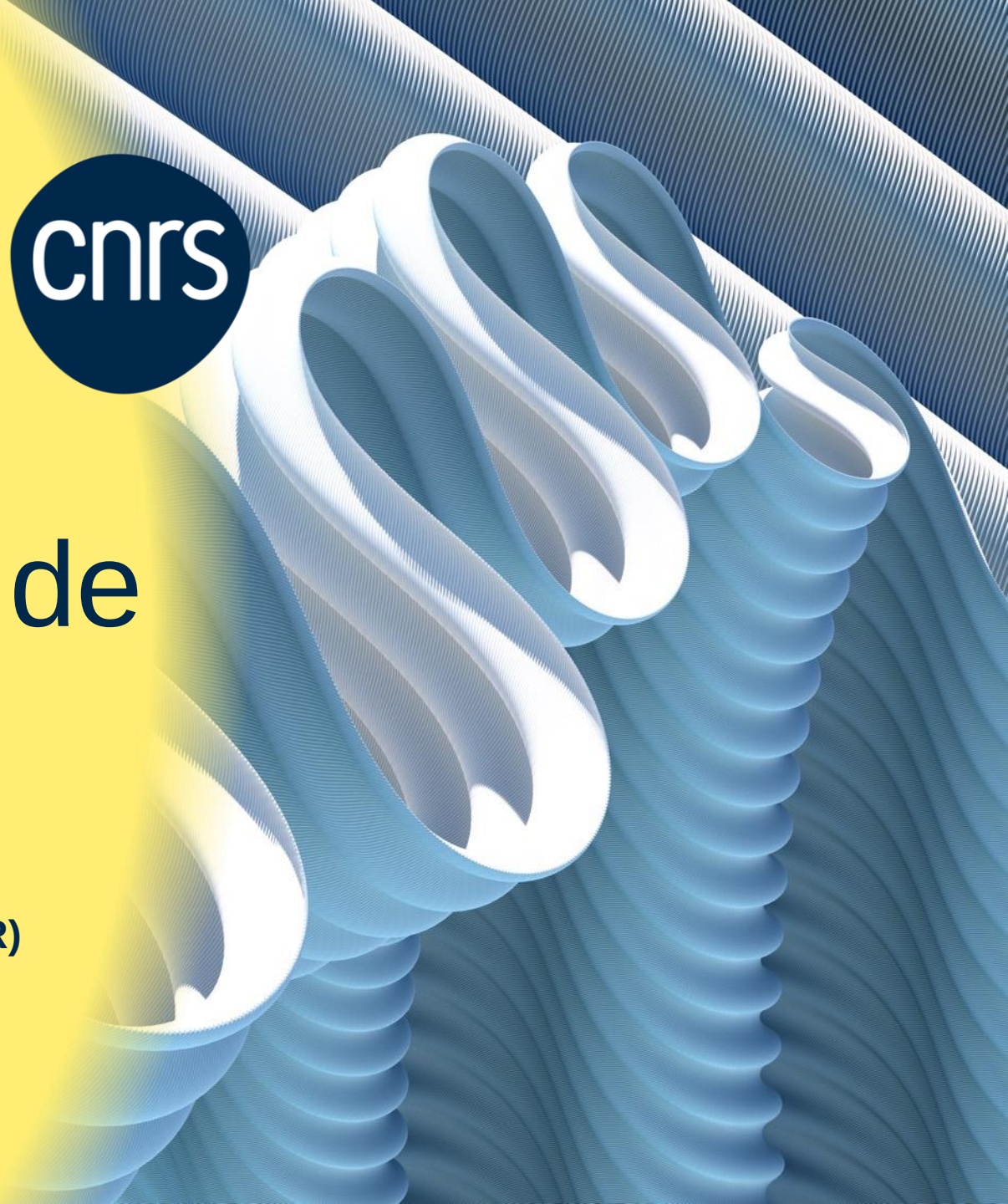
Initiatives en cours autour du stockage de données au CNRS

Denis Veynante

Direction des données ouvertes de la recherche (DDOR)

→ RIP DATA 1-2 octobre 2024



Motivation : ouverture des données

- **Assurer l'intégrité scientifique** (reproductibilité et validation des résultats)
- **Rendre la recherche plus efficace et non redondante** (pas de duplication inutile)
  - taux de perte des données estimé à 20 % / an
- **Être en capacité de réutiliser les données même sans en être à l'origine**
- **Croiser les données** (nouvelles analyses, voire nouvelles thématiques)
- **Satisfaire le cadre légal d'ouverture des données a priori :**
  - *« Ouvert autant que possible, fermé autant que nécessaire »*
  - *Obligation contractuelle (ANR, Europe, …)*

En pratique…

- **Des communautés très organisées**
  - Physique des particules, Astronomie, Sciences de la terre …
- **Une offre générique : Recherche Data Gouv**
  - Entrepôt et catalogue, 20 ateliers de la donnée, 6 centres de références thématiques, 4 centres de ressources
- **… qui ne répondent pas à tous les besoins**
  - Communautés qui ne disposent pas d'entrepôts thématiques
  - Volumétrie limitée (Recherche data gouv : 50 Go par dépôt, 5 To par organisme)
  - Besoin de capacité de traitement à proximité des données volumineuses (limiter les transferts)
- **Mutualiser et rationaliser infrastructures informatiques et ressources RH**
  - *Datacentres labelisés*
  - *Optimisation et réduction des coûts et de l'empreinte environnementale*
  - *Pas de doublons inutiles, ni de trous*
  - *Nouveaux métiers (« data stewardship », …)*

Ne pas développer sa propre solution !!!

CNRS et infrastructures numériques

➢ Opérateur de deux des quatre datacentres d'envergure nationale

- **IDRIS** (Orsay) Calcul intensif
  - Opère le calculateur Jean Zay, financé par GENCI
  - Centre de ressources pour la recherche en intelligence artificielle
  - Projet CLUSSTER
  - Hébergement : mésocentre Paris-Saclay, données CLIMERI, IFB, …
- **CC-IN2P3** (Villeurbanne)
  - Traitement de données massives pour les activités IN2P3 (LHC, LSST, …)
  - Hébergement : DSI CNRS, HAL, HumaNum, BBEES, …

➢ Deux mésocentres rattachés au CNRS (UAR)

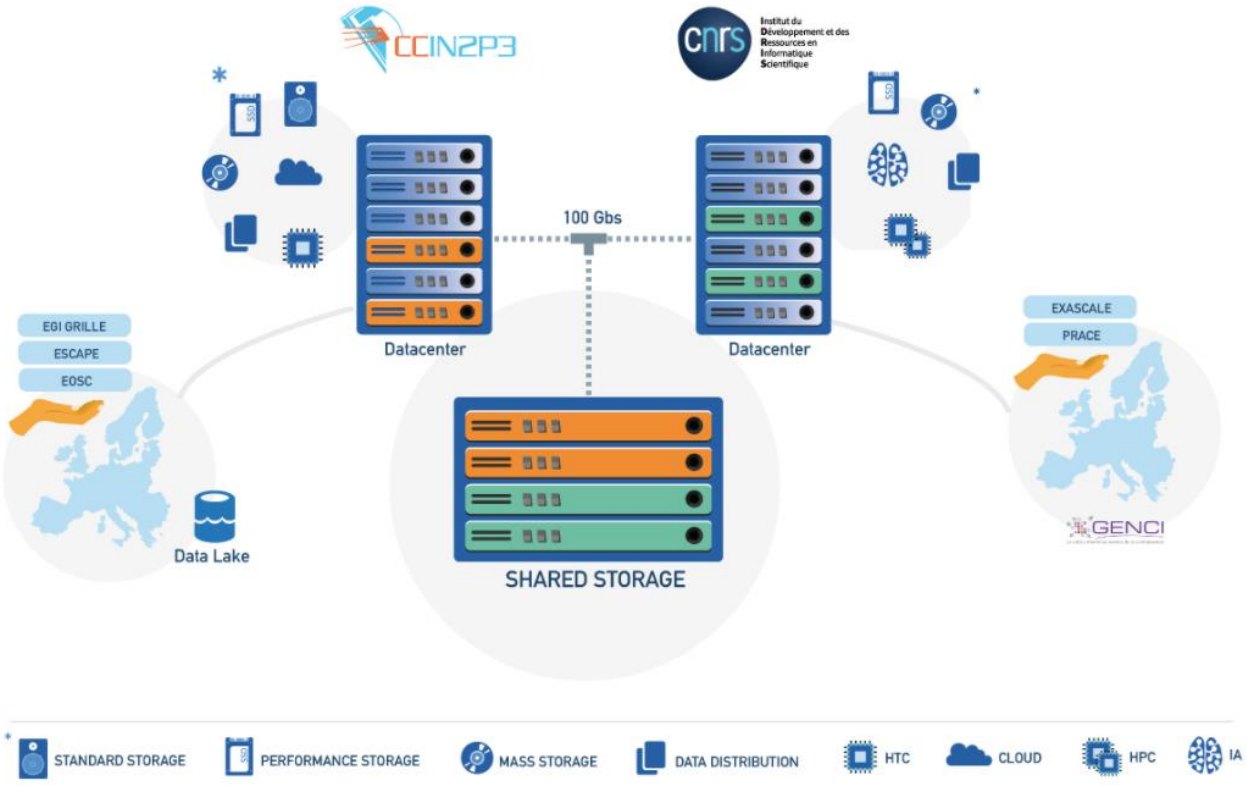
- **CALMIP** (Toulouse)
- **GRICAD** (Grenoble)

+ Demandes d'association d'autres mésocentres

Equipex+ FITS (CNRS Federated IT services for Research Infrastructures)

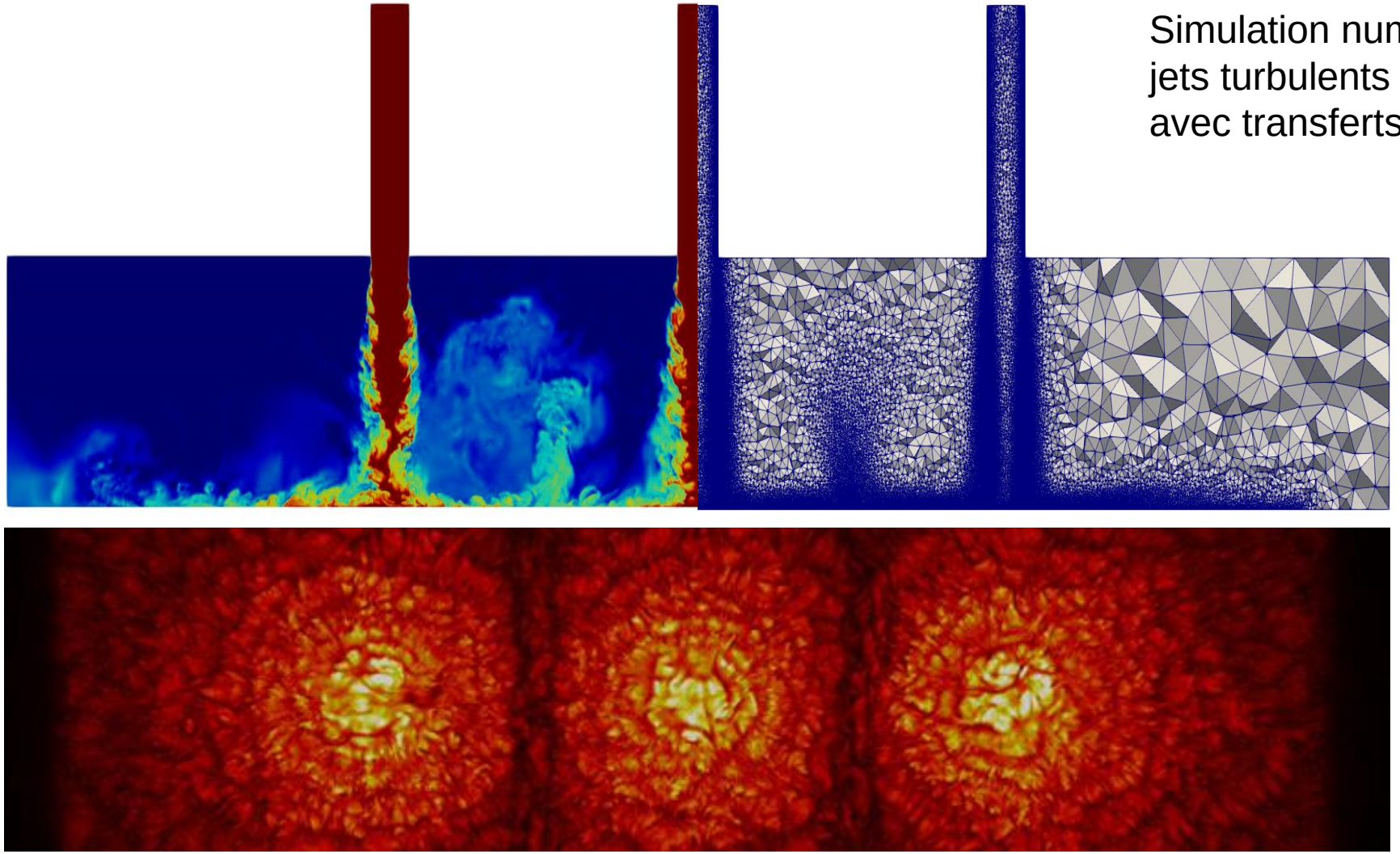
Infrastructure répartie de stockage, traitement, mise à disposition, diffusion et valorisation des données au service des IR/IR*

- Basée sur CC-INP2P3 et IDRIS + partenariat GENCI (calculateur Jean Zay)
- Portail unique d'accès aux ressources
- 4 cas d'usage : Soleil, HL-LHC, LSST, IFB
- Extension des capacités des centres
- 15,4 M€ dont 11,4 M€ de travaux
- 8 ans (juin 2021 – juin 2029)
- Mise en place d'un modèle économique
- www.fits.cnrs.fr



Offre réservée aux IR/IR*

Un autre besoin : cas d'usage HiFiLES4ML



Simulation numérique directe de jets turbulents impactant une paroi avec transferts thermiques

Objectifs :

- Compréhension physique
- Validation de modèles

Base (1 configuration) : 10 – 20 To

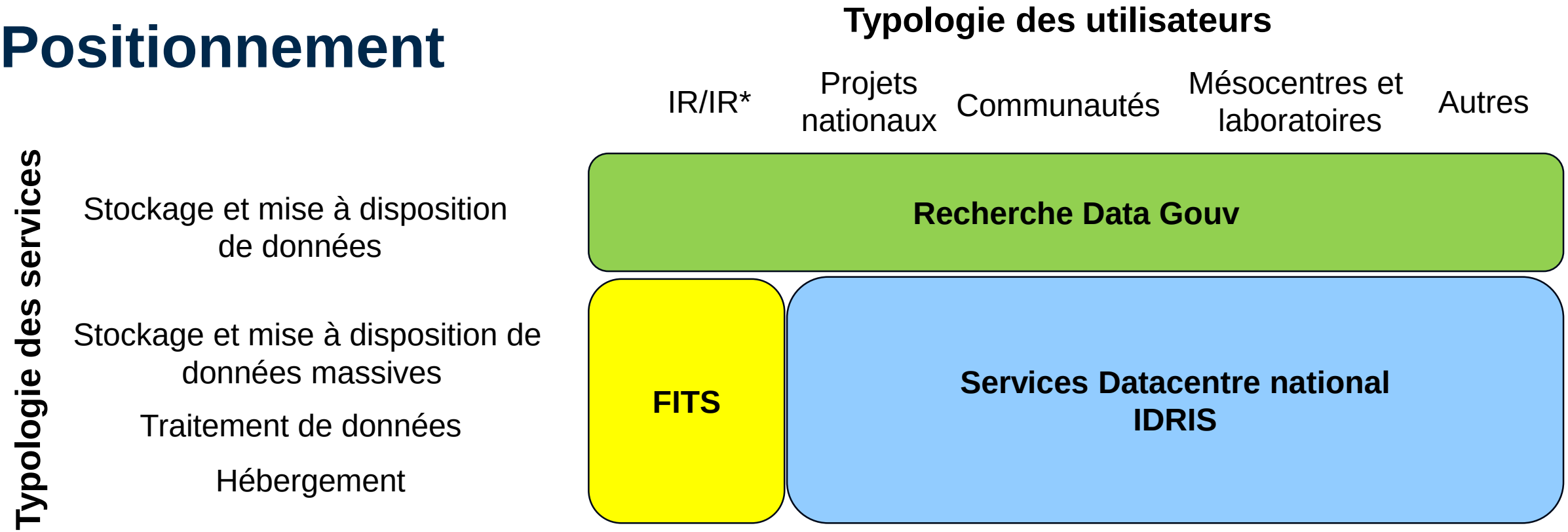
V. Moureau, UMR CORIA (Rouen)

Projet : services datacentre national à IDRIS

- **Développement d'une offre de service pour les données:**
  - Trop volumineuse pour relever de Recherche Data Gouv
  - Issues de communautés ne disposant pas d'entrepôts thématiques
- **Trois types de services :**
  - Stockage de données massives *(volumétrie cible de plusieurs Po)*
    - Chaud *(technologie de type disques)*
    - Froid *(bandes magnétiques)*
  - Traitement de données avec une puissance cible de plusieurs PFlops
    - CPU *(nœuds classiques et nœuds grosse mémoire)*
    - GPU
  - Services d'hébergement de matériels informatiques
- **Rationalisation, structuration et extension de services existants :**
  - Développés pour répondre à des demandes ponctuelles
    - Hébergement et mise à disposition de données du climat *(IR CLIMERI)*
    - Hébergement de calculateurs *(Mésocentre UPSaclay, IFB, …)*
  - Empreinte environnementale maîtrisée et à l'état de l'art

En veillant à la cohérence avec les services et projets existants (FITS, CLIMERI, …)

Positionnement



- **Infrastructure de services extensible et flexible**
  - Adaptation aux nouvelles demandes et nouveaux besoins
  - Hors ZRR
- **Problématiques communes et convergence avec FITS, Clusster, Numpex**
  - Authentification, portail, cybersécurité
  - Mutualisation logicielle et matérielle autant que possible

Evolution : projet commun

Déploiement d'une nouvelle génération d'offre de service de stockage, traitement et mise à disposition de données scientifiques

Data Terra – France-Grilles - IDRIS

- **Analyse des besoins des infrastructures de service aux données (ISD)**
  - Recommandation du document d'orientations stratégiques du CoSIN
- **Déploiement d'une offre nationale de services :**
  - Stockage de données massives
  - Traitement
  - Hébergement
- **Interconnexion des infrastructures de stockage**
  - IDRIS, mésocentres de Clermont-Ferrand et Strasbourg
- **Analyse des coûts et modèle économique**
  - Assurer la pérennité de l'infrastructure

Statut du projet

- **Sollicité et soutenu par la DGRI**
  - Embryon du « cloud stockage et traitement de données »
  - Retenu dans le cadre du « fonds d'amorçage CoSIN » à hauteur de 2 M€ *(attente de versement !)*
- **Co-financé par le CNRS**
  - Pour 500 k€, acquis et notifiés
- **Calendrier**
  - A préciser en fonction des ressources humaines disponibles
  - Déploiement espéré pour les premiers utilisateurs courant 2025
- **Remarques complémentaires**
  - Le projet est extensible et pourra intégrer d'autres mésocentres à l'avenir
  - L'insertion harmonieuse avec les autres acteurs du « cloud » sera un point d'attention

Conclusions - résumé

- **Développement d'une offre stockage et traitement de données**
  - Répondre aux besoins non-couverts par les infrastructures actuelles
  - Infrastructure mutualisée, optimisée et à l'empreinte environnementales maîtrisée
- **Deux projets ambitieux et complémentaires…**
  - FITS
  - Offre de services datacentre national
- **… Qui prendront du temps**
  - FITS attendu pour juin 2029, au-delà des 4 cas d'usage intégrés
  - Déploiement progressif de l'offre datacentre national espéré à partir de mi-2025
- **Modèle(s) économique(s)**
  - Ces infrastructures ont un coût : investissements, fonctionnement, jouvence, extension…
  - Seul « l'amorçage » est aujourd'hui couvert
  - Facturations ? Qui paie ?

Ne pas disperser les efforts ni construire de solutions individuelles ad hoc !!!

Merci de votre attention

→ RIP DATA 1-2 octobre 2024