INRAE

Cati Sicpa
Systèmes d'Informations et Calcul pour le Phénotypage Animal

Etude d'une solution d'interopérabilité pour l'exploitation et l'analyse de données hétérogènes animales

Données utilisées au sein de l'UMR GenPhySE

UMR GenPhySE

Présentation de l'unité

- Génétique, Physiologie et Systèmes d'Elevage
  - Rattachée aux départements GA (Génétique Animale) et PHASE (Physiologie Animale et Systèmes d'Elevage)
  - Recherches : Du gène au phénotype et au système d'élevage
  - Espèces étudiées : ovins, caprins, lapins, porcins, abeilles, cailles, (palmipèdes)

UMR GenPhySE

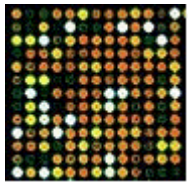
Présentation de l'unité

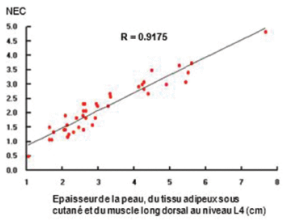
- Génétique, Physiologie et Systèmes d'Elevage
  - Objectifs
    - Améliorer les connaissances sur la structure et l'organisation fonctionnelle des génomes
    - Explorer la variabilité génétique des caractères complexes chez les animaux d'élevage
    - Comprendre les mécanismes biologiques dans l'élaboration des phénotypes
    - Comprendre et modéliser les interactions
    - Améliorer les populations animales par la sélection génomique et la conception de programme de sélection
    - Comprendre les effets environnementaux sur les phénotypes
    - Concevoir des systèmes de production animale plus durables

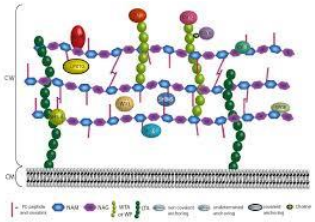
UMR GenPhySE

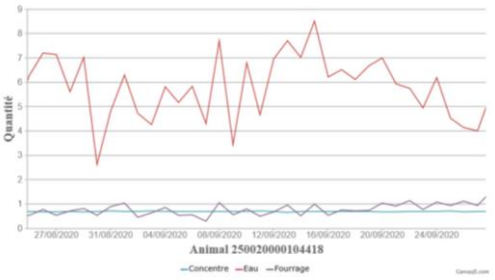
Les données étudiées

- Des données très hétérogènes
  - Types de données
    - Données génétiques : génotypage (mutations ponctuelles, puces de SNP)
    - Données de séquence : séquençage d'ADN et d'ARN, microbiotes, métagénome
    - Données biochimiques : dosages hormonaux, activités enzymatiques
    - Données spectrales : MIR, NIR…
    - Imagerie : caryotype, imagerie cellulaire ou d'organoïdes en 3D
    - Données phénotypiques : pesées, consommation, qualité lait/viande, …
    - Données environnementales : températures, hygrométries, gaz, …
    - Données spatiales : géolocalisation de l'animal, posture de l'animal, …





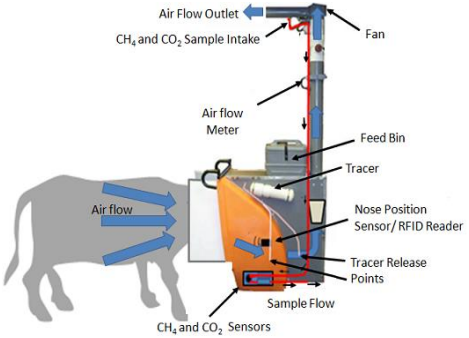




UMR GenPhySE

Les données étudiées

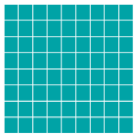
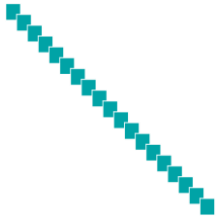
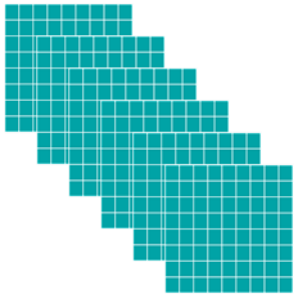
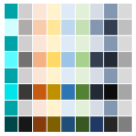
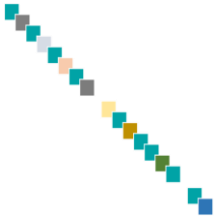
- Des données très hétérogènes
  - Modes de collecte
    - Observations humaines : naissances, maladies, notations, comportements, …
    - Prélèvements ponctuels : sangs, fèces, rumens, tissus, lait, …
    - Enregistrements continus : DAC, capteurs, accéléromètres, surveillance vidéo, radars, …



UMR GenPhySE

Les données étudiées

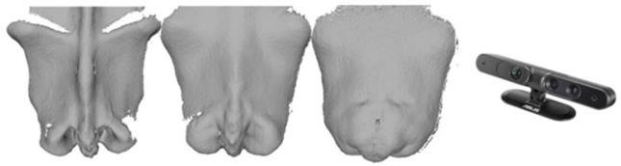
- Des données très hétérogènes
  - Fréquences et répétitions

	Ponctuellement	Longitudinales	Répétées ou continues
Homogènes			
Hétérogènes			

UMR GenPhySE

Les données étudiées

- Des données très hétérogènes
  - Format
    - Base de données relationnelles
    - Fichiers plats : csv, xls, …
    - Fichiers binaires : bam, …
    - Photos, images 3D, vidéos
  - Stockage
    - Base de données MySQL, Oracle, PostgreSQL
    - PC ou disques durs perso
    - Serveurs de l'unité
    - Différents entrepôts publics



Le projet d'interopérabilité

Les cas d'usage

- Cas d'usage 1a : à partir du numéro animal, connaitre l'existence de la donnée
  - Savoir que des données existent déjà sur un animal et quels types de données
    - Pour élargir mon champ de recherche
    - Pour éviter de génotyper plusieurs fois le même animal
- Cas d'usage 1b : … et savoir où se trouve la donnée
  - Avoir l'information du lieu de stockage de la donnée
    - Existe-t-elle toujours ?
    - Est-elle accessible ? Sous quelle condition ?
- Cas d'usage 2 : récupération de la donnée brute et/ou pré-traitée
  - Données brutes ou données déjà agrégées
  - Accès aux publications

Le projet d'interopérabilité

Les cas d'usage

- Autres cas d'usage
  - Accès aux données à partir
    - D'une race particulière
    - D'un lot d'animaux
    - D'un stade physiologique particulier
    - D'une maladie
    - D'un comportement
    - …

Etude d'une solution

Les 3 V du Big Data

- Volume
  - Plusieurs To répartis sur différents serveurs, entrepôts, bases de données
- Vélocité
  - Acquisition de données par automates, capteurs, caméras, …
- Variété
  - Hétérogénéité des types de données et des formats

Etude d'une solution

Des solutions autour des technologies Big Data

- Data Lake
  - Stocker les données brutes, agrégées ou transformées
  - Durée indéterminée
  - Permet la cohabitation entre différents types de données
- Bases de données NoSQL
  - Cassandra, MongoDB, InfluxDB, …
- Accéder, calculer, …
  - Spark, MapReduce, …
  - Utilisation d'ontologies : ATOL, EOL, AHOL, …

Etude d'une solution

Nos points forts

- Donnée centrale
  - L'animal
- Compétences internes
  - T. Heirman a validé une formation diplômante de Data Architecte
  - Stagiaire 5e année INSA, option « Systèmes Distribués et Big Data »
- Des partenaires
  - Cati Codex, collègues ex-Irstea, GenoToul (plateforme Bio-info), LIPM
- Des premières expériences
  - Stockage données DAC : Cassandra, Spark
  - Enregistrement données capteurs : InfluxDB
  - Interopérabilité : Web service Java
- Des scientifiques motivés et organisés
  - Groupe Qualité : préconisation sur les métadonnées
  - Groupe de réflexion sur le sujet depuis 2018

Etude d'une solution

Les difficultés

- Définition du besoin
  - Réflexion chronophage pour mieux caractériser les données
  - Pour chaque type de données : établir les métadonnées à associer
  - Discussions entre scientifiques et informaticiens
- Numéro Animal
  - Pas toujours le même en fonction de la source
- Interprétation des données
  - Domaine de compétences différents des scientifiques
    - Génétique, physiologie, approche systèmes
  - Peu de compétence Data Analyst, IA

Etude d'une solution

Perspective

- Un portail d'accès aux données
  - Un point d'entrée unique pour visualiser les données
  - Une application qui permette de requêter sur les données

- D'autres unités ont les même problématiques
  - Domaine animal : UMR Gabi, Pegase, Herbipôle, …
  - Domaine végétal surement aussi