3 et 4 décembre 2014 – Montpellier

compte-rendu

Journées communes des CATI Codex et Sicpa

Participants CATI SICPA: Laperruque François, Journaux Alexandre, Reichstad Matthieu, Heirman Thierry, Labrune Yann, Furstoss Vincent, Robert Pierre-Emmanuel, Lagant Hervé, Bompa Jean-François, Pierre Chalier, Urban Bernadette, Ricard Edmond, Battut Marie-Christine

Participants CATI CODEX: Negre Vincent, Hilgert Nadine, Philippe Abbal, Boulet Jean-Claude, Mineau Jonathan, Tireau Anne, Latrille Eric,Schorgen Antoine , Neveu Pascal

- 14h40 Interconnexion de bases de données et webservices : Thierry Heirman (CATI SICPA)

Projet GAniMed : 7 systèmes d'information : pour les faire communiquer on passe par une plate-forme commune (intergiciel) qui s'occupe du routage des données. C'est une architecture SOA (Service Oriented Architecture) en interopérabilité au fil de l'eau (on clique et on récupère les données). ESB : Enterprise Service Bus basé sur XML pour le format des données et JMS (Java message service) pour le transport. Plusieurs solutions ont été testées (Apache service mix, JBoss, Petals ESB). Petals ESB est le produit retenu (openSource avec support professionnel). Déploiement de webservice en SOAP uniquement avec un serveur GlassFish. On peut aussi répartir la charge sur plusieurs services. Développement en JAVA.

En REST, on utilise les fonctions put, get de http alors qu'en SOAP il faut définir une enveloppe d'entête, de corps et de signature (un peu plus complexe).

- 15h15 : Modélisation systémique et analyse multi-échelles à l'unité MoSAR : savoir-faire et perspectives : Pierre-Emmanuel Robert (CATI SICPA)

Modélisation systémique qui considère les systèmes dans leur ensemble. Téléologie : on s'intéresse plutôt aux résultats qu'à décrire le fonctionnement. Approche systémique de Daniel Sauvant de la nutrition animale. MoSAR : modélisation systémique appliquée aux ruminants. Trois exemples : la nutrition avec plusieurs compartiments (les réserves, la reproduction, la lactation, …). Approche multi-échelles : phénotypage fonctionnel. Approche par les variables latentes (modèles de mesure indirecte). Les chèvres sont pesées deux fois par jour. Base de données PostGreSQL. Le troupeau est aussi considéré comme un système en soit.

- 15h55 : Introduction au web sémantique Pascal Neveu (CATI CODEX)

Il faut 5 étapes pour rendre accessibles les données sur le web.

Le web sémantique change nos façons de rechercher de l'information, faire des achats, échanger entre personnes. Les limites sont la sensibilité au vocabulaire, la précision et la spécialisation. Il faudrait intégrer des données de sources différentes mais avec des noms différents. Mais à un moment, il faut changer de méthodes. Le web sémantique est un framework qui permet de nommer les choses. Les URI sont des identifiants unique. RDF resource description framework (sujet, predicat, objet) : #plante12 #isLocated #parcelle5. Ensuite #parcelle5 #variety #syrah. L'élément de base est le triplet. Mais, il faut une couche supérieure de vocabulaire (sémantique) qui est le RDF Schema (RDFS). Au-dessus, il faut organiser les connaissances, c'est en OWL. Puis, il faut interroger avec SPARQL, c'est un peu comme du SQL. On peut faire des inférences, trouver des concepts voisins …

Quels sont les liens avec Twitter et la gestion des hTag et avec le projet Wolfram alpha ? Facebook utilise effectivement les ontologies RDF.

- 17h00 : DESIRR, une bibliothéque de partage de scripts R : Anne Tireau (CATI CODEX)

Définition d'une ontologie pour définir le contexte avec un vocabulaire contrôlé. Cela permet de faire le graphe des appels des fonctions. PHP, versioning et corese.

- 17h30 : Validation des données issues de phénotypage végétal : Antoine Schorgen (CATI CODEX)

Nettoyage des données avec des filtres. Détection des valeurs aberrantes et remplacer les données manquantes. Utilisation de ggplot2 et du lissage par un modèle logistique. Les données sont-elles remplacées ou bien de nouvelles variables sont-elles créées ? Intervalle de confiance à 99%.

- 9h10 : Créer ses packages avec Rstudio : Vincent Nègre (CATI CODEX)

IDE de développement de R. Créer un projet de type R-Package et sélectionner les scripts. R crée d'arborescence du package. Puis, vérifier Build-check le package. Pour déposer sur le CRAN, il y a des règles supplémentaires (un nom unique, un auteur, les termes de la licence). Installer Roxygen2 pour construire la documentation. Le seul travail à faire est de documenter les scripts. Mettre un @ pour définir les mots clés. Un package peut être en source ou en binaire afin de protéger les scripts.

Gestionnaire de versions installé sur un serveur. Rstudio supporte git et subversion et SVN. Création d'un projet sur MULCYBER (INRA, MIA). Un onglet svn est créé dans R pour gérer les versions. Puis ajout et commit d'un nouveau script.

- 9h45 : Utilisation d'une application de classification et d'analyse multivariée (PLS) sous R-studio installé sur un serveur et connecté à une base de données : EricLatrille (CATI CODEX)

  Petite présentation et démonstration à partir du serveur Foix.

- 10h10: knitr : création de documents dynamiques avec R : Yann Labrune (CATI SICPA)

Création de fichier de sorties pdf ou html en utilisant les markdown de Rstudio. Knitr est plus facile à utiliser que Seawe. On insère des chunks qui sont les balises de markdown. Cette syntaxe est celle utilisée par les wikis.

- 10h45 : le projet ATOL, référentiel pour le phénotypage des animaux d'élevage : Matthieu Reichstadt (CATI SICPA) – Animal Trait Ontology for Livestock

Ontologie : représentation formelle d'un ensemble de concepts et des relations entre ces concepts dans une discipline spécifique. L'ontologie la plus connue en bioogie est celle de geneOntology. Objectif : sélection animale. EOL : ontology pour les environnements d'élevages. www.atol-ontology.com

Il faut surtout faire un effort pour lister l'ensemble des synonymes de chaque trait. Contruction des owl avec webProtégé (développé en Java). Des développements ont été rajoutés à webProtégé (spécifier si une espèce possède un trait). Les ontologies ATOL et EOL serviront à annoter des données de phénotypage (publications).

Tout est publié sur le site BioPortal qui regroupe de très nombreuses ontologies.

- 11h15 : PHIS : Phenotyping Hybrid Information System : Jonathan Mineau (CATI CODEX)

Sur PhenoArch on s'approche du big data en open data. 600 000 données par an. PHIS est composé de bases de données relationnelles qui stockent les données offLine (16 tables). Dans la table Objet, il y a l'échantillonnage et la localisation. C'est l'équivalent de nos tapPoints.

Les données Online seront stockées dans une base NoSQL. Choix du NoSQL : Not only SQL pour gérer des gros volumes de données. MongoDB orienté Documents. Les données OnLine sont gérés par le NoSQL car le volume journalier est très gros et la donnée est très standardisée. On gère des liens et des métadonnées vers des fichiers images et video.

Les objets sont stockés dans une base de données sémantique (triple store).

A partir de 100 millions d'enregistrements dans une base de données relationnelle, on commence à avoir des problèmes sur les serveurs.

- **Discussions** :

- Les perspectives à mettre dans les rapports d'activités de nos 2 CATIs pourraient être réfléchies ensemble.

- Pour 2015, nous prévoyons de renouveller les rencontres entre le deux Catis.

Plusieurs thématiques se sont dégagées :

- statistiques;
- PHP;
- Web services;
- développement mobile.

Les objectifs à atteindre sont à préciser : veille technologique; formation; partage de code; projet commun.

- Les stats en invitant des scientifiques sachant que SICPA utilisent surtout SAS et Codex plutôt R.
- Les développements Web, y compris les applications mobiles, et autour d’un outil commun php. Comment pourrait-on avoir des interfaces homogènes ?
- Les méthodes d’annotation de courbes : Comment donner un retour aux utilisateurs pour qu’ils aient un attrait à saisir des données d’observaStion ?