchaînement d'un traitement de façon interprétable par la machine et compréhensible par un utilisateur.

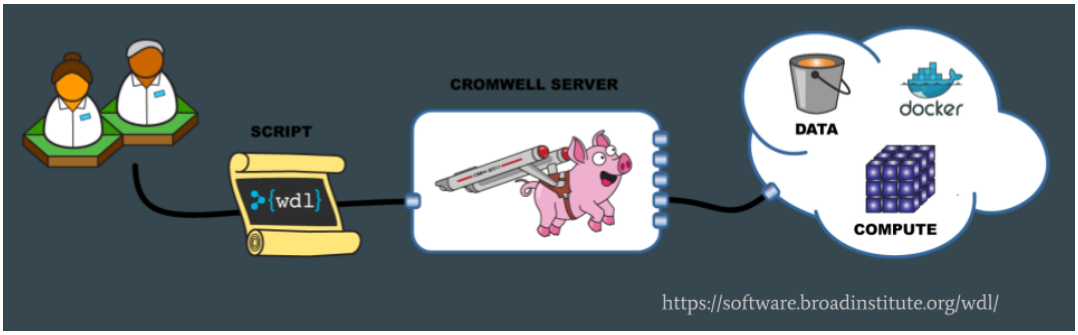


Figure 6 - Fonctionnement général du workflow CromWell.

Une chaîne de traitement également appelée *macro* est donc constituée d'un fichier WDL qui va décrire les enchaînements des différentes étapes de calculs (ou *modules*) ainsi que leurs entrées et les sorties.

Afin d'être portables les modules sont encapsulés dans des images Docker[10]. Les modules sont disponibles dans un dépôt d'images privées (Docker registry) partagé entre les partenaires.

L'utilisateur dispose d'une interface graphique qui permet de créer sa macro en enchaînant les différents modules nécessaires aux traitements de ses données (voir un exemple sur la **Figure 7**). On peut voir dans cette figure que l'exécution de certaines macros est parallélisée (icône).

8. https://cromwell.readthedocs.io/en/stable/
9. Workflow Description Language
10. https://www.docker.com/

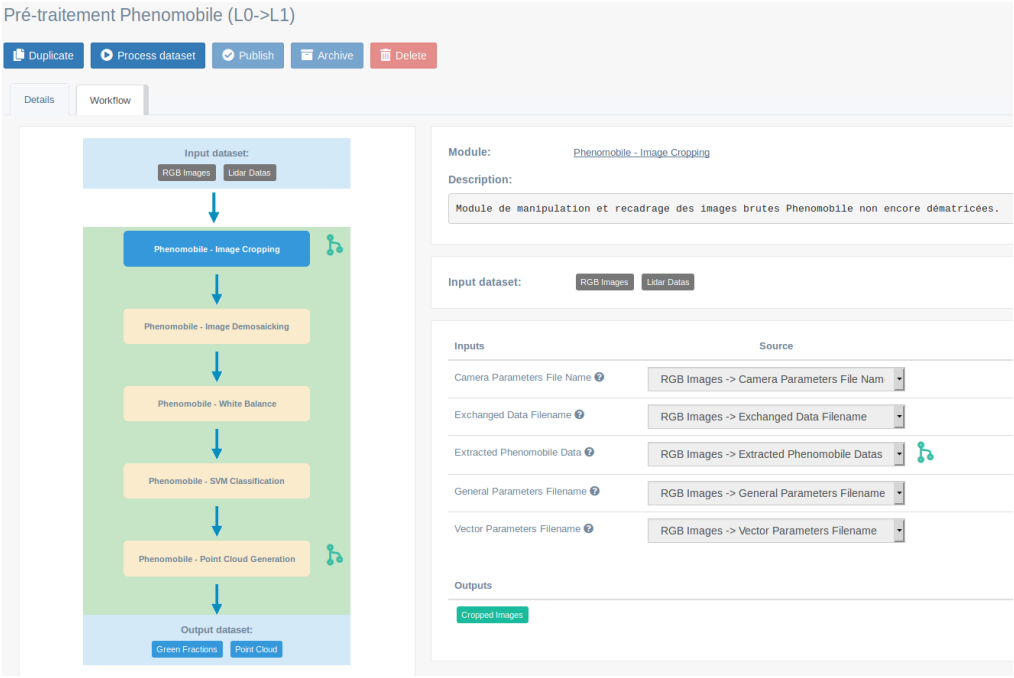


Figure 7 - Interface de création d'une macro dans 4P.

Lorsque l'utilisateur crée une macro sur l'interface de 4P, l'application génère automatiquement un fichier WDL décrivant les modules et leur enchaînement (voir **Figure 8**) ainsi qu'un fichier JSON décrivant les paramètres d'entrées et de sorties.

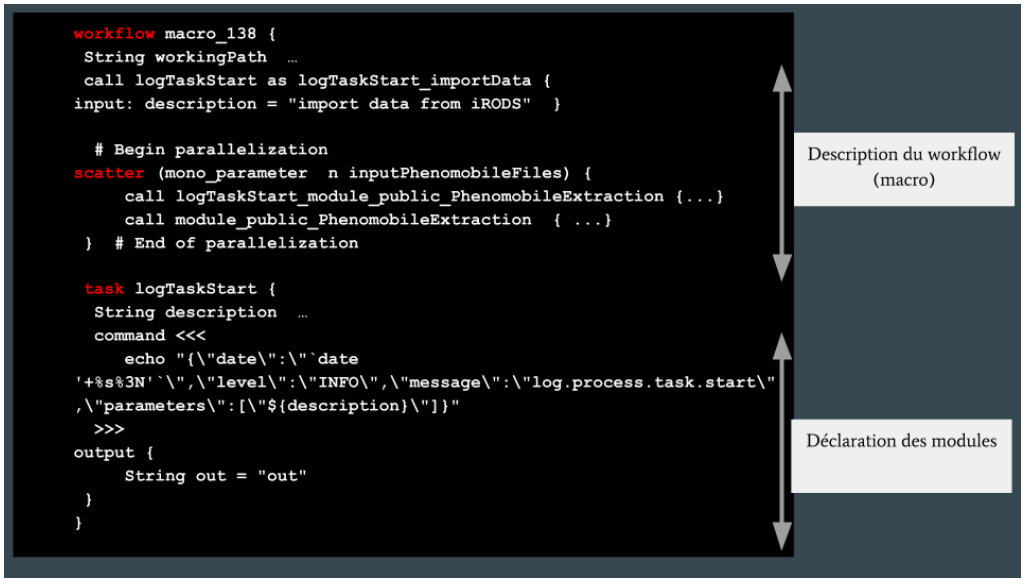


Figure 8 - Exemple de fichier WDL généré par 4P.

Le bloc **workflow** correspond à la description de l'enchaînement des différentes tâches. Chaque tâche (ou module) est décrite dans un bloc **task** qui va correspondre à l'exécution d'un container Docker (voir **Figure 10** plus bas). Le mot clé **scatter** permet de paralléliser le traitement.

Le lancement du traitement se fait très simplement (voir **Figure 9**) .

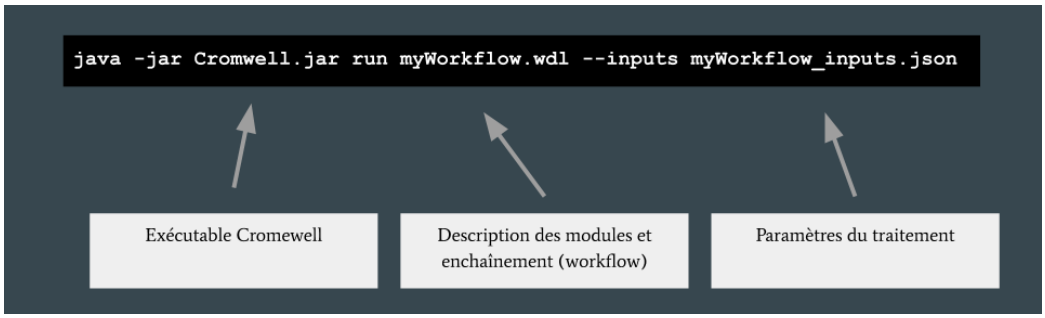


Figure 9 - Exemple d'exécution d'un workflow avec CromWell.

L'exécution proprement dite des traitements s'effectue dans un cluster Docker Swarm.

La communication entre l'application Java et le moteur de traitement s'effectue via l'API REST disponible dans Cromwell.

4.3 Les modules de calcul

Un *module* de calcul (correspondant à une étape du calcul) peut être écrit dans n'importe quel langage (Java, C, C++, Python, ...), mais doit respecter les conditions suivantes:

- il doit être embarqué dans une image Docker incluant tout l'environnement d'exécution ;
- il doit être exécutable en ligne de commande sans intervention de l'utilisateur (voir **Figure 10**) ;
- les paramètres doivent être fournis en arguments et doivent correspondre à un type de données déclaré dans 4P.

Le module doit être publié dans un dépôt git (Gitlab, Bitbucket, Github, ...) accessible en lecture par 4P.

CLI Example:

```
docker run -v <host>:<mount_point> -w <mount_point> <module_name>:<tag>
[arguments]
```

Figure 10 - Exemple d'exécution d'un module Docker

4.4 Transfert des données

Depuis une interface web, l'expérimentateur peut transférer les images acquises lors des expérimentations au champ. Les images sont transférées sur un système de stockage distribué basé sur la technologie iRODS[11]. Les fichiers allant jusqu'à plusieurs dizaines de Go sont transférés par le protocole http grâce à la W3C file API implémentée dans les dernières versions des navigateurs (IE 11+, Firefox 60+, Chrome 49+, Safari 10.03+).

Lorsque l'utilisateur lance un traitement (le transfert peut être découplé de l'upload), les images sont copiées sur un système de fichier local pour être traitées. Une fois le traitement terminé les résultats sont copiés vers iRODS et le système de fichier local est nettoyé.

Le système iRODS est maintenu par l'infrastructure France Grilles[12] à travers son service de gestion de données FG-IRODS (voir **Figure 11**) . C'est un système qui permet de stocker de grandes quantités de données (pouvant aller jusqu'à plusieurs Po) et de les sécuriser grâce à un mécanisme de réplication.

Dans notre cas les données sont automatiquement répliquées parmi 3 sites répartis géographiquement sur le territoire (le data center INRA de Toulouse; le data center du CEA en Ile de France et le site du CINES à Montpellier).

11. https://irods.org/

12. http://www.france-grilles.fr/accueil/

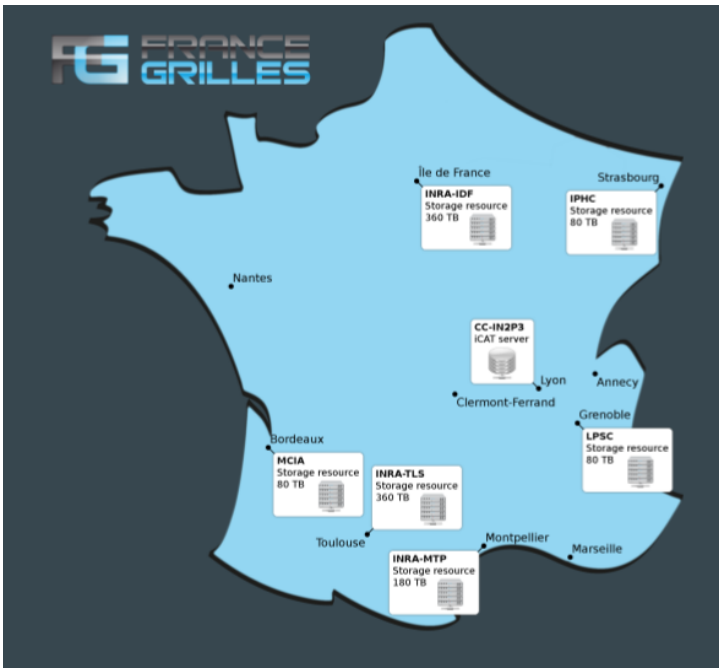


Figure 11 - Infrastructure iRODS opérée par FranceGrilles.

Les métadonnées nécessaires aux traitements (droits utilisateurs, type de capteurs, contexte d'acquisition des images) sont récupérées sur le système d'information PHIS grâce à l'appel aux web services REST.

4.5 Le déploiement sur l'infrastructure FG-CLOUD

La plateforme 4P est déployée sur l'infrastructure cloud opérée par France Grilles à travers son service FG-Cloud[13].

A travers ce service basé sur la technologie OpenStack[14], nous utilisons six machines virtuelles (CentOS 7) afin de répartir les services et la charge :

- une machine virtuelle hébergeant l'application web ;
- une machine virtuelle qui gère l'exécution des traitements et héberge le nœud manager du cluster Swarm ;
- 3 machines virtuelles noeud worker du cluster Swarm dont un noeud hébergeant un logiciel propriétaire (Agisoft Photoscan) ;
- un serveur de licence flottante pour le logiciel Agisoft Photoscan.

13. http://www.france-grilles.fr/catalogue-de-services/fg-cloud/

14. https://www.openstack.org/

4.6 Monitoring

La plateforme 4P faisant appel à de nombreuses technologies il est nécessaire de s'assurer du bon fonctionnement de chaque service.

Ce monitoring est effectué grâce au logiciel Prometheus[15]. Sur chaque machine virtuelle, une tâche permet d'envoyer des infos sur l'état des services à la machine manager. Les résultats sont agrégés et affichés sur l'interface de monitoring (voir **Figure 12**) .

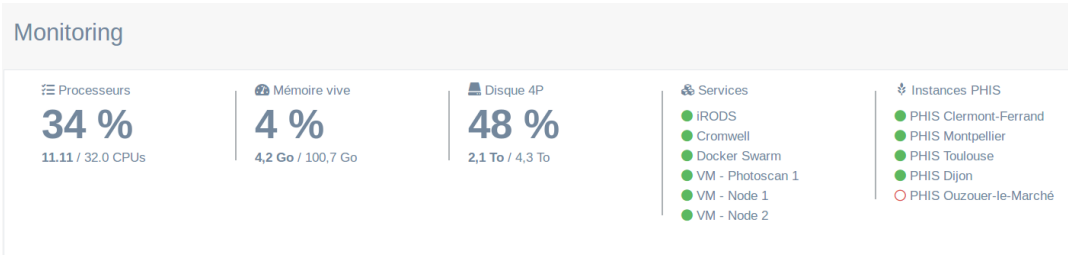


Figure 12 - Monitoring de la plateforme 4P.

5 Exemple d'utilisations

A titre d'illustration les traitements suivants ont été réalisées :

- analyse de données LIDAR pour le suivi de suivi de la hauteur de couvert (S.Madec, S. Thomas, voir **Figure 13**).

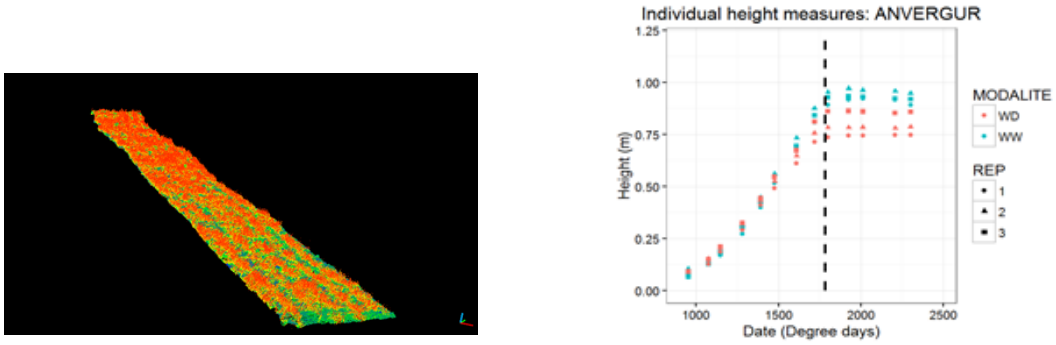


Figure 13 - Nuage de points 3D généré par le LIDAR sur la phénomobile sur une parcelle de blé (à gauche) et évolution de la hauteur de couvert déduite au cours du temps (à droite).

15. https://prometheus.io/

- analyse de données issues d'une caméra multispectrale pour le suivi spatial et temporel de l'indice foliaire (surface de feuilles sur 1m² de sol) (voir **Figure 14**). On observe bien des différences de niveau d'indice foliaire entre micro-parcelles, à la fois dues à des différences génétiques mais aussi à la différence entre les deux modalités: pas de stress hydrique (WW) ou stress hydrique (WO)

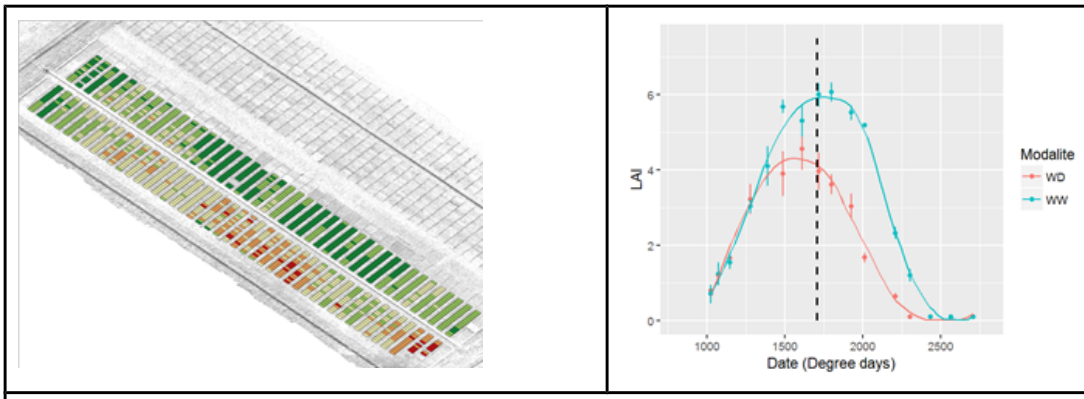


Figure 14 - Carte d'indice foliaire générée à partir d'une caméra multispectrale sur une plateforme de phénotypage (à gauche) et évolution de la surface foliaire au cours du temps pour deux micro-parcelles correspondant au même génotype avec (WO) ou sans stress hydrique (WW (à droite).

6 Conclusion et perspectives

Le cahier des charges a été respecté et les fonctionnalités prévues ont été implémentées dans les temps grâce notamment à :

- **une bonne communication** entre les partenaires du projet, notamment avec le prestataire informatique qui a su prendre en considération les demandes (parfois complexes) des utilisateurs et proposer une solution intégrant de multiples technologies dans la plateforme ;
- **au support fourni par France Grilles** qui a également été déterminant pour déployer rapidement la plateforme.

Les perspectives de ce projet sont les suivantes :

- **développer de nouvelles macros** (données phénomobile, ré-encodage modules matlab, passage en python) ;
- **renforcer l'intégration** avec le système d'information scientifique PHIS (récupération des variables, publication des traitements) ;

- **optimiser l'exécution des workflows** (instanciation de machines virtuelles ou de containers à la volée).

7 Remerciements

Nous tenons à remercier l'ANR qui a financé le projet PHENOME-EMPHASIS ainsi que l'équipe France Grilles pour son support technique.

8 Bibliographie

[1] Li L, Zhang Q, Huang D. A review of imaging techniques for plant phenotyping. *Sensors (Basel)*. 2014;14(11):20078–20111. Published 2014 Oct 24. doi:10.3390/s141120078.

[2] Blancon, J., Dutartre, D., Tixier, M.-H., Weiss, M., Comar, A., Praud, S., & Baret, F. (2019). A High-Throughput Model-Assisted Method for Phenotyping Maize Green Leaf Area Index Dynamics Using Unmanned Aerial Vehicle Imagery. Frontiers in Plant Science, 10, 685.

[3] Dutartre, D., Weiss, M., Thomas, S., Baret, F., De Solan, B., & Maupas, F. (2015). Green fraction (GF) estimates from RGB images: automatic classification based on Support Vector Machine In E.P.P.N. (EPPN) (Ed.), Plant Phenotyping Symposium. Barcelona, (Spain).

[4] Jay, S., Baret, F., Dutartre, D., Malatesta, G., Héno, S., Comar, A., Weiss, M., & Maupas, F. (2018). Exploiting the centimeter resolution of UAV multispectral imagery to improve remote-sensing estimates of canopy structure and biochemistry in sugar beet crops. Remote Sensing of Environment

[5] Jin, X., Liu, S., Baret, F., Hemerlé, M., & Comar, A. (2017). Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. Remote Sensing of Environment, 198, 105-114.

[6] Liu, S., Baret, F., Abichou, M., Boudon, F., Thomas, S., Zhao, K., Fournier, C., Andrieu, B., Irfan, K., Hemmerlé, M., & Solan, B.d. (2017). Estimating wheat green area index from ground-based LiDAR measurement using a 3D canopy structure model. Agricultural and Forest Meteorology, 247, 12-20.

[7] Madec, S., Baret, F., de Solan, B., Thomas, S., Dutartre, D., Jezequel, S., Hemmerlé, M., Colombeau, G., & Comar, A. (2017). High-Throughput Phenotyping of Plant Height: Comparing Unmanned Aerial Vehicles and Ground LiDAR Estimates. Frontiers in Plant Science, 8, 2002.

[8] 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. Dealing with multi-source and multi-scale information in plant phenomics: the ontology-driven Phenotyping Hybrid Information System. New Phytol. 2019 Jan;221(1):588-601. doi: 10.1111/nph.15385.