Direction du Système d'Information

Service **"Informatique scientifique et Appui aux Partenaire du Sud"**

Formulaire de demande

SPIRALES

« Soutien aux Projets Informatiques dans les Equipes Scientifiques »

Appel à projets 2015

Date de clôture: 16 janvier 2015

La mise en œuvre de l'appel à projets est réalisée par la DSI de l'IRD

Contact

spirales@ird.fr

1 Nature du projet

1.1 Titre du projet

Développement d'un outil générique d'indexation pour optimiser l'exploitation de données biologiques structurées, semi-structurées et non-structurées (BIO eSAI).

1.2 Résumé du projet proposé (5 lignes maximum)

Des études de la diversité des riz vietnamien sont conduites au LMI RICE dans le but d'identifier des gènes d'intérêt pour l'amélioration de variétés locales. Ces études requièrent la manipulation d'un volume important de données hétérogènes (fichiers textes, images et métadonnées associées, bases de données relationnelles). Dans ce contexte, le LMI RICE souhaite développer un outil d'indexation générique afin de pouvoir naviguer, partager et annoter ces données dans l'intérêt de les exploiter et les diffuser au mieux.

1.3 Type de projet

o Continuum (préciser année de démarrage : 2014)

2 Porteur(s) et collaborateur(s) du projet

2.1 Unité

o UMR N° 532 Nom : **DIADE** et **LMI** Rice Functional Genomics and Plant Biotechnology (RICE) – Hanoi, Vietnam

2.2 Département

o Environnement & Ressources

2.3 Statut et coordonnées du porteur de projet

Pierre LARMANDE – Permanent / IE2 IRD – UMR DIADE, Montpellier – 0467416290 – pierre.larmande@ird.fr

2.4 Nom et coordonnées du Directeur d'Unité (si différent)

Lebrun Michel – Permanent /Professeur UM 2– Hanoi, Vietnam – lebrun@univ-montp2.fr

2.5 Avis du directeur d'unité (obligatoire)

Le DU doit être garant de l'esprit incitatif de SPIRALES, et confirmer qu'il est prêt à assurer sur les fonds propres de l'unité la vie de l'outil (hébergement et maintenance) après la phase de développement soutenue par "SPIRALES". La DSI soutiendra financièrement l'hébergement de 3 projets par unité.

Ce projet de système d'indexation de données multi-formats est très important pour exploiter au mieux nos résultats de phénotypage et de génotypage visant à valoriser ou mieux connaitre les ressources génétiques

locales (riz, pathogènes) dans un contexte de travail multi-partenarial et international qui caractérise le LMI RICE (IRD, CIRAD, AGI, USTH). L'aspect novateur qui réside dans la mise au point d'un système évolutif permettra d'intégrer les données qui seront obtenues dans le cadre de projets en émergence qui seront réalisés par les partenaires en s'appuyant sur des infrastructures du LMI. Le LMI RICE soutient fortement ce projet, plusieurs projets financés sont en cours pour l'acquisition des données (phénotypage, génotypage). Les membres permanents du LMI RICE consacreront une partie de leur temps pour développer le projet. Cette seconde phase de développement de l'outil BIOeSAI nécessite des compétences spécifiques. Par conséquent, ce travail sera principalement effectué à Montpellier dans le cadre de l'UMR DIADE (P. Larmande) en collaboration avec le LMI RICE.

2.6 Site(s) de déroulement du projet

1. UMR DIADE, Montpellier (France)
2. Agricultural Genetics Institute (AGI), Hanoi (Vietnam), site d'implantation du LMI RICE

2.7 Site administratif à partir duquel se feront les dépenses budgétaires

Centre IRD Montpellier

2.8 Liste des unités (ou organismes partenaires) du projet

UMR DIADE (IRD/UM2/CIRAD) - Montpellier
UMR MISTEA (INRA/Supagro) - Montpellier
UMR IMPE (IRD/UM2/CIRAD) - Montpellier
University of Science and Technology of Hanoi (USTH)
UMI UMMISCO (IRD/UPMC)
Localisation géographique : Montpellier (France) et Hanoi (Vietnam)

2.9 Liste des intervenants impliqués de manière effective dans la réalisation du projet

Prénom Nom - Statut / Catégorie – Organisme (unité/laboratoire) - Localisation géographique - Email – ***Contribution en % de temps homme ou en jours*homme (ETP total ou pour une période)***

LE Ngoc Luyen - Stagiaire Master 2 UMII – IRD – Montpellier – ***6 mois plein-temps soit 100 jours ETP***
Pierre LARMANDE – Permanent / IE2 – IRD (UMR DIADE) - Montpellier – pierre.larmande@ird.fr – ***2 jour/semaine soit 70 jours ETP***
Stéphane JOUANNIC – Permanent / CR1 – IRD (UMR DIADE) – Hanoi – stephane.jouannic@ird.fr – ***0,5 jour/semaine soit 20 jours ETP***
Michel LEBRUN – Permanent / PR1 – UM2 (UMR DIADE-LMI RICE) – Hanoi – pascal.gantet@univ-montp2.fr – ***0,5 jour/semaine soit 20 jours ETP***
Stéphane BELLAFIORE – Permanent / CR2 – IRD (UMR RPB) – Hanoi – stephane.bellafiore@ird.fr – ***10 jours ETP***
Anne Tireau – Permanent / CR1 – INRA (UMR MISTEA) – Montpellier supagro - tireau@supagro.inra.fr ***0,5 jour/semaine soit 20 jours ETP***
Pascal Neveu – Permanent / IR0 – INRA (UMR MISTEA) – Montpellier supagro - tireau@supagro.inra.fr ***0,5 jour/semaine soit 20 jours ETP***
LUONG Chi May - Permanent / PR (IOIT) – Hanoi – lcmai@ioit.ac.vn – ***5 jours ETP***
PHAN VU Trung – Administrateur système USTH, Hanoi – ***5 jours ETP***

3 Moyens / appuis demandés à la DSI

3.1 Soutien demandé à la DSI pour 2012

Soutien demandé :

- soutien en accompagnement:
  - ▮ à la préparation du projet informatique (expression des besoins, étude de faisabilité)
  - ▮ à la réalisation du projet informatique – *préciser si possible les compétences attendues (web, SIG, SGBDR…)*
  - ☐ à la rédaction d'un dossier de financement (H2020, ANR, autre)
  - ☐ à l'identification d'un opérateur informatique répondant à des besoins de stockage massif, archivage, ou d'hébergement
- soutien pour l'hébergement:
  - ▮ de l'applicatif scientifique sur une machine virtuelle (accès root autorisé)
  - ☐ d'un serveur physique (cas particulier où une machine virtuelle ne conviendrait pas)
- soutien pour l'utilisation d'outils:
  - ☐ Redmine[1] pour le suivi du développement des fonctionnalités et des bugs, et le suivi de projet
  - ▮ SVN [2] pour le partage du code source au sein de l'équipe de développement
  - ☐ PowerAMC[3] pour la modélisation de bases de données
  - ☐ La suite ArcGIS[4] (ArcView, ArcEditor, ArcInfo) de l'éditeur ESRI
  - ☐ ENVI[5] pour l'analyse de données géospatiales
- soutien financier (pour un besoin différent des soutiens précédents): **7000** € HT
  - ▮ justification:

    - Indemnité supplémentaires d'accueil de 2 stagiaires Vietnamiens pour 6 mois en France (Montpellier) : le premier pour travailler sur le projet SPIRALE en continuité de son travail en 2014 (**cf. sujet de stage 1**). Le deuxième pour travailler sur un projet de développement informatique qui sera connecte au projet SPIRALES **(cf. sujet de stage 2**). **3000** € (2 x 250 euros x 6 mois)

    - Mission de 2 semaines au LMI RICE (Vietnam) et au centre international du Riz (IRRI, Philippines) pour Pierre LARMANDE. Au LMI RICE, Pierre Larmande participera a la rédaction du cahier des charges et des tests a grande échelle (cf. Gant chart). Au Philippines, Pierre Larmande effectuera une mission pour développer un partenariat avec l'IRRI en présentant ce projet. **4000** € (vols France-Hanoi AR puis Hanoi – Manilles AR, per diem)

3.2 Montant(s) précédemment attribué(s) par la DSI - en euros HT

	2012	2013	2014
Montants attribués (€ HT)	0	0	4000

[1] http://www.redmine.org/
[2] Système libre de gestion de versions (http://subversion.apache.org/)
[3] http://www.sybase.fr/products/modelingdevelopment/poweramc
[4] http://www.esri.com/software/arcgis/arcgis-for-desktop/index.html
[5] http://www.exelisvis.fr/ProduitsetServices/LesproduitsENVI/ENVIpourArcGIS.aspx

3.3 Moyens affectés au projet et Cofinancements acquis hors SPIRALES (€ HT)

Autres sources de financements **acquis** (interne ou externe IRD) **pour ce projet** (ex. ANR, CE…)

Montant (€ HT) : 5 232 € pour le financement des 2 stagiaires Vietnamiens par le projet IDEX IBC

4 Description des besoins et du projet

4.1 Objectifs scientifiques (en précisant les aspects innovants)

Dans le cadre du laboratoire LMI RICE, des études de la diversité génotypique et phénotypique de variétés traditionnelles de riz vietnamien sont conduites dans le but d'identifier des gènes d'intérêts pour la compréhension de processus biologiques (développement et plasticité de la plante, résistance aux maladies, interactions bénéfiques) mais également pour des futurs programmes d'amélioration conduits par les partenaires vietnamiens. Ces études requièrent la manipulation d'un important volume de données hétérogènes de séquençage, de génotype, de phénotype. Ces données sont pour parties déjà disponibles et peuvent être stockées sous la forme de fichier Excel, texte structurée, images ou bases de données relationnelles.

Dans ce contexte, l'équipe du LMI RICE souhaite, par la mise en place d'un outil d'indexation, organiser ses propres jeux de données afin de pouvoir plus facilement les exploiter et les partager dans un contexte multi partenarial (LMI). Les projets développés au sein du LMI RICE à Hanoi sont à l'interface de trois UMR "Plantes" de l'IRD (DIADE, IPME et LSTM), unités qui bénéficieront directement des retombées de ce projet Spirales, notamment via le plateau bioinformatique de Montpellier (sous la co-responsabilité de Pierre Larmande). Par ailleurs, le LMI RICE est un laboratoire associé à l'USTH (université franco-vietnamienne à Hanoi), pour laquelle le développement d'un tel outil est d'un grand intérêt pour la formation et les différents laboratoires de recherche qui y sont associés. De plus le LMI RICE est en relation avec l'UMI UMMISCO et l'IoIT (Institute of Information Technology, Hanoi, Vietnam) qui pourront apporter leur support et leur expertise au développement du système d'indexation multi-données.

L'objectif scientifique de ce projet est donc de proposer et d'implémenter une solution de stockage et de gestion de fichiers de natures diverses (Excel, texte structurés, images, bases de données relationnelles), grâce à la conception d'un système « souple » (c'est a dire supportant le changement) en fonction des besoins des utilisateurs.
Au cours de l'année 2014 une première phase de réalisation a été effectue dans le cadre d'un stage de master finance par SPIRALE.
Un prototype d 'application a été développé en répondant aux objectif initiaux. Ce prototype est disponible sur une VM mise a disposition par l'équipe IS de l'IRD a l adresse http://vmbioesai-dev.ird.fr:8080/Syspherice/.

Les objectifs scientifiques de cette demande de renouvellement SPIRALES sont d'améliorer le fonctionnement de l'application pour la porter en phase de production. Cette phase nécessitera de tester l'application a plus grande échelle (jeu de données et nombre d'utilisateurs plus important) mais également de définir avec les utilisateurs des fonctionnalités leur permettant de couvrir l'ensemble de leurs besoins (meilleure gestion des requêtes, connexion avec des systèmes existants ; cf. 4.4)

Un des objectifs de la phase 2 sera d'améliorer la généricité de l'outil afin que son utilisation dépasse les besoins de l'unité LMI RICE. Pour ce faire, nous travaillerons en collaboration avec des équipes extérieures telles que l'unité MISTEA.

4.2 Description de l'existant (moyens – outils – compétences)

4.2.1 Nom de votre outil (dans le cas d'un développement d'application)

Biological Electronic Scientific Assistant Index (BIO eSAI)

4.2.2 Description de l'existant (moyens – outils – compétences)

Aujourd'hui le système comporte une interface de recherche (figure 1) afin de fouiller parmi les données indexées (image, fichiers Excel, documents texte). Toute la partie de gestion des documents et des index s'effectue dans un espace sécurisé (figure 2). Elle comprend notamment les fonctions de création de projets auxquels les documents sont associés, les fonctions d'import (figure 4) et de gestion des documents. Un exemple de gestion des images

est donne en figure 3. L'administration permet également de définir les index sur les documents et les requêtes qui permettent d'agréger plusieurs documents entre eux.
L'originalité du système tiens par la possibilité d'ajouter des annotations ou « tags » sur les documents et les données.
La partie des interfaces, constitue la majeure contribution du travail. Elle est directement liée aux besoins exprimés en terme de requêtes et visualisation. Nous nous sommes assuré que cette couche soit suffisamment générique pour être utilisée dans un autre contexte scientifique. C'est une phase que nous souhaiterions évaluer dans ce deuxième volet SPIRALE.

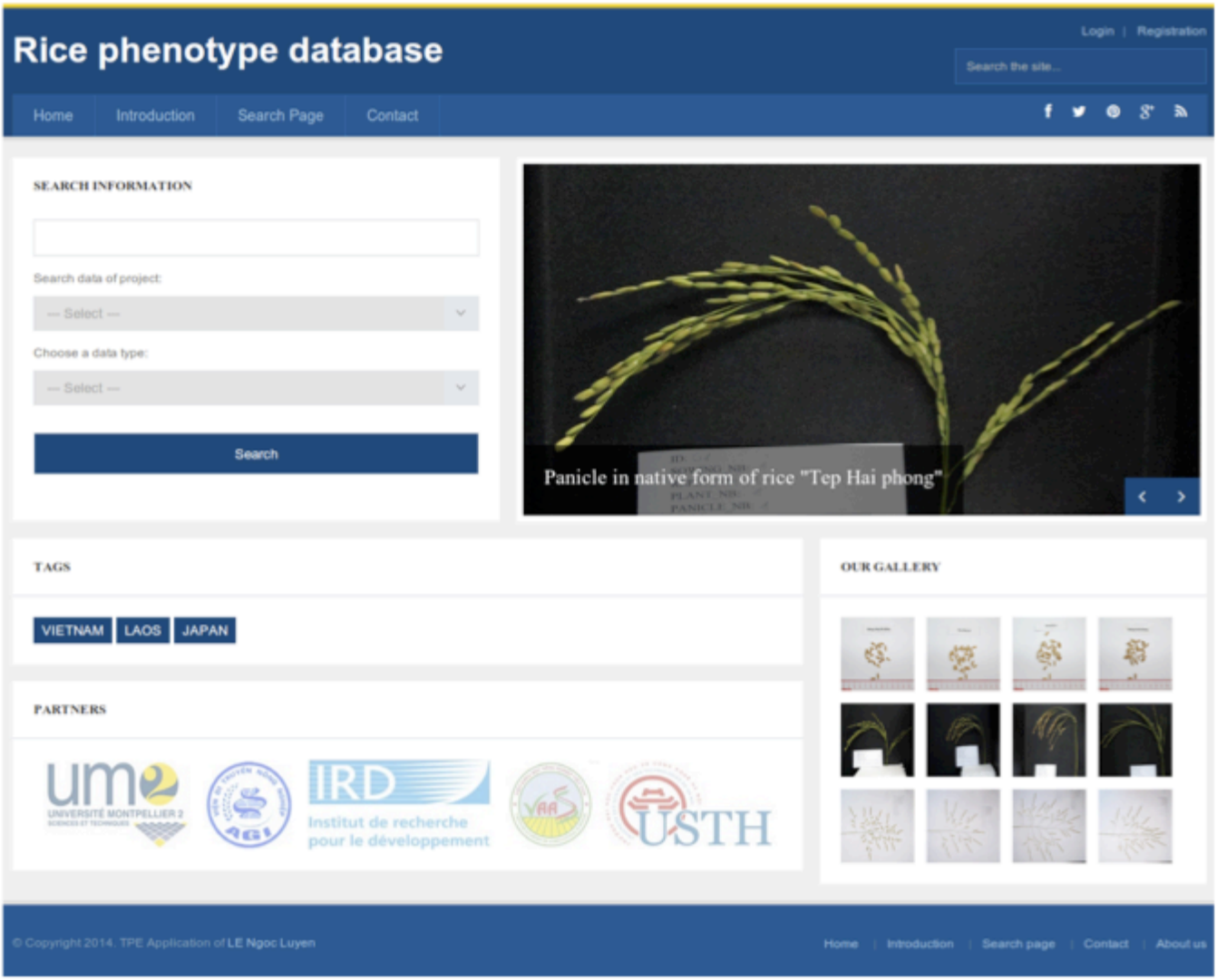


Figure 1: Interface de recherche principale

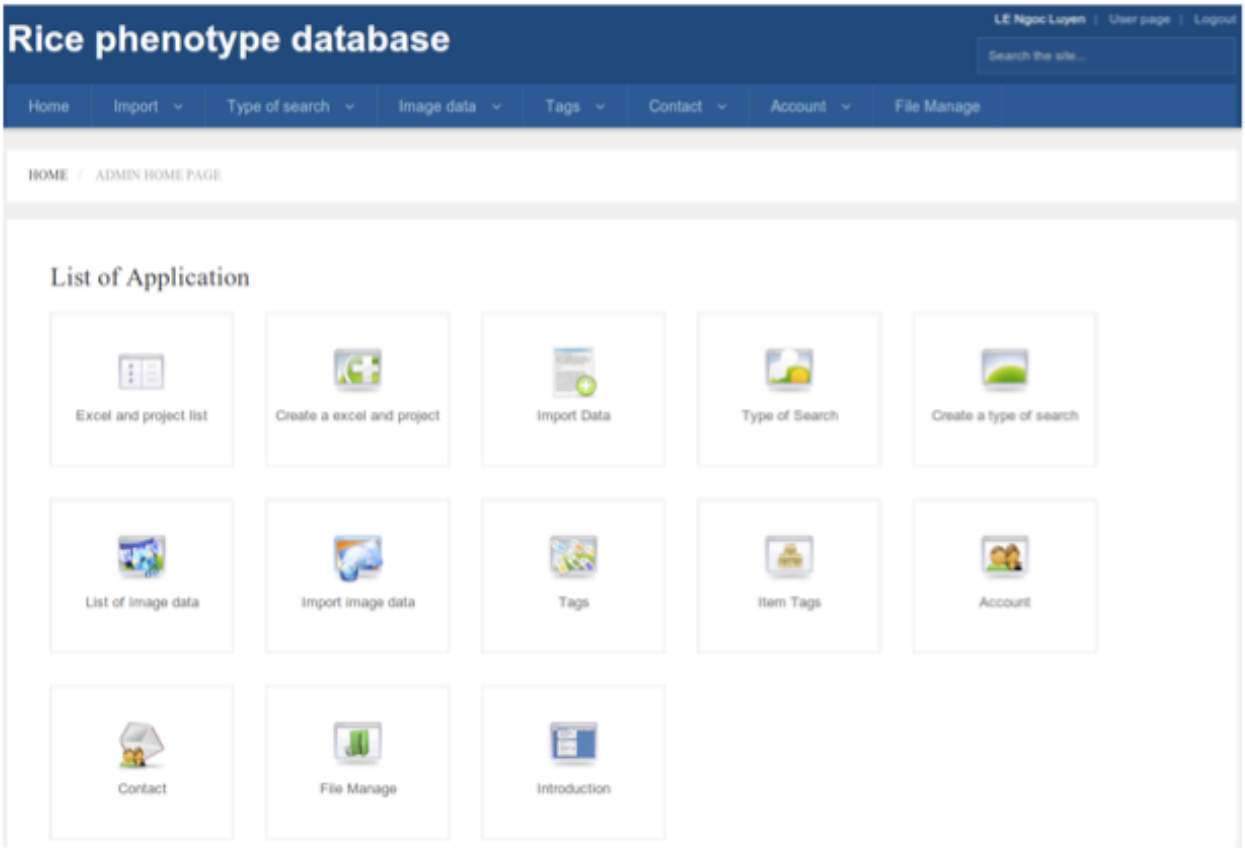


Figure 2: Interface de gestion du système

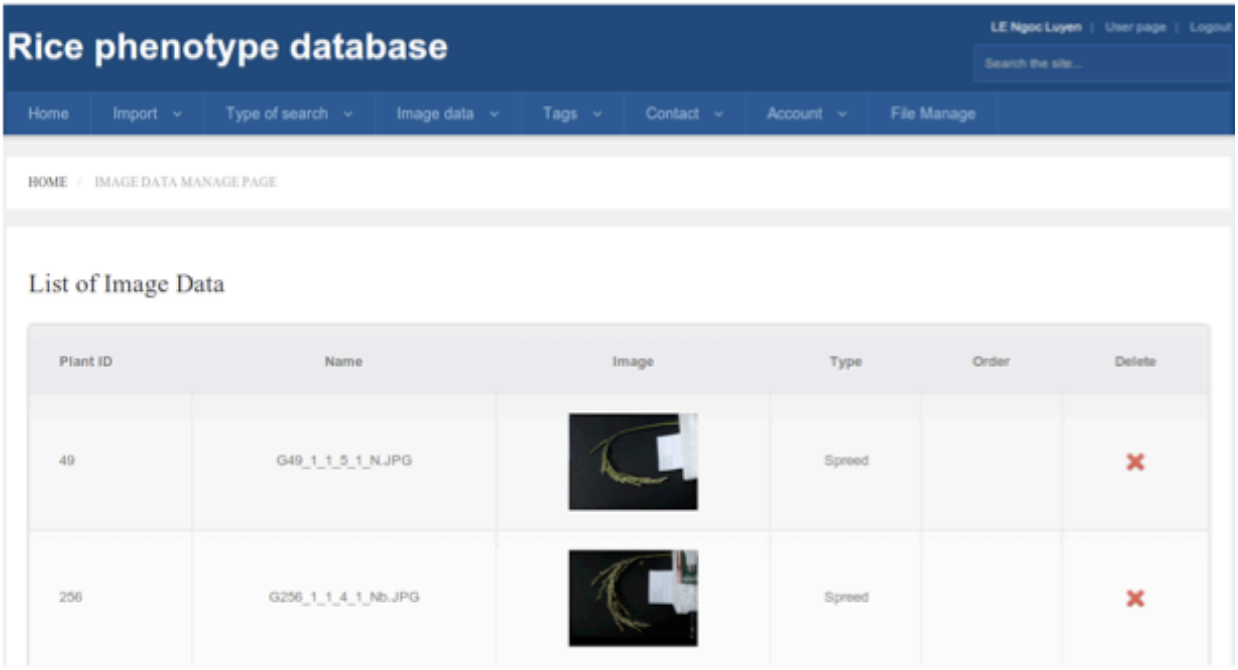


Figure 3: Interface de gestion des images

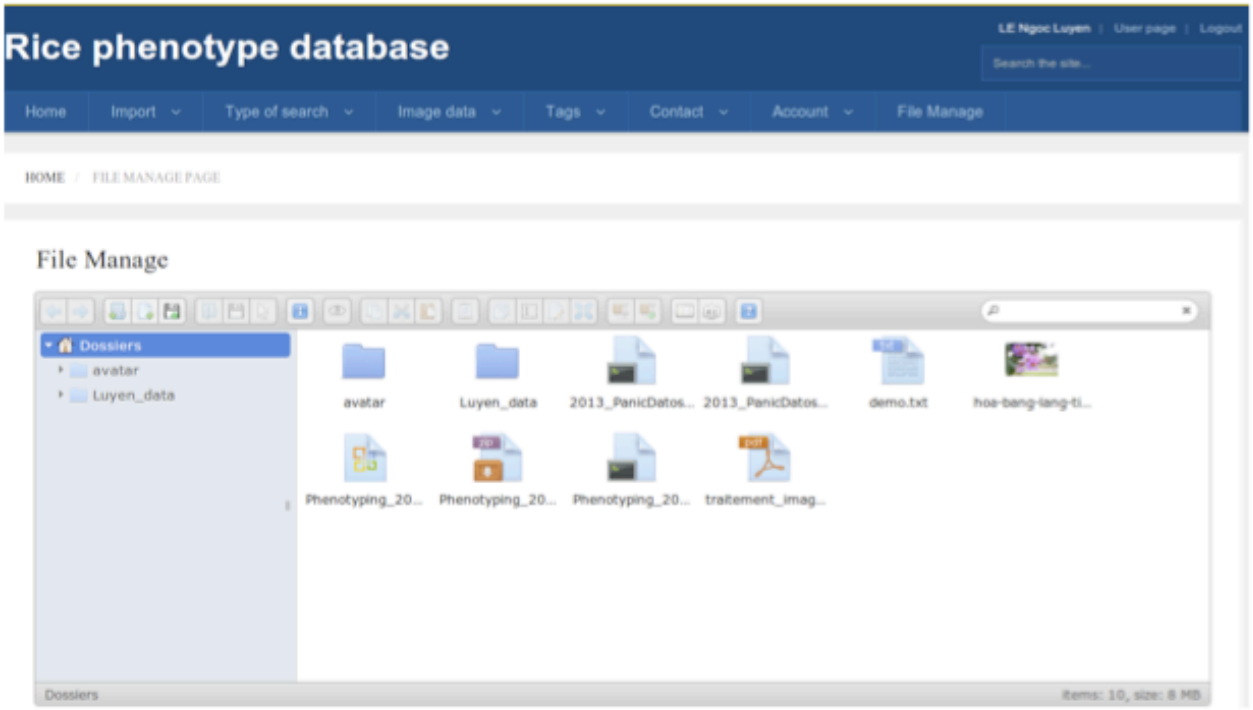
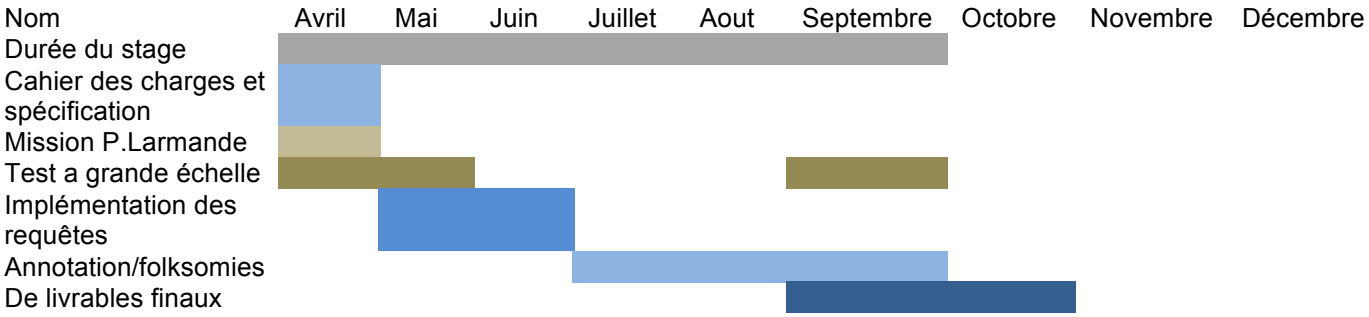


Figure 4: Interface d'import et de gestion des fichiers

4.3 Calendrier du projet (digramme de Gant souhaité)

4.3.1 Calendrier du projet (digramme de Gant souhaité)

Le planning du projet s'étale sur la durée d'un stage de Master 2 en informatique, soit du 1er avril au 31 aout 2015. Un sujet a été proposé à l'Institut de la Francophonie pour l'Informatique (IFI) d'Hanoi. Le candidat sélectionné sera le mémé étudiant qui a travaille sur l'application en 2014.

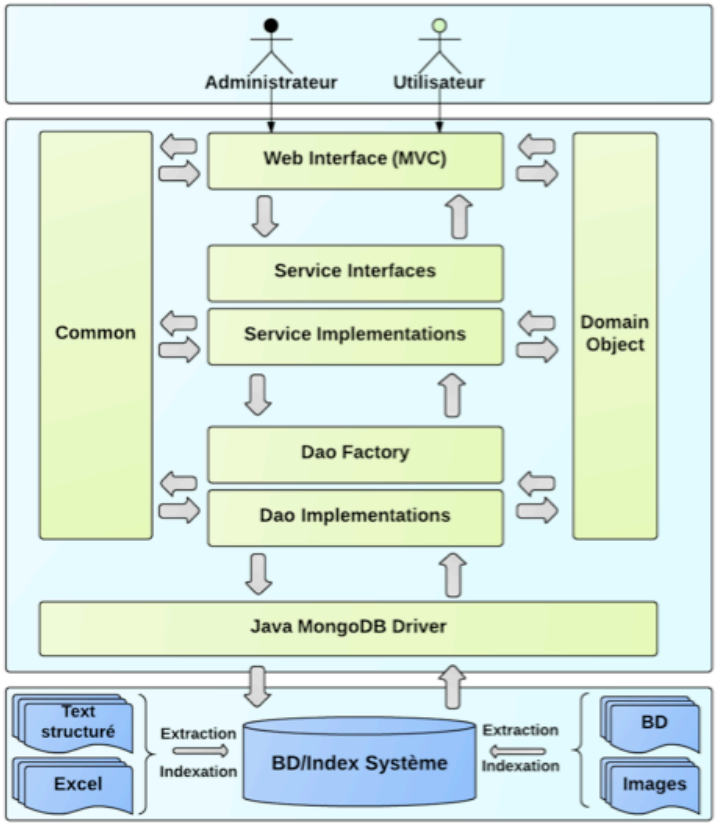


- La phase initiale sera d'effectuer l'inventaire des besoins en termes de fonctionnalités (cf. 4.4).
- Une mission au Vietnam de Pierre Larmande sera prévue lors du début de la phase de spécification.
- Une phase de test sera nécessaire pour évaluer la volumétrie et la vélocité de l'application.
- Une phase d'implémentation des requêtes sera réalisée.
- Le travail sur les annotations sera réalisé en dernière partie de projet

4.4 Décrire l'architecture envisagée pour votre outil (*un schéma sera apprécié*)

Le système est composé de 3 parties (n-tiers/architecture) programmées en Java:
La couche d'accès aux données, gère le plus souvent des données persistantes au sein d'un Système de gestion de base de données (SGBD). Dans le modele ci-dessous, la couche de l'accès consiste à DAO Factory,

DAO implémentation et les API de “mongodb java driver” pour interagir avec SGBD. Dans notre cas nous avons réutilisé des API existantes telles que Java SPRING.
La couche de métier est indépendante de toute forme d'interface avec l'utilisateur. Ainsi elle est utilisable aussi bien avec une interface console, une interface web, une interface de client riche. C'est généralement la couche la plus stable de l'architecture. Elle ne change pas si on change l'interface utilisateur ou la façon d'accéder aux données nécessaires au fonctionnement de l'application.
La couche interface utilisateur est l'interface qui permet à l'utilisateur de piloter l'application et d'en recevoir des informations par l’utilisation des requêtes et les réponses du protocole HTTP. On utilise le modèle de MVC pour construire les interactions avec utilisateur et les données du système.



A l’issu de la première phase du projet (SPIRALE 2014) nous avons identifier des points d’amélioration du système :

- Tester l’application a grande échelle, c est a dire avec une volumétrie plus importante et en augmentant le nombre d’utilisateurs.
- Améliorer les fonctionnalités de recherche
    - inclure la possibilité de faire des requêtes conditionnelles (OR – AND – NOT)
    - inclure la recherche par intervalle de valeur sur les champs numériques
    - gestion des recherches multi langues

- Utiliser les annotations pour mettre de la sémantique (folksomie[6]).

6 http://fr.wikipedia.org/wiki/Folksonomie

4.5 Énumérer et décrire les données/méta données de votre outil (thématique, format, volume, …)

Gérer des données de recherche fait partie intégrante du projet de recherche. Connaître les jeux de données utilisés pendant le projet est une information importante.

Jeux de données qui seront utilisés pendant le projet	Type de données	Format	Volume
Données phénotype	images	jpeg	400Go
Données phénotype	texte	Excel et Txt	200Go

4.6 Stockage, sauvegarde, Lister les méthodes/référentiels, langages de programmation…

Gérer des données de recherche fait partie intégrante du projet de recherche. Connaître l'organisation mise en place pour stocker, sauvegarder et diffuser les données du projet est une information importante.

Note : l'application est actuellement hébergée sur une VM fournie par IS-DSI, nous évaluons également la proposition de stockage et de partage des données.

Lieu de stockage des données du projet : LMI RICE une copie sur le serveur d'analyse du LMI

Plan de sauvegarde : Un backup hebdomadaire sur disque dur externe + redondance de disque RAID

Personne ou équipe responsable de la sauvegarde : Plateau bioinformatique (contact Pierre Larmande)

Mécanisme de partage des données (en ligne ? conditions de restriction ? …) : Pas encore évalué

Logiciels utilisés par l'application :

- NoSQL databases (mongoDB)
- UML (Modeling language)
- Java (programming language)
- Javascript, Ajax, jQuery (web langages)
-

La plupart des logiciels utilisés sont soumis à des licences libres (open GL, BSD, Apache ou GPL). De manière générale nous privilégierons les logiciels avec ce type de licence.
Le langage Java est utilisé pour réaliser les couches d'accès aux données ainsi que les interfaces de visualisation et d'administration. Dans ce cas, nous sommes dans un environnement web (Java, Javascript, Ajax, jQuery).

4.7 Liste des livrables et documents (spécifications fonctionnelles, techniques, API, manuel utilisation…)

Le logiciel final sera mis en ligne sous la forme d'un package web application (+ dépendances) à déployer sur un serveur Apache Tomcat.

Nom du document	Date de réception	Descriptif du document
Cahier des charges	20/04/2015	Synthèse des besoins des utilisateurs.
Document de planification	29/04/2015	Document de planification du travail
Document d'installation technique et manuel d'utilisation	15/08/2015	Fiche technique décrivant l'installation de l' application

5 Bénéfices pour le Sud (cf objectifs dans le “guide du candidat”)

5.1 Sites de déroulement du projet au Sud

Le site principal d’évaluation de cet outil sera l'**AGI** (Agricultural Genetics Institute, Hanoi, Vietnam), site d'implantation du LMI RICE alors que la majorité des développements s’effectuerons sur le site de Montpellier par la supervision de Pierre Larmande.

5.2 Sociétés publiques/privées du Sud impliquées

Néant

5.3 Liste exhaustive des partenaires au Sud

La recherche de synergie ou de partenariat (projet inter-unités impliquant des partenaires du Sud) et la mobilisation de compétences extérieures à l'unité doit être recherchée.

Prénom Nom – Organisme (laboratoire/unité) – Lieu géographique – Email – Type de bénéfice

Pr. Do Nang Vinh, Agricultural Genetics Institute (AGI), Hanoi Vietnam, nangvinhdo@gmail.com, application de l'outil pour activités de recherche

Dr. Tran Thu Hoai, Plant Resource Centre (PRC), Hanoi Vietnam, thuhoai70@yahoo.com, application de l'outil pour activités de recherche

Pr. Michel Lebrun, University of Science and Technology of Hanoi (USTH), Hanoi Vietnam, formation des étudiants

Ho Tuong Vinh - UMI UMMISCO – IFI (IRD/UPMC) - Hanoi Vietnam - ho.tuong.vinh@ifi.edu.vn - Formation des étudiants

5.4 Pérennité du projet

Le LMI RICE accueille plusieurs projets de recherche centrés sur le riz, impliquant plusieurs UMR IRD (DIADE, IPME, LSTM) mais également en collaboration avec d'autres UMR de Montpellier (BPMP, AGAP). Les projets scientifiques développés au sein du LMI RICE ont pour objectif de produire de la connaissance (génotypage, phénotypage) relative à l'amélioration des variétés de riz au Vietnam et de former les futurs chercheurs de l'AGI. Ces projets ne peuvent se développer que sur le long terme en partenariat avec différentes institutions vietnamiennes. L'outil développé contribuera à cette pérennisation des projets scientifiques en facilitant la mémorisation, la valorisation la diffusion et l’utilisation des données acquises. L'objectif est également de déployer cet outil vers différents partenaires, différentes UMR. Le LMI RICE alimentera et encadrera le développement de ce projet. Par ailleurs tout nouveau projet/demande de financement extérieur impliquant l'utilisation de cet outil, son développement et son alimentation en terme de données devront consacrer une partie de leur budget à cette fin.
L'administration du serveur de stockage des données et de l'outil développé dans le cadre de ce projet se fera grâce au soutien de l'administrateur système de l'USTH et du plateau bioinformatique de l'IRD-Montpellier. De plus, l’encadrement des ingénieurs du plateau bioinformatique garantira la maintenance de l’application à la fin du projet.

5.5 Renforcement des capacités des partenaires

L'objectif est que les partenaires acquièrent de plus en plus d'autonomie quant à l'archivage et la valorisation des données relative à leurs projets scientifiques via l'utilisation de cet outil. Les chercheurs de l'AGI ainsi que les étudiants formés au sein du LMI (les futurs chercheurs de l'AGI) seront les premiers contributeurs et utilisateurs de

cet outil. Cet outil constituera également une vitrine pour la communication des données acquises en partenariat avec le LMI et sera transféré vers d'autres instituts vietnamiens tels que le PRC et le PPRI. Plus globalement, l'objectif est de produire un outil d'indexation générique utilisable pour un grand nombre d'applications, de projets basés sur l'utilisation de fichiers multi formats.

6 Actions transversales

Un projet "SPIRALES" ne peut être le projet d'une unité ; il a vocation à être valorisé et être réutilisé au sein de l'institut, et à l'extérieur. Une démarche de capitalisation doit être recherchée.

6.1 Protection de code

La plupart des logiciels évalués sont soumis à des licences libres (open GL, BSD, Apache ou GPL). De manière générale nous privilégierons les logiciels avec ce type de licence. Pour répondre aux besoins de propriété intellectuelle, les codes feront l'objet d'un dépôt à l'Agence de protection des programmes.

6.2 Transfert de technologie

Le projet BIOeSAI a travers le soutien des projets SPIRALES souhaite transférer sa technologie avec ses partenaires Vietnamien de l'AGI travaillant au LMI RICE. De plus, nous avons la volonté de développer des collaborations avec l'IRRI dans ce contexte d'utilisation et de développement logiciel.

6.3 Ré-utilisation d'anciens SPIRALES

Ce projet s'appuie sur les acquis du projet Spirale BIOeSAI 2014.

6.4 Communications

Une publication dédiée au développement et à l'exploitation de cet outil est prévue en 2015. Les participants à ce projet ont déjà l'expérience de valorisation d'outil informatique : OryzaTagLineDB (Larmande et al. 2008), P- TRAP (AL-Tam et al. 2013). Par ailleurs, cet outil pourra être également valorisé au travers de plusieurs publications qui porteront sur les différents résultats scientifiques faisant références aux analyses associées à cet outil. De mémé, ce produit sera mentionné systématiquement lors de participation à des congres (communication orales ou poster) relatifs aux résultats scientifiques obtenus.

7 SUJET DE STAGE 1 : Développement d'un système connaissances pour BIG DATA application aux données de phénotypage chez le riz (*O. sativa*)

Encadrement: IRD – Pierre LARMANDE

Collaboration externe : Stéphane JOUANNIC, JONQUET Clément, Patrick Valduriez , UMR MISTEA - INRA

Contexte

Dans le cadre de l'équipe Génome et Développement des Riz (GDR) et du LMI RICE (Hanoi), des études de la diversité génotypique et phénotypique de variétés traditionnelles de riz vietnamien sont conduites dans le but d'identifier des gènes d'intérêt pour la compréhension de processus biologiques (développement et plasticité de la plante, résistance aux maladies) mais également pour des futur programmes d'amélioration. Ces études requièrent **la manipulation d'un important volume de données hétérogènes. Ces données peuvent être stockées sous la forme de fichier Excel, texte structuré, images ou bases de données relationnelles. Dans ce contexte, l'équipe de GDR souhaite organiser ses propres jeux de données afin de pouvoir naviguer, partager et annoter ces dernières afin de les exploiter au mieux.**

L'implémentation de systèmes d'information utilisant les bases de données relationnelles n'est pas adaptée à notre problématique car cette méthode n'est pas assez flexible et évolutive. L'objectif scientifique de ce projet est donc de proposer et d'implémenter une solution de stockage, de gestion et de consultation de fichiers de natures diverses (Excel, texte structurés, images, bases de données relationnelles), grâce à la conception d'un système « souple » (c'est a dire supportant le changement) en fonction des besoins des utilisateurs.

La difficulté réside dans la définition de systèmes « souples », c'est à dire supportant une évolution des besoins utilisateurs avec un minimum de développement. L'importance des données médias (images dans ce cas) est à prendre en compte. En effet, leur association avec les jeux de données « textuelles » est évidente, mais elle nécessitent également la prise en compte de « méta-informations » d'abords basique comme l'auteur, la date, le lieu, géolocalisation, puis élaborée comme un système de « tagging » permettant de rechercher des associations entre les jeux de données.

Objectifs

Un système d'information a été implémenté lors d'un stage de Master 1 en 2014. Ce système est basé sur un SGBD NoSQL incluant également la gestion des métadonnées et des tags. Toutefois, la méthode mise en place ne permet pas de détecter des relations explicites/implicites entre les données gérées par le système.

L'objectif du stage proposé sera d'évaluer la faisabilité de gestion des BIG DATA couplé au technologies du Web Sémantique en s'appuyant sur les articles de synthèse du domaine (Shiri, 2014; Wu & Yamaguchi, 2014). Par ailleurs, un état de l'art de solutions existante telles que les technologies proposées par Duraspace (http://www.duraspace.org/) sera envisagé. Le sujet s'inspirera également de solutions développées dans le domaine biologique (Kawano et al., 2014).

Profil recherché:
- Master 2 informatique, bioinformatique
- Solides connaissances du langage de programmation Java
- Bonnes connaissances des systèmes de gestion de bases de données (NoSQL, NewSQL, SQL)
- Bonnes connaissances du Web sémantique (RDF, SPARL, RDF Triple store)
- Autonomie
- Bon relationnel

Gratification: 436 euros
Candidature: CV and covering letter
Contacts: pierre.larmande@ird.fr

References:

Kawano, S., Watanabe, T., Mizuguchi, S., Araki, N., Katayama, T., & Yamaguchi, A. (2014). TogoTable: cross-database annotation system using the Resource Description Framework (RDF) data model. *Nucleic Acids Research*, *42*(Web Server issue), W442–8. doi:10.1093/nar/gku403

Shiri, A. (2014). Linked Data Meets Big Data : A Knowledge Organization Systems Perspective, *24*, 16–20. doi:10.7152/acro.v24i1.14672

Wu, H., & Yamaguchi, A. (2014). Semantic Web technologies for the big data in life sciences. *BioScience Trends*, *8*(4), 192–201. doi:10.5582/bst.2014.01048

8 SUJET DE STAGE 2: The Agronomic Linked Data (AgroLD) project.

Supervisors: Pierre Larmande, IRD and Aravind Venkatesan, IBC

Keywords: Data Integration, Information extraction, Knowledge management, Semantic Web, Linked Open Data, Bioinformatics

Background:
Agronomy is an overarching field that consists of various areas of research such Genetics, Plant Molecular Biology, Ecology and Earth Science. To effectively develop applications to improve crop production through sustainable methods, it is important to over lay research findings from these fields as they are highly inter connected. We are currently witnessing rapid advancements in information technologies that continue to drive a flood of information and analysis techniques within the domains mentioned above. However, the information currently available are highly distributed and patchy in nature. Using these resources more effectively and taking advantage of associated cross-disciplinary research opportunities poses a major challenge to both domain scientists and information technologists.
At the Institute of Computational Biology[7] (IBC), we are involved in developing methods to aid data integration and knowledge management within the plant biology domain to improve information accessibility, sharablity within the domain. We address this challenge by pursuing several complementary research directions in: distributed, heterogeneous data integration.

Objective:
To build on this momentum, we at IBC are currently building a RDF knowledge base, **Agro**nomic **L**inked **D**ata (AgroLD). The knowledge base is designed to integrate data from various publically available plant centric data sources such as Gramene[8], Oryzabase[9], TAIR[10] and resources from the SouthGreen platform[11], to name a few. The aim of AgroLD project is to provide a portal for bioinformatics and domain experts to exploit the homogenized data model towards filling the knowledge gaps. To this end, we plan to engage with stakeholders in demonstrating the advantages of SW in answering complex domain relevant questions that were unapproachable using traditional methods, strategically filling knowledge gaps.

The internship proposal is to contribute to this project by:

- Be involved in the building of AgroLD, such as identifying additional data resources for integration.
- Develop parsers and wrappers for existing public RDF triplestores.
- Develop a web application allowing queries to remote and local resources.
- Develop a query builder to improve information retrieval.
- The student will also work on documentation of the work and manuscript writing.

Candidate profile:
- Master 2 informatics,
- Strong knowledge in Java and Python,
- Good knowledge of RDBMS,
- Good knowledge of the Semantic Web (RDF, SPARL, RDF triple store),
- Ability to work independently,
- Good interpersonal skills and ability to work in a group.

Compensation: 436 euros
Application: CV and covering letter
Contacts: pierre.larmande@ird.fr, Aravind.Venkatesan@lirmm.fr

[7] IBC: http://www.ibc-montpellier.fr/
[8] Gramene: http://gramene.org/
[9] Oryzabase : http://www.shigen.nig.ac.jp/rice/oryzabase/
[10] TAIR : http://www.arabidopsis.org/
[11] SouthGreen platform: http://southgreen.fr/databases