De la production de la donnée à son exploitation scientifique

Défi 13

DÉFI 13 – CONTEXTE

PROBLÉMATIQUES ET ENJEUX

- **Comment gérer les données de la Science ?**

  Quelles informations et procédures doivent permettre de garantir, dans le temps, l'intégrité, la qualité et la véracité d'une donnée, ainsi que sa réutilisabilité ?

- **Comment utiliser les données de la science ?**

  Quels outils, quels flux et quelles infrastructures vont permettre aux chercheurs d'analyser leurs données, de les combiner des données existantes afin de produire de nouveaux résultats ?

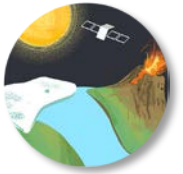
- **Comment répondre à ces questions dans un contexte d'avalanche de données ?**

  Quelles sont les nouvelles méthodes à mettre en place (IA, Machine Learning, …)

VERROUS SCIENTIFIQUES

- **Hétérogénéité des contextes/communautés au sein de l'INSU**
- **Limitations des ressources**
  - Humaines (effectifs, nouvelles compétences :data scientist, data manager)
  - Matérielles (conservation et circulation des données)
- **Besoins de communiquer au-delà de l'INSU: INS2i, INRIA…**
- **Problématiques liées aux données massives**
  - Traitements embarqués
  - Méthodes d'analyse génériques
  - Logistique
- **Manque d'outils:**
  - Outils d'analyse: comment rivaliser avec les GAFA
  - Outils de gestion intégrés au cycle de vie de la données
- **Verrous humains:**
  - Prise en compte tardive de la problématique de la donnée dans les projets scientifiques
  - Volonté de conserver toutes les données

DÉFI 13 – RECOMMANDATIONS



ORGANISATION

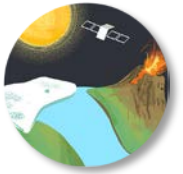
- **Gouvernance nationale des données : mise en place d'une « IR Donnée »**
  - Coordination technologique des mésocentres de données
  - Financement long terme de la conservation de la donnée
  - Expertise vis-à-vis des communautés scientifiques
  - Lien avec les pôles de données et observatoires virtuels.
- **Gouvernance locale des données: imposer sur l'ensemble des projets des DMP indiquant:**
  - La durée de conservation des différents produits
  - Les volumes générés
  - Les entrepôts de données associés au projet

  → Les coûts de conservations annuels.

INSTRUMENTATION, MODÈLES ET DONNÉES

- **Entrepôts de données interconnectés**
  - Les données doivent être regroupées dans des mésocentres aux capacités adaptées au contexte Big Data.
  - Ces entrepôts doivent prouver leurs capacités
  - Ces centres doivent mettre en place des flux de données privilégiés permettant un accès interopérable aux données
- **Rapprochement avec les centres de calcul**
  - Le lien avec les centres de calcul doit être anticipé via la mise en place d'infrastructures réseaux suffisantes
- **Logiciels communs**
  - L'IR Donnée doit piloter la mise en place d'outil communs dédiés à
    - L'analyse des données
    - La supervision des données stockées

DÉFI 13 – RECOMMANDATIONS



LIENS AVEC LES INDUSTRIES

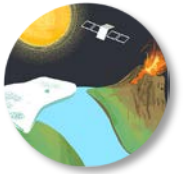
- **Les GAFA proposent des écosystèmes logiciels à la fois adjuvant et concurrent (stockage, analyse des données, …) vis-à-vis desquels l'INSU doit se positionner:**
  - quand les utiliser ?
  - dans quelles limites ?
  - ...

LIENS AVEC LES ODD

- **La conservation des données sur le long terme à un fort impact écologique.**

  La gouvernance de la donnée – nationale ou locale - doit avoir comme objectif d'en réduire le coût: mutualisation des équipements, proximité des utilisateurs principaux, destruction régulière, …

DÉFI 13 – RECOMMANDATIONS



COMPÉTENCES ET INTERDISCIPLINARITÉ

- **Compétences**

Favoriser la formation, le recrutement (temporaire/pérenne) et la valorisation des profils hybrides:

- Data scientists: Data science + Thématique scientifique (sujets très valorisants pour des doctorants/ postdoctorants)

- Data managers: Gestion de la donnée + Thématique scientifique

- Spécialistes infrastructure: Calcul + stockage

- **Interdisciplinarité**

De fortes interactions hors de l'INSU seraient bénéfiques, notamment pour le traitement des données. Par exemple avec l'INS2i ou l'INSMI au sein du CNRS ou d'autres organismes comme l'INRIA.

La création d'une commission interdisciplinaire dédié à l'analyse des données massives et la mise en place de « Data Challenge » seraient pertinents

COMMUNICATION ET DIFFUSION

- **La diffusion des données fait l'objet du défi 14**