Inserm

La science pour la santé
From science to health

Problématiques de stockage à l'Inserm

Gilles MATHIEU

Inserm DSI/SSDUN, domaine Informatique Scientifique

21/01/2020 – DOMA-FR meeting

Quelques mots sur l'Inserm

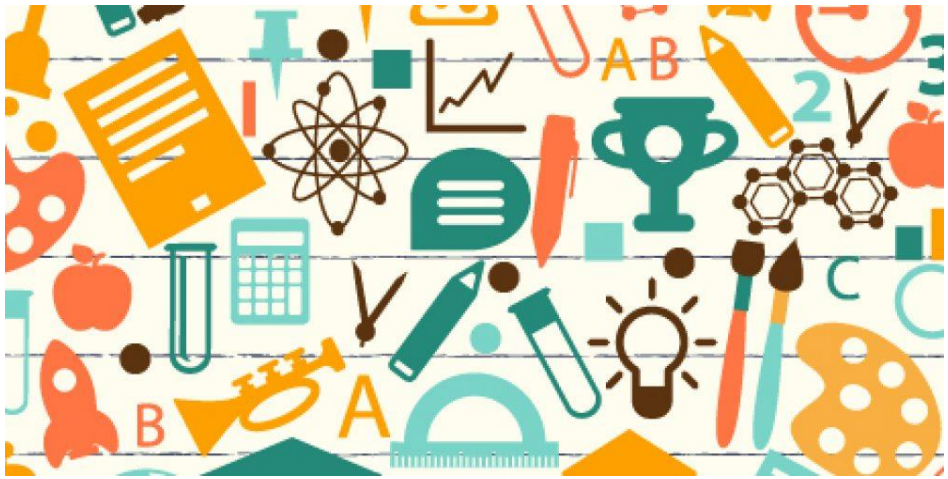
- Institut dédié à la recherche biologique, médicale et à la santé humaine

- Organisation et gouvernance
  - Cotutelle MESRI et Ministère de la Santé
  - 9 Instituts Thématiques (axes scientifiques)
  - 13 délégations régionales (axes administratifs)
  - Des départements transversaux, dont le DSI

- Structures de Recherche
  - 281 Unités de Recherche (presque toutes mixtes)
  - 36 Centres d'Investigation Cliniques
  - 34 Unités de Service

Les données à l'Inserm

- Une grande hétérogénéité
  - Des types de données
  - Des formats
  - Des volumétries
  - Des profils d'utilisation
  - Des solutions en place



Constat à l'échelle de l'Institut (1)

- **Stockage et volumétries**
  - 50%+ des unités ont entre 10 et 100To de données actives, et autant de données historiques
  - 68% des unités produisent moins de 10To/an
  - Evolution variable
    - de stable à fortement exponentielle.
    - Accroissement massif anticipé pour les données « omiques » (plusieurs Po/an)

- **Infrastructures**
  - Solutions locales fortement majoritaires
  - Centres régionaux : utilisation croissante
  - Centres nationaux et infrastructures nationales : utilisation encore marginale

Constat à l'échelle de l'Institut (2)

- **Données nécessitant une protection**
  - 50% des unités manipulent des données à caractère personnel
  - 75% des unités manipulent des données considérées comme « sensibles »
  - Ces données sont très majoritairement stockées en interne et/ou dans les CHU

- **Services hébergés/déployés par les unités**
  - 30% des unités hébergent un serveur web et/ou de bases de données
  - D'autres services déployés à moindre échelle (annuaires, outils, serveurs d'applications…)
  - Quelques systèmes de sauvegarde

Besoins remontés par nos unités de recherche

- **Besoins en termes de gestion de la donnée:**
  - Stockage classique
  - Stockage sécurisé
  - Sauvegarde et Archivage
  - Partage
  - Gestion de données ouvertes (publications/DOI)
- **Besoin en termes de traitement**
  - Capacité de calcul
  - Solutions de partage logiciel
  - Flexibilité de l'environnement
- **Besoins en termes d'accompagnement**
  - Support de proximité et de qualité
  - Formations techniques

Eléments de stratégie

- Objectif 4 du plan stratégique Inserm 2025
  - Mouvement global vers la science ouverte
  - Implique la mise en place des services associés
  - https://www.inserm.fr/connaitre-inserm/documents-strategiques

- Volonté d'une solution Inserm pour stocker/gérer les données de l'institut
  - Un « portail de la donnée » Inserm
  - La mise en place d'infrastructures sous-jacentes

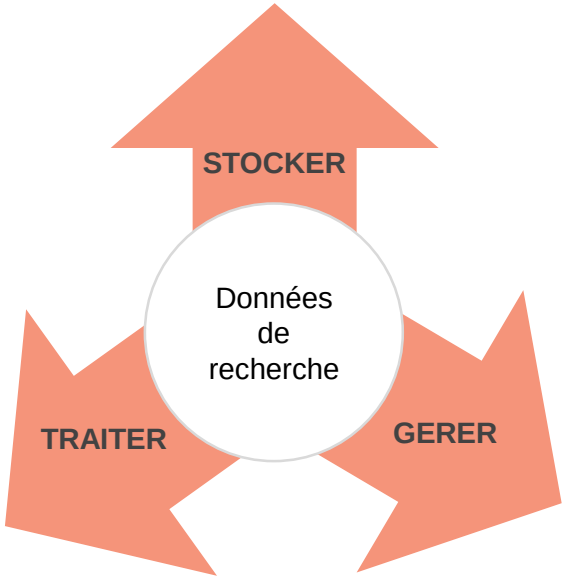
Réflexions et interrogations

- Challenges
  - Dimensionnement global difficile
  - Pas de solutions « *one size fits all* »

- Nécessité de s'interroger sur :
  - La popularité actuelle des solutions locales
  - Les difficultés rencontrées pour amener les unités à utiliser les solutions mutualisées
  - les coûts
  - les risques juridiques pour les données
  - Les enjeux du partage des données et de leur protection

Le projet SCaaS: Scientific Computing as a Service

- Objectif : définir, concevoir et mettre en place une offre de service « informatique scientifique » à l'Inserm
  - Complémentaire aux offres existantes
  - Mutualisée si possible
  - Ouverture progressive
- 3 axes de services centrés sur la donnée de recherche
  - Stockage
  - Gestion
  - Traitement



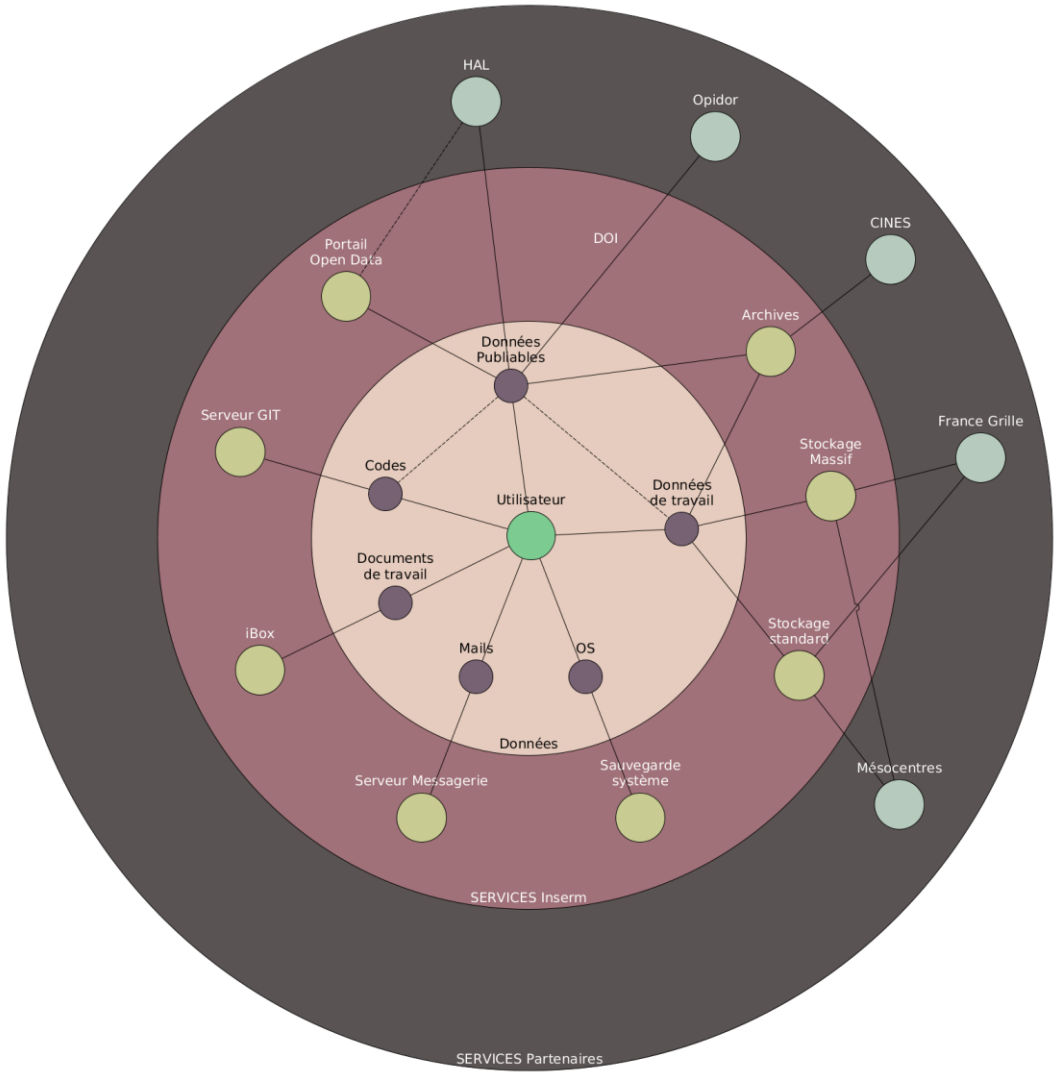
Stocker les données

- Stockage massif et pérenne (data store)
  - sécurisé ou non
- Stockage non pérenne (data scratch)
  - proche du calcul
- Partage de données et de documents
  - Un service existant : iBox
- Archivage long terme

Gérer les données

- Un portail de la donnée Inserm
  - Solution Dataverse envisagée
- Des solutions pour la publication de données
  - DOI, liaison avec HAL…
- Un service de gestion des DMP
  - En lien avec OpidOR
- Le Cahier de Laboratoire Electronique
  - Solution existante : CLE Inserm
- Un repository de partage de code
  - Solution déjà à l'étude au DSI

Stockage et gestion



Mise en route

- Calendrier global pour SCaaS
  - 2020 : Analyse, pré-étude, architecture et design, conception, prototypage
  - Déploiement de services cibles à horizon 2021

- Axes de travail associés
  - Réflexions conjointe au niveau régional, national, inter-EPST
  - Contribution à EOSC-Pillar
  - DOMA ?