



# Ingenum/DTN

**DataPartage**

22 novembre 2016

Vers une **e-Infrastructure** pour les données scientifiques

AgroDataRing

Christophe Caron *et al.*

# Plan

- Etude « e-infra storage »
- Enjeux technologiques et organisationnels
- Conclusion

# Un travail collectif !

- P. Moreau (EA), O. Langella (CNRS), T. Hotelier (SupAgro), A. Dehne-Garcia (SPE), O. Filangi (BAP), D. Tessier (CEPIA), T. Rabemanantsoa (EA), L. Legrand (SPE)
- G. Salin (GA), F. Giacomoni (AlimH), P. Adenot (PHASE), A. Hofstetter (SAE2)
- H. Quesneville (BAP), AF Adam-Blondon (BAP), P. Neveu (MIA), V. Loux (MIA), K. Hugot (CNOC)
- Ingenium/DTN



Besoins

# Etude

- **Objectif**

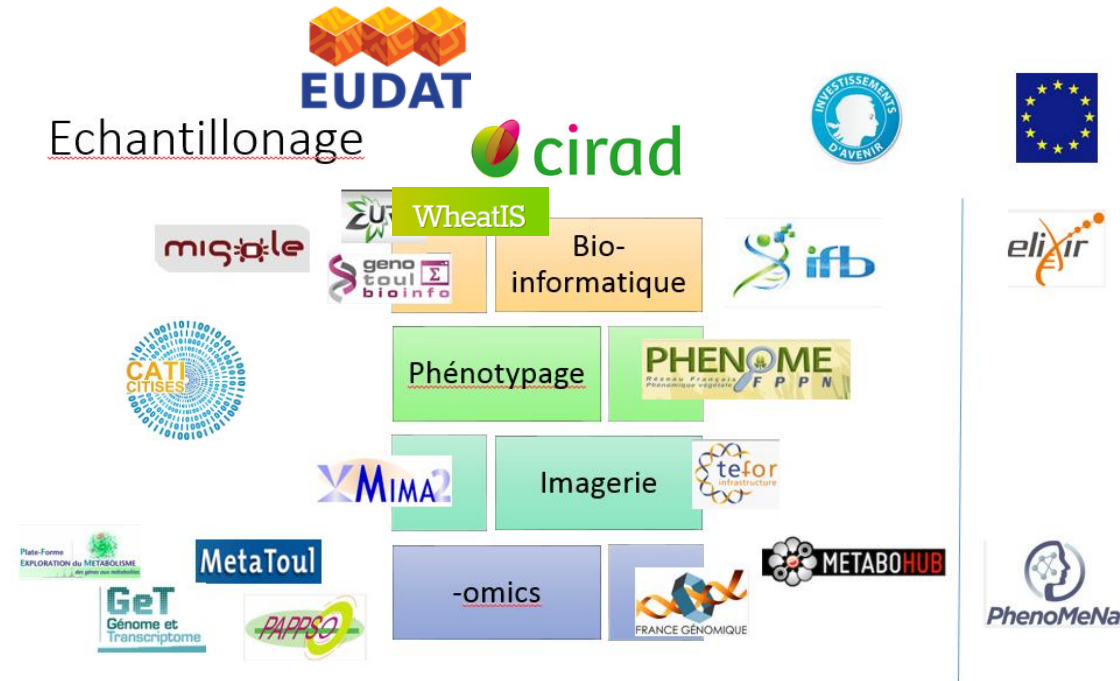
- Evaluer les besoins et élaborer des scénarios

- **Constats et contexte**

- Stockage adapté à la **production en masse de données**
- **Complémentarité** offre de stockage DSI
- Des infrastructures DC et salles de proximité
- **Financement** de la recherche par projet

- Comment approcher la réalité en étant pragmatique ?

- **Échantillonnage** et croisement avec les éléments de contexte



# Points saillants



- ~ Centaines To /an
- Diversité des solutions actuelles de stockage

Volumétrie maîtrisable



- **Maturité** des scientifiques sur la capacité à qualifier la donnée : Conservation vs. « jetable »
- **mois / années / décennies**

Pérennité



CONTRAT DE PLAN  
Etat - Région



- Assurer la récurrence des **financements** ?
- « **Facilité** » à **obtenir des financements équipements vs. fonctionnement**

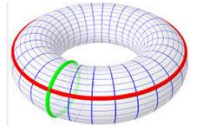
Financement



- Evolutions très rapides des technologies (acquisition/stockage)
- **Passage à l'échelle face aux ruptures technos ?**

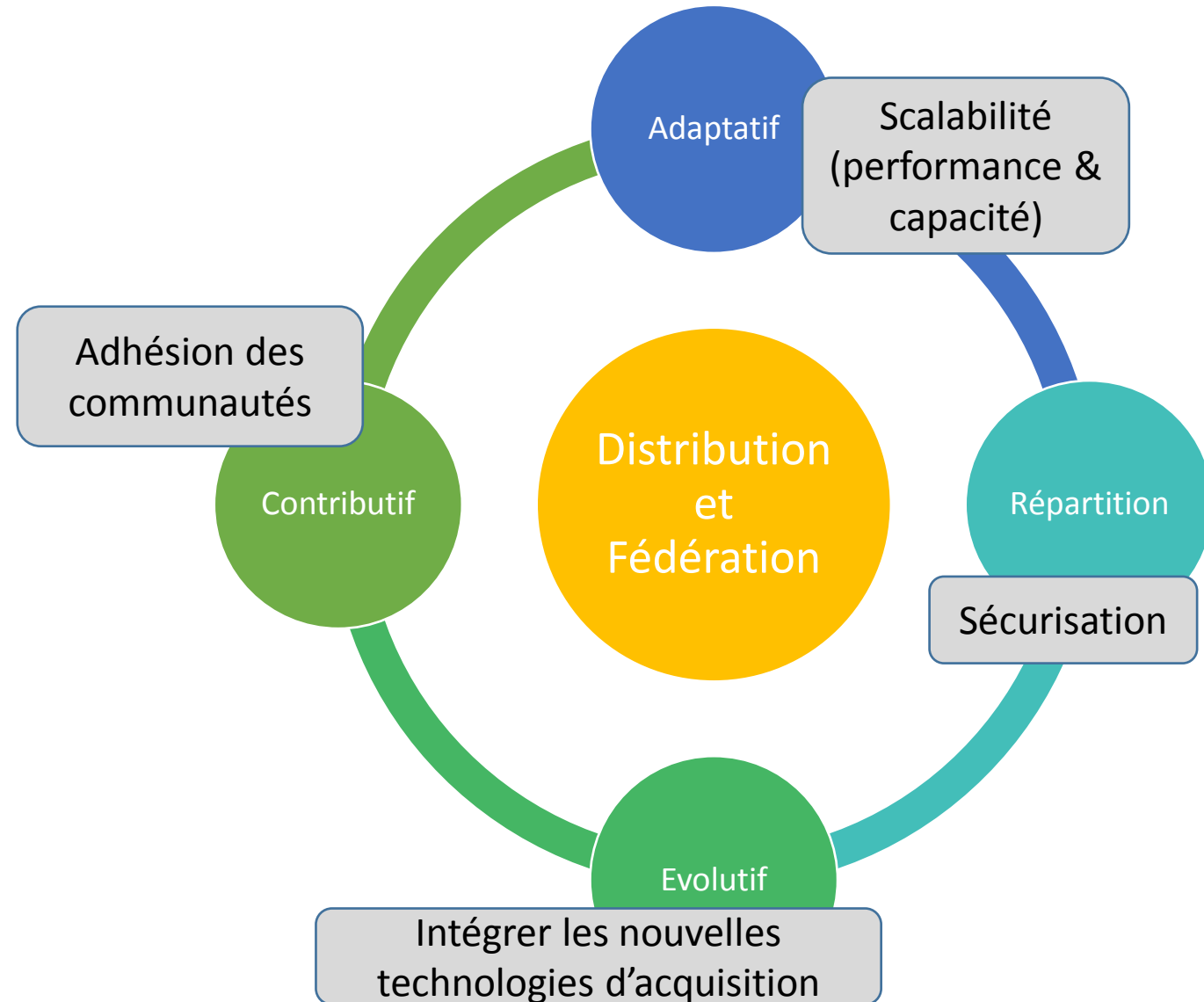
Réactivité/Souplesse

# Enjeux technologiques



# Vers un modèle distribué ?

- Utilisation d'un stockage « banalisé »
- Avec un logiciel de gestion dédié
- Lien naturel : distribution des sites INRA
- PoC Ingenium/DTN : Environnement stockage distribué
  - **Gestionnaire : iRODS**
  - *P. Moreau*
  - *Aspect meta-données*





# Une illustration INRA : Phénome

- Besoin : Stockage sécurisé **distribué** sur plusieurs sites
  - T- 12 : Tour d'horizon des types d'offres : DSI, France Grilles, etc.
  - T : Acquisition des équipements
  - T+3 : Déploiement et test
  - T+6 : Ouverture à la communauté scientifique

720 To utiles

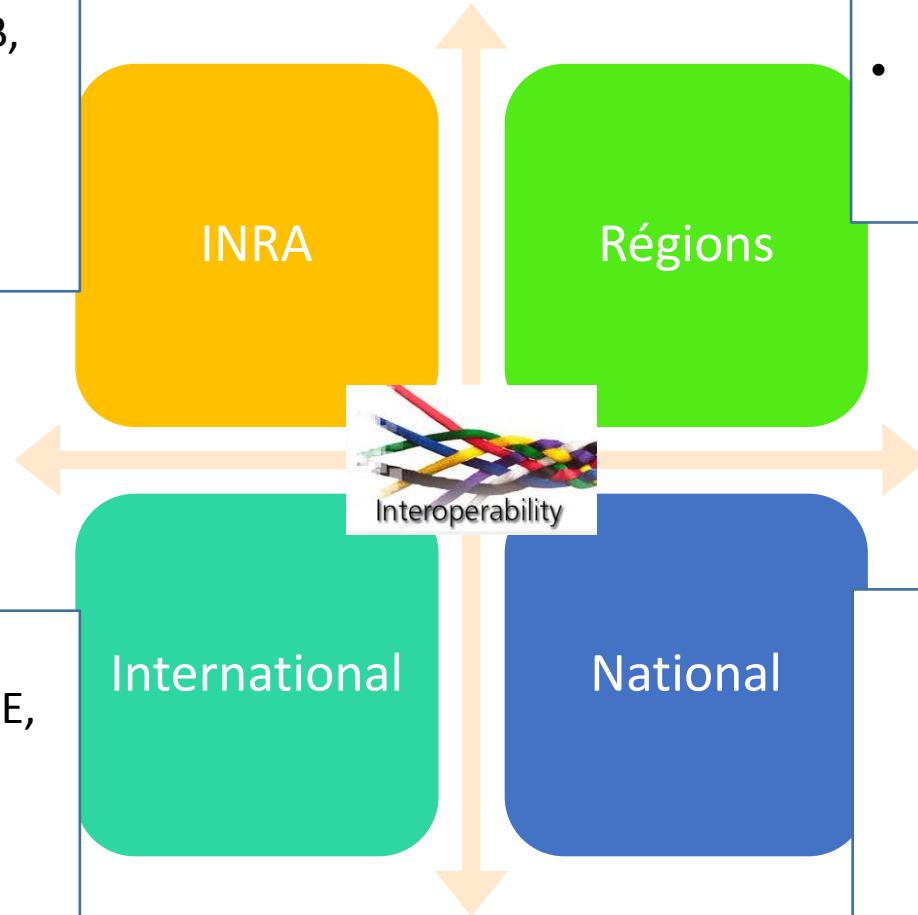
Coût **environné** (incluant les fluides)

16 k€/an = **23 €/To/an** (46 €/To/an avec réplication)

Enjeux organisationnels

# Organisation : tendances

- Grandes infrastructures de production de données : Phenome, MetaboHUB, France Génomique...
- CRBs, Plateformes ...
- Equipes



- BioDataRing



- wheatIS, IFB/ELIXIR, Emphasis, ANAEE, etc.
- NSF : OSiRIS (Open Storage Research InfraStructure)



# Conclusion

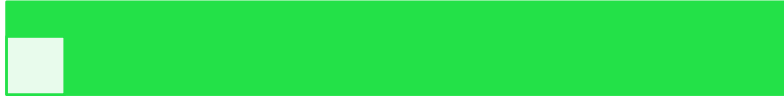
Il n'existe pas de solution unique pour le stockage ... mais les solutions distribuées émergent

## Atouts



- ✓ **Souple et adaptatif** : Big Data, financement équipement, etc.
- ✓ Allier capacitif et performances
- ✓ Coût : 20 € / an / To

## Opportunités



- ✓ Meta-données : **catalyseur** potentiel
- ✓ Compétences en interne : expertise, favorise l'adhésion
- ✓ **CATI 3G** ?!

## Points de vigilance



- ! Data post projet ? Une e-infra institutionnelle pour le stockage ?
- ! Gestion de l'anneau !

**Une e-Infrastructure stockage « INRA » 2 Po pour 300 k€ (60 k€/an) ?**