AG du Cati CODEX

14 et 15 février 2013. Campus SupAgro à Montpellier

Participants : 23 personnes dont Isabelle Blanc (Supsis), Patrice Buche (ICAT), Nicolas Donès (IUMA), Cyril Pommier (CGI), Hervé Richard (CaScisDi).

Programme :

JEUDI 14 février

9h30-10h00. Laurent Bruckler – président du centre de Montpellier *: métiers de l'informatique à l'INRA, nouveaux enjeux*

Laurent Bruckler présente le Cati Codex comme une force « locale » sur l'analyse des données et des connaissances d'expérimentations. Puis, il trace le portrait de l'informatique à l'INRA qui, en 2013, représente 4 millions d'euros. Les 276 agents de BAP E (calcul scientifique) dans les unités de recherche, constituent la 2ème branche d'activité à l'INRA après la BAP A des 395 biologistes. L'informatique se répartit en 3 activités : 156 agents en développement, 70 en calcul scientifique et 50 pour la maintenance du réseau. L'âge moyen des agents de la BAP E est de 43 ans alors que pour l'ensemble des BAP elle est de 47 ans indiquant que c'est en BAP E que l'INRA a recruté le plus, récemment. L'objectif est de rapprocher la biologie de l'informatique/statistique/mathématique. D'autant plus que la biologie devient très dépendante des sciences numériques.
Cela induit des changements d'organisation tels que :

- L'émergence de Data Centers tels qu'à Toulouse, et probablement à Montpellier et Jouy en Josas (confirmée depuis).
- La mise en place de réseaux de regroupement de métiers : PEPI.
- Les Catis pour la production informatique.
- Le schéma directeur des systèmes d'information INRA.

Parallèlement, la configuration de la science est en train de changer : elle devient participative. En témoigne, le réseau Nutrinet Santé qui regroupe 250 000 personnes volontaires qui répondent à des enquêtes annuelles. A raison de 4 heures/an par personne, cela représente 600 postes de permanents !! Il faut faire face à une production de données massives, hétérogènes, incertaines. Cela génère des questions de recherche en math/info.
Nous sommes au cœur de ses évolutions.

Parmi les questions avec les participants, notons le besoin exprimé de proposer des formations en statistique et en mathématique pour les informaticiens qui n'ont ni les compétences, ni les connaissances en modélisation.

10h00-10h30. Isabelle Blanc : Open Science et Open Data à l'INRA

L'INRA participe à des groupes internationaux d'Open Data en agriculture et en science en général. Le premier groupe était en 2004 l'Open Data de la promotion du libre accès aux publications scientifiques.
Au niveau plus global : Big Data = ensemble des données interprétées et publiées et aussi celles (la plus grosse partie) qui sont brutes.
Il y a des revues qui favorisent la mise en accès des données.
Le Data citation index est en pleine émergence : évaluation des liens et du nombre de publications issues de données disponibles.
Enjeux du partage de données : accessibilité des données et réutilisation des données dans un objectif de transparence et création de valeur. La DG est très sensible à cette accessibilité des

données.
Pourquoi partager ses données : le Royaume-Uni et l'Australie sont très avancés car tout ce qui est financé sur fonds publics doit être accessible sur des entrepôts de données. Cela augmente l'impact et la visibilité de la recherche.
Les différentes échelles de la qualité des données ouvertes : le premier niveau est le fichier PDF, puis un tableur, puis sous forme de format non propriétaire (cvs, txt), puis utiliser des URI pour pointer les données, jusqu'à la mise en lien des données.
L'Open Science est une science collaborative et ouverte à des scientifiques et des utilisateurs de tous les pays dans le cadre d'une science “non concurrentielle”.
Quelle est la volonté de l'INRA de participer à ce projet ? ProdInra2.0 des archives ouvertes est la première réalisation. Il y a maintenant un chantier ouvert en 2012 qui doit proposer les grands principes pour la création des entrepôts. Les réponses ne seront pas uniques, cela dépendra des domaines et de la nature des données.
Au niveau européen, il y a le chantier IGLO dans le cadre de Horizon 2020.

L'INRA tiendra un Séminaire dans le cadre de la stratégie de valorisation des données sur 3 jours du 15 au 17 avril 2013.
Questions : la pratique de l'INRA est bien loin de ça et a plutôt tendance à « protéger » les données.
Patrice Buche : concernant la voie Gold des publications, les coûts et le modèle économique des éditeurs n'est pas encore stable et connu.
Isabelle Blanc : il faut faire l'ouverture progressivement.
Patrice Buche : même sur la partie visible des données, elles ne sont pas vraiment disponibles car il n'y a pas de format. Il faudrait que les données des publications soient environnées et sémantisées.
Il n'y a pas que l'aspect réutilisation des données mais aussi des calculs : il faut aussi sémantiser les calculs et algorithmes.
Marie : comment les données peuvent-elles être évaluées ? Isabelle : il faudra une commission interne d'approbation de la validité des données et aussi une acceptation du principe de partage de la données par les producteurs de la donnée.
Il y aura certainement de nouveaux métiers à créer avec une recomposition des métiers.

11h35. Présentation générale du Cati CODEX : Nadine Hilgert et Pascal Neveu

Le Cati Codex est un Cati « centré compétences ».
Effectif : 19 personnes (16 permanents + 3 non-permanents).
Départements : MIA, EA, CEPIA, AlimH, SPE, MICA, BAP
2 axes de compétences : informatique et math/statistiques.
Communauté bénéficiaire large : phénotypage haut débit, filière, bioprocédés avec comme point commun la donnée expérimentale qui évolue avec le temps.
Mode d’interaction avec la communauté bénéficiaire : développement de logiciels, publications et formations.

La production du Cati doit être articulée avec les Catis « connexes » au travers des correspondants chargés d'informer et de s'informer sur les productions respectives.
Les interlocateurs/correspondants du Cati Codex vers les autres Catis :
IUMA : Nadine Hilgert
ICAT : Brigitte Charnomordic
CaSciSDi : Pascal Neveu
Diisco : Eric Latrille
Sioea : Vincent Nègre
CGI : Yann Serrand
Sicpa : Pascal Neveu
Autres Cati : ??

Le rôle du Cati est aussi le positionnement, les trajectoires et les formations au profit des membres du Cati.

Quelques caractéristiques du Cati :

- à 70% Montpelliérain
- Point fort : analyser et gérer les données.
- Diffusion élargie au-delà des unités.

Financements : 200 €/pers financés par les départements pour les déplacements à l'AG annuelle.
Il faudra aussi compter sur un soutien des unités pour les autres déplacements. Une possibilité est aussi d'inclure directement l'UMR Mistea et le Cati Codex dans les nouveaux projets d'intérêt collectif afin de pouvoir attribuer une ressource au Cati.
Pascal et Nadine se proposent d'offrir leur appui auprès des DU des membres de Cati lors des entretiens annuels.

Tour de table : les membres du Cati Codex

Anne Tireau : Dept MIA, MISTEA Montpellier, développement informatique.
Virginie Rossard : Dept EA, LBE Narbonne, développement d'applications.
Yann Serrand : Dept BAP, Versailles, développement d'applications en phénotypage.
Roger Boll : Dept SPE, Sophia-Antipolis, entrepôt de données et interface.
Marie Weiss : Dept EA, EMMAH Avignon, caractérisation de couvert végétal.
Philippe Abbal : SPO, dept CEPIA, IR et doctorant sur la gestion des connaissances.
Aurélie Thébault : Dept MIA, MISTEA Montpellier, CDD, statisticienne filière viniticole
Jonathan Mineau : Dept EA, LEPSE Montpellier, automates de plate-forme de phénotypage.
Emilie Genari : Dept MIA, MISTEA Montpellier, CDD, développement d'applications.
Alexandre Mairin: Dept MIA, MISTEA Montpellier, CDD, développement d'applications.
Martine Marco : MISTEA SupAgro, développement d'applications d'intérêt générale.
Brigitte Charnomordic: Dept MIA, MISTEA Montpellier, outils d'aide à la décision. .
Isabelle Sanchez. Dept MICA, SPO Montpellier, statisticienne, microbiologie, levure en œnologie.
Thomas Moyon : Dept AlimH, PhAN Nantes, étude protéomique et métabolomique.
Vincent Nègre. : Dept EA, LEPSE Montpellier, développement d'applications.
Luc Verdier : Dept CEPIA, SPO, données de spectrométrie de masse.
Nadine Hilgert : Dept MIA, MISTEA Montpellier, statisticienne.
Pascal Neveu : Dept MIA, MISTEA Montpellier, responsable du Cati Codex.
Eric Latrille : Dept EA, LBE Narbonne.

Patrice Buche : Dept CEPIA, IATE Montpellier, cati ICAT.
Hervé Richard : Dept MIA, Avignon, Responsable Cati Cascisdi + mission informatique du département.
Isabelle Blanc : unité Supsis , support informatique auprès des unités. En lien avec tous les Catis.
Nicolas Donès : Dept EA, responsable IUMA
Cyril Pommier : Dept BAP, URGI

14h10 Patrice Buche : Cati ICAT (méthodologie).

Le Cati ICAT a 2 objectifs :

- Capitalisation et modélisation de connaissances et de données textuelles.
- Exploitation, analyse et traitement de ces données.

Il regroupe des agents issus de 7 unités : MIG, METARisk, IATE, I2M, IST.
7 CR, 15 ingénieurs, 3 doctorants, 1 post-doc.

A l'unité MIG, application ALVIS : extraction de terminologie et construction d'une ontologie basée sur l'habitat et les bactéries. Annotation semi-automatique
Aux unités IATE et MetaRisk : extraction de tableaux contenus dans des articles stockés dans Mendeley.
Aux unités Sens et Ecodéveloppement : cartographie de partenariat de projets basés sur la co-occurence de termes contenus dans les résumés.
A I2M (Bordeaux) : l'objectif est de construire des livres de connaissance. Exemple sur le pétrissage de pain.
Aux unités MISTEA et IATE : sur l'exemple d'une opération unitaire de cuisson des pâtes, construction d'arbres de décision basés sur des connaissances. L'objectif est de pouvoir faire des expérimentations virtuelles sans modèle et avec uniquement des données.

Quelles productions peuvent être partagées ?

- des retours d'expérience sur l'OpenData et les repository RDF.
- Des retours d'expérience sur les éditeurs d'ontologie.
- Distribution et transfert de plateforme (ALVIS) après 10 ans de développement à l'INRA. Il y aura un déploiement sur un serveur de Montpellier.
- Chantier cartographie/recensement des ontologies disponibles au sein du Cati. Projet d'ouverture d'un portail INRA regroupant l'ensemble des ontologies à construire entre des Catis.
- Les 2 premiers points sont plutôt de la compétence des PEPI.

14h40 : Nicolas Donès. Cati IUMA : objet : modèles d'agroécosystèmes.

Objectif de développer des SI en soutien au développement des modèles.
5 plateformes sont visées : OpenAlea (INRIA/CIRAD) plante virtuelle, Sol virtuel, RECORD (cultures/élevages), Capsis (peuplement forestier), paysage virtuel.
On regarde surtout l'interopérabilité des plate-formes.
62 membres dans ce Cati répartis en 4 pôles (48 Bap E + 14 CR).
Quelques pistes de liens avec Codex : outils collaboratifs (forge), calcul scientifique lié à la modélisation (analyse de sensibilité, estimation de paramètres, optimisation de code), liens plates-formes/modèles avec bases de données.

15h00 : Hervé Richard. Cati CaSciSDi (centré compétences) : calcul scientifique, statistique et informatique.

31 personnes réparties dans 16 unités. Uniquement des agents Bap E, les chercheurs ne sont pas intégrés directement.
Développement de bibliothèques, notamment en analyse d'images.
Recouvrement important avec le PEPI-MACS.
Le Cati peut avoir un rôle d'expertise au cours du montage de projets type ANR.
C'est un Cati qui part de loin et qui reste modeste.
IUMA et CaSciSDi propose de mettre en place des déclarations d'intention individuelles en accord avec les DU.
Besoin d'outils collaboratifs tels que la Forge.
Créer des fiches techniques sur les fonctions développées sous R.
Retour d'expérience sur les développements sous R.

15h25 : Cyril Pommier. Cati CGI en génomique info.

Pérennisation de données génomiques de plantes et champignons pathogènes et symbiotes et données phénomiques.

Chaînes de traitement et clusters de calcul. Conseil dans le montage de projets.
Stockage délocalisé des données primaires et intégration dans les laboratoires.
44 personnes réparties dans 13 unités.
Impliqué dans le projet Phénome et les projets bioressources.
Participation dans des groupements internationaux d'ontologie (Crop ontologie, …).
Lien avec Codex :

- système d'information
- intégration et échange de données :
  - ontologies pour bien identifier les objets biologiques
  - identification de génotype/accessions.
- Format de données accessibles (humain et machine).
- Réseau , Web service (API).

16h00-18h00 : réunions en 2 groupes : un informatique, l'autre math/stat.

Les participants à l'axe math/stat : Isabelle S., Marie, Thomas, Roger, Philippe, Nadine, Eric. Chacun avait préparé une présentation de 15 minutes de ses réalisations, première étape à la conception et à la réalisation d'applications communes.

Les participants à l'axe info : Anne, Martine, Pascal, Brigitte, Virginie, Luc, Yann, Jonathan, Vincent. Il y a eu quelques exposés du groupe Silex et un exposé sur Phenoscope. A noter qu'une majorité des informaticiens du groupe sont déjà membres du meta-projet Silex, sauf Yann qui va l'intégrer.

Fin de journée : Visite des plateformes de phénotypage végétal du LEPSE et diner au restaurant.

VENDREDI 15 février

9h00-10h00 Pascal Neveu : le Web sémantique

Le Web sémantique a pour objectif de faire comprendre aux machines des données et des connaissances formalisation de concepts et des relations entre ces concepts. (sous forme de mots-clés).
Il y a 4 langages différents : RDF (métadonnées), RDFS (créer des vocabulaires), OWL (couche logique) et SPARQL (interroger un graphe RDF).

Exemple de dépôt de fonctions R

C'est une nouvelle façon de stocker les fonctions R qui sont décrites et organisée avec de la sémantique. Le dépôt intègre un générateur de requêtes SPARQL qui permet aux utilisateurs de disposer de raisonnements pour rechercher des fonctions R
Démonstration de l'outil DESIRR par Anne pour le dépôt sémantique de fonctions R.
Pascal propose que cet outil soit utilisé dans le cadre du Cati Codex.

10h00-10h30 Vincent Nègre : Site Web Codex

Cati-codex-bureau@listes.inra.fr
Cati-codex@listes.inra.fr
https://listes.inra.fr pour la consultation des archives.
Site Web : trouver un groupe pour définir les contenus du site.

Forge logicielle : pour mettre en ligne les sources des réalisations.

10h30 Anne Tireau. SILEX

Système d'Information pour L'EXpérimentation.
SPO, LEPSE, LBE, MISTEA : actuellement 7 informaticiens.
Les différents modules :

Mesures en ligne, mesures hors ligne, mesures at-line.
Modèles de bases de données.
Gestion des connaissances.
Applications Web : graphiques,
Gestion des utilisateurs.
Graphiques avec HighCharts.

Proposition d'adapter le module « at-line » pour les données d'images de Marie.
Discussions sur les difficultés d'utilisation du module « hors-ligne ».

10h50 Nadine Hilgert : projet PHENOME

Projet d'investissement d'avenir sur 8 ans.
Objectif : Fournir à la communauté végétale française une infrastructure de phénotypage haut-débit (public, privé, instituts techniques).
Etude du comportement des plantes face aux changements climatiques.
3 Projets méthodologiques :

- capteurs et méthodes (architecture, phénomobile, nouvelles techniques d'imagerie, caractérisation de l'environnement).
- Systèmes d'information distribuée (ontologie, base de données, webservice, archivage).
- (Pré-)traitement des données (qualité, analyse des données « courbes », interprétation).

11h10 Pascal Neveu : Méthodes agiles pour la recherche.

Dans le développement logiciel, la vraie difficulté est dans la relation avec les utilisateurs.
La première difficulté est de bien faire exprimer aux utilisateurs leurs besoins.
Méthode Agile : développement itératif avec un pas de 1 mois environ.
4 valeurs :

- Simplicité : rechercher la simplicité en dialoguant avec l'utilisateur. Faire simple mais ne pas simplifier. Trouver le bon compromis entre générique et spécifique. Le générique complexifie souvent trop.
- Feedback : le pas d'itération de 3 semaines à 2 mois. Documentation légère mais surtout à jour. Faire des tests individuels.
- Communication : priorité aux personnes et aux interactions. Risque de mal ou trop solliciter les utilisateurs. Conception collective, responsabilité collective du code. Développement en binôme.
- Courage : s'engager, proposer et être capable de changer. Parfois repartir à zéro ! Ne pas repousser une rencontre ou une livraison. Nécessite une disponibilité.
-

Conclusion : les codes sont de meilleure qualité et utilisés. C'est une méthode qui dépend clairement de la dimension humaine.

11h40 Clôture restitution

En visioconférence avec Frédéric Garcia (MIA).

Nous allons travailler sur l'organisation en 2 axes Info et Math/Stat.
Nous avons tenu le programme et l'intervention d'Isabelle Blanc a été appréciée. Les chiffres de Laurent Bruckler ont montré que les informaticiens étaient la catégorie la plus en croissance à l'INRA.
Frédéric Garcia indique qu'il a beaucoup défendu la création du Cati Codex en montrant que nous étions une communauté qui voyait où elle allait. Le Cati Codex commence avec un projet qui est bien vu. Faire tout de même attention à ne pas devenir le Cati de « toutes » les données à l'INRA.
Nous allons commencer humblement en sachant que nous sommes attendus dans le projet PHENOME.

Positionnement des membres suivant les 2 pôles.

Prochaines étapes :

- Pour le pôle math/stat :
  - réunion par visio dans moins de 2 mois entre Nantes-Montpellier-Nice
  - Marie peut commencer à voir ce qu'elle pourrait faire avec le site de dépôt R, DESIRR.
  - Roger peut suivre de plus près le groupe de développement autour de SILEX et voir aussi les conditions de dépôt APP du Data_Market.
  - Isabelle et Thomas ont des méthodes et des développements sous R à partager. Utiliser DESIRR semble très pertinent.
- Pour le pôle informatique : Silex est le cadre pour fédérer notre production