



RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

anr[®] agence nationale de la recherche



DATA
TERRA



Gaia Data

Infrastructure technique intégrée, équipement et interconnexion de sites

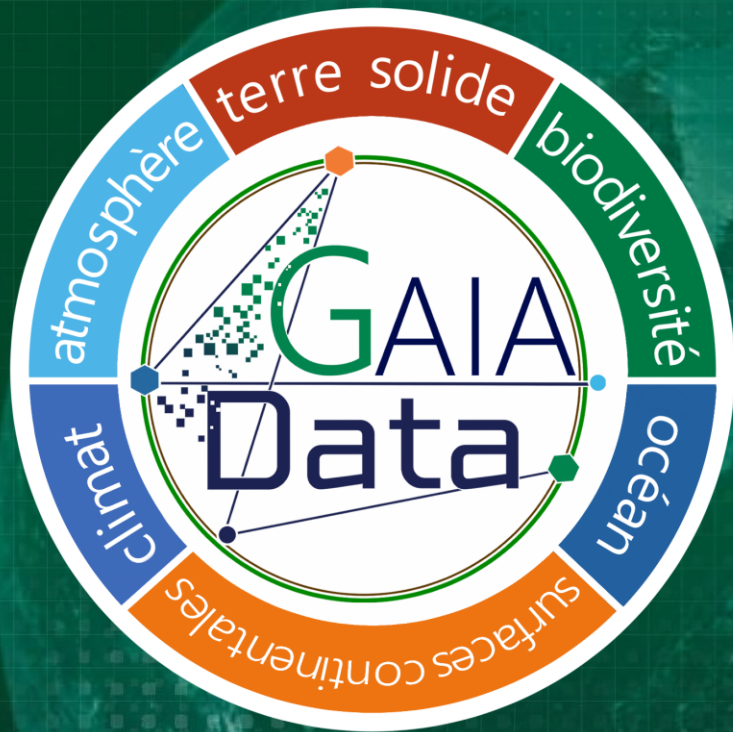


Frédéric Huynh (Data Terra)
Responsable Scientifique du projet Gaia Data
Joel Sudre (CNRS/CPST)
Richard Moreno (CNES)
Karim Ramage (CNRS/IPSL)
Coordinateurs techniques du projet GAIA DATA

Atelier OMP - HPC et Stockage

10/02/2023





Porté par 3 Infrastructures de Recherche



OBJECTIF

Développer et mettre en œuvre **une infrastructure/plate-forme intégrée de données FAIR** et de services distribués pour **l'observation, la modélisation et la compréhension du Système Terre, de la Biodiversité et de l'Environnement**

- **sur l'ensemble du cycle de la donnée**, de son **acquisition** (spatiale, sols, in-situ) jusqu'à ses **multi-usages** (qualification/validation, stockage, accès, traitements/croisements de données multi-sources/extraction de connaissances, produits/services)
- **pour la communauté scientifique** contribuant à la connaissance du système Terre, de la biodiversité et de l'environnement ; **acteurs publics et privés**



0 1

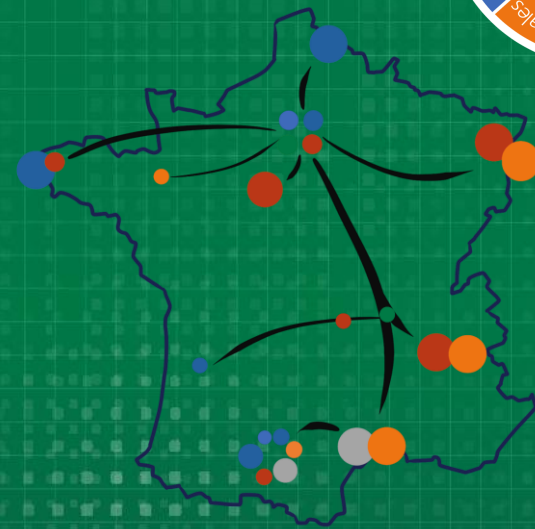
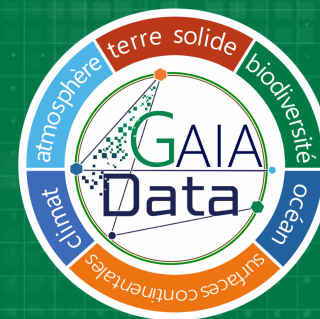
CONTEXTE ET ENJEUX

Contexte des 3 Infrastructures de Recherche

- 30 Centres de données et Services regroupant les experts :
 - ingénierie de la données
 - ingénierie logicielle
 - ingénierie systèmes et réseau
- 50 Po (2020) ; 100 Po (2023) ; 150 Po (2025)
- 50 000 cœurs de calcul cumulés sur les centres de traitement

Stratégie Gaia Data pour les Equipements

Concentrer les investissements en **équipement de Gaia Data** sur les Centres de Calcul et Données des pôles de **Data-Terra**, de **ClimERI** et du **PNDB** compatibles avec la politique InfraNum, tout en **gardant un ancrage régional** essentiel au fonctionnement et au financement des activités des trois IRs.



8 sites principaux
30 sites existants
Dynamique régionale

Projets Nationaux et Européens connexes au projet GAIA DATA



Projets Equipex+ ou PIA4 infra
FITS / MesoNet / Clusster

Projets Equipex+ ou PEPR
thématiques
Obs4Clim / TerraForma /
Marmor

Projets PEPR thématiques
OneWater / Irima / Fair Carbon
Numpex / Traccs

Projets H2020 – Horizon
Europe

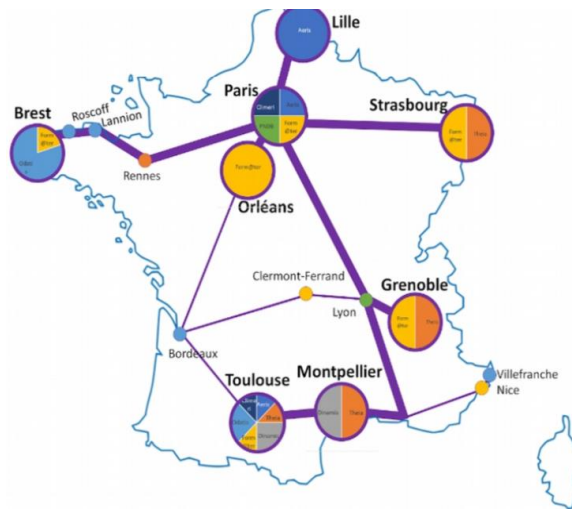
- IS-ENES
- PHIDIAS
- EOSC-Pillar
- FAIR EASE
- FAIR IMPACT



Projets CPER en région



En relation avec des projets connexes



CDS GAIA data

- CDS ossatures multipôles
- Autres CDS

Réseau Renater/GAIA data

- Principal
- Secondaire

IRs impliquées dans GAIA data

- PNDB
- Climeri
- Aeris
- Theia
- Odatis
- Form@ter
- Dinamis

— Data Terra



Intégré dans le paysage international / Européen



02

La solution technique



Services découverte, Accès et Gestion de données

Catalogue (métadonnées, vocabulaires, ontologies), systèmes évolués d'accès et de recherche

Consultation et accès aux données via **web services interopérables** (INSPIRE, Opensearch, STAC, ...)

Services avancés de visualisation

Accompagnement des communautés pour la **FAIRisation**



Services transversaux pour faciliter les travaux transdisciplinaires

Portail connaissances,

Authentification unique

Support utilisateurs & formation – animation communautés

Support aux campagnes

Analysis Ready Data
Datacubes, ...

Grille de données,

Cloud GAIA Data,



Earth Analytics Lab

Exploration de la donnée, Bac à sable, Virtual Research Environment

Capacité à se connecter directement sur les centres

Traitements à la demande

Notebook/PANGEO/STAC

Low code / NoCode : Galaxy-E, FG/VIP



Services de production réguliers

Optimisation des traitements (outils orchestration) et formats de données (Zarr, CoG, ...)

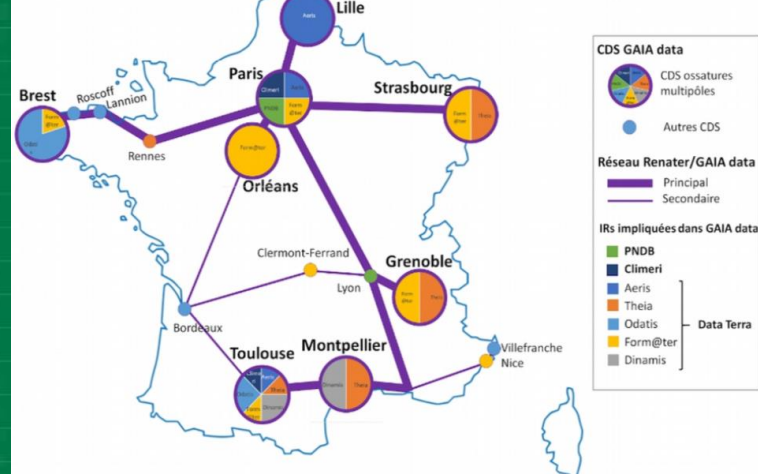
Supporté sur un continuum d'infrastructures partagées

INFRASTRUCTURE GAIA DATA

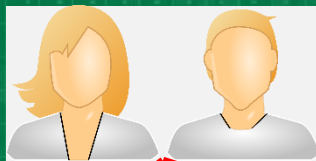
8 Sites regroupant :

- Centres de Calcul Nationaux (CINES, IDRIS)
 - Centres de Calcul et données d'organismes (CNES, BRGM)
 - Mésocentres Régionaux (GRICAD, UniStra, Univ Lille, Meso@LR)
 - Mésocentres Thématiques (Datarmor, ICARE, ESPRI, IPGP-Dante)
- en développant **les liens avec les infrastructures nationales** et en tenant compte des **évolutions du paysage numérique national et européen** au cours du projet.

- **Renforcer les moyens** dans les Centres de Calcul et Données pour assurer les missions des pôles et IR
- **Développer** les infrastructures pour permettre **l'interopérabilité des accès aux données**
 - Mise en place d'un **réseau dédié haut-débit et sécurisé**
 - Déploiement d'une **grille de données** et de **datalakes** pour permettre un **accès distant** aux données et le **transfert rapide et automatique** de grands ensembles de données d'un centre vers un autre
- Développer les systèmes pour assurer **l'interopérabilité des services** entre les Centres de Calcul et Données
 - Système d'authentification fédéré
 - **Interopérabilité des traitements** entre les 8 centres de Gaia Data, avec les centres HPC en France, EuroHPC et avec les clouds commerciaux (DIAS 2.0, WekEO – OVHcloud – Orange BS, ...)
- Renforcer les architectures spécialisées pour le service de la données (visualisation, VRE/VAP, IA, ...)



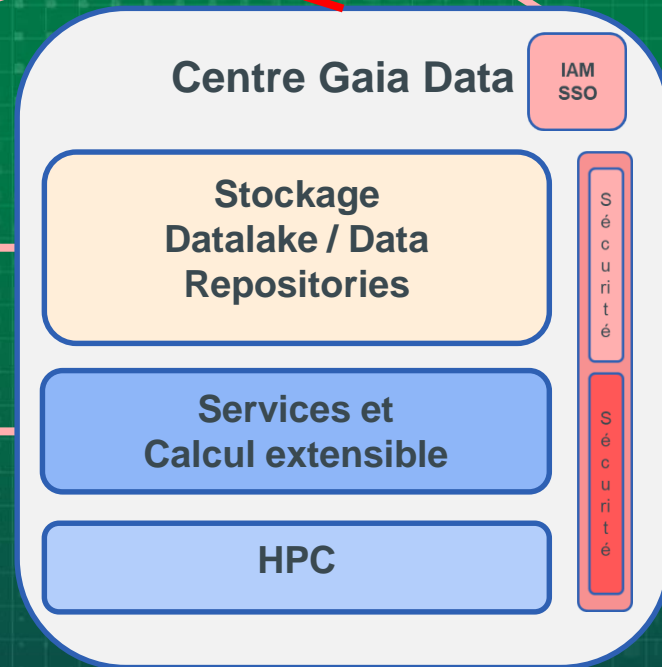
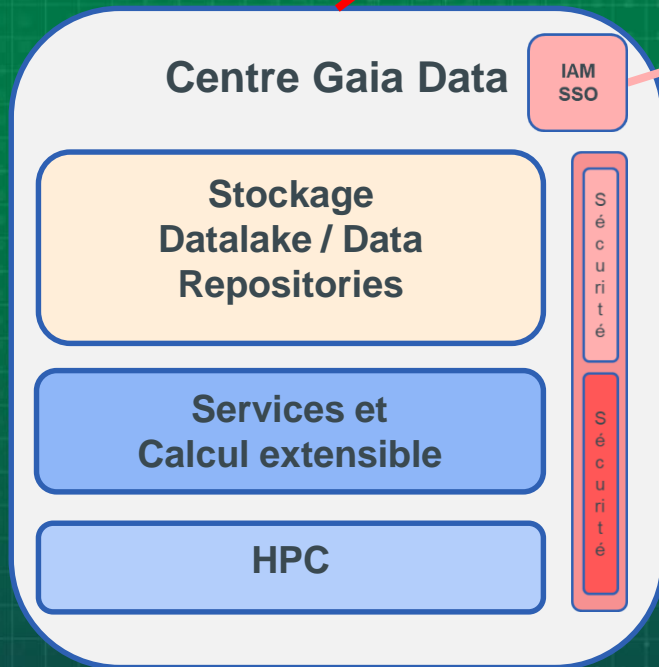
Architecture Gaia Data



IAM
SSO

Accès aux données :
S3, OpenDAP, API, ...

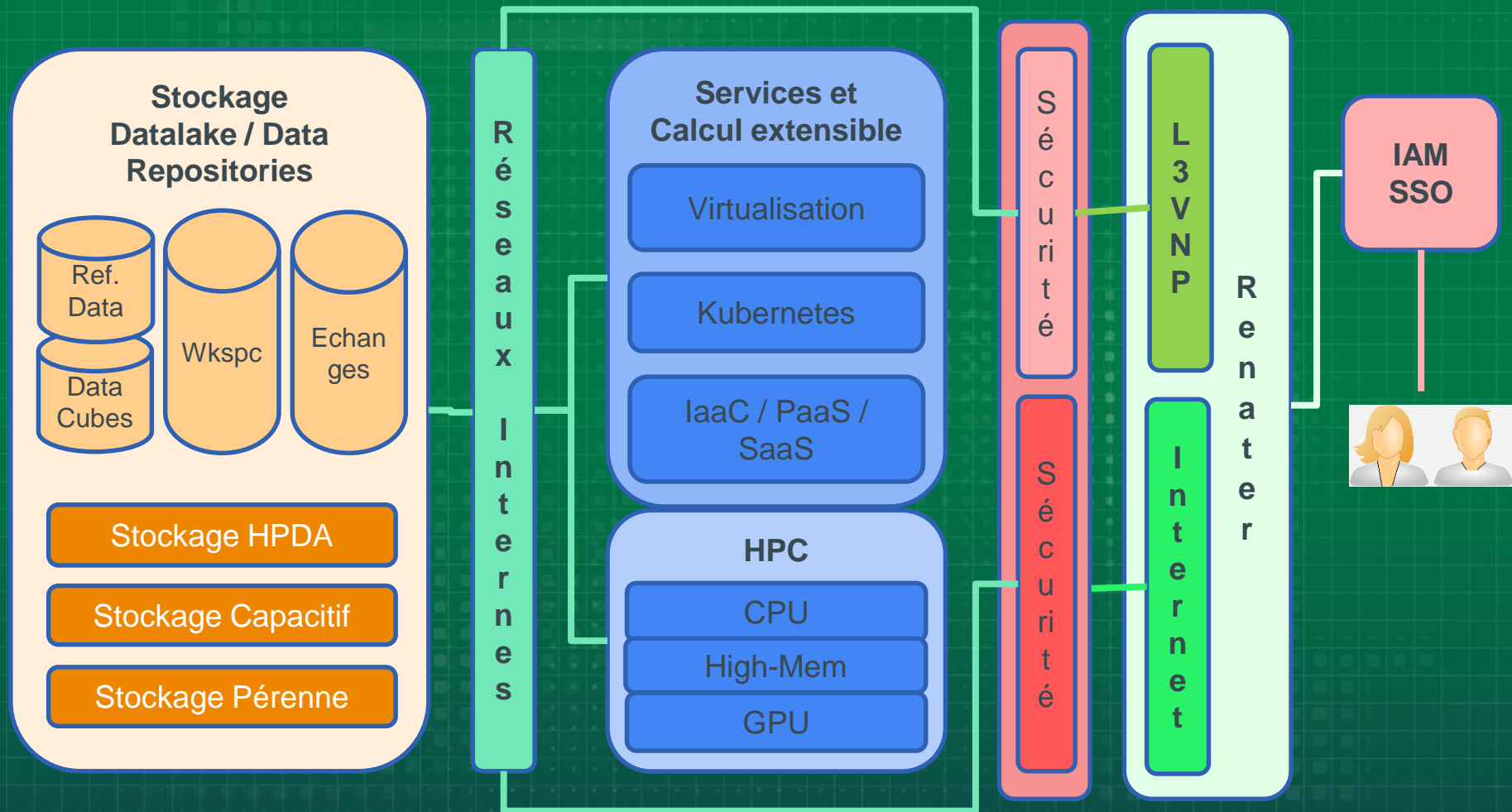
Accès aux Services :
OGC, notebooks, ...



iRods / S3

Docker
Singularity

Architecture Cible des Centres de Gaia Data



Extension des systèmes de stockage capacitif des 8 centres pour les données de références

- **Garantir la capacité des pôles et des IR** à assurer leurs missions pour l'acquisition, le traitement, **l'hébergement** et la distribution des données du système Terre et de la biodiversité

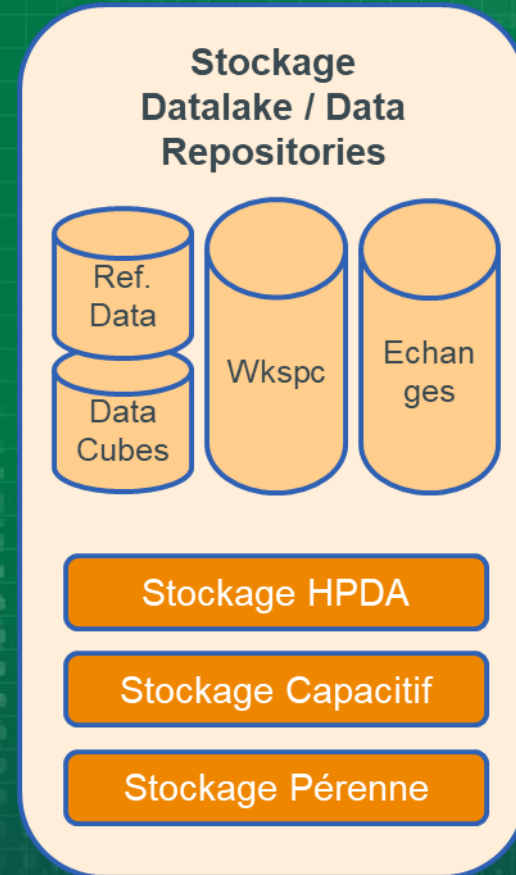
Renforcement des espaces de disques rapides :

- **Stockage au plus près des ressources de calcul** : espaces tampons pour héberger des copies des données du format d'origine, proche de l'observation ou de la simulation, vers des formats de type « **DataCubes** » pour l'analyse (algorithmes parallèles de type IA, par exemple) ou pour la visualisation : images tuilées, facettes 3D, ...

Acquisition d'espaces de stockage « tampon » pour les échanges inter-centres

- Hébergement de **données de références communes** et utiles à plusieurs centres
- **Transfert de données** d'un hébergeur vers un centre de calcul (GENCI par exemple) pour des (re-)traitements massifs
- **Regroupement de jeux de données** multi-centres pour les traitements à la demande de Earth Analytics Labs

Evolution des infrastructures Ressources de Stockage



- **Renforcement des ressources HPC pour la production régulière**
- **Acquisition de nœuds à large mémoire pour:**
 - les traitements rapides à la demande
 - sélections immédiates des données d'intérêt parmi des données extrêmement nombreuses (moteurs d'indexation, TripleStore et bases NoSQL)
- **Nœuds pour la visualisation des données**
 - Processeurs graphiques pour générer des images à la volée
- **Développement des plateformes de virtualisation / containerisation pour l'hébergement des services**
- **Interopérabilité des traitements entre les centres de Gaia Data**
 - environnements logiciels reproductibles, standardisés et paramétrables
 - Packaging des applications et environnements (Guix, Spack, ...)
 - Conteneurisation des codes et applications (docker, singularity)
 - Orchestrateurs d'infrastructure (Terraform, Openstack, K8S) pour faciliter les déploiements de services
- **Développements en lien avec France-Grilles, Clusster, EOSC-Association TF, ESA Cloud**

Evolution des infrastructures

Ressources de Calcul



Déploiement d'interconnexions réseaux pour la grille de données et les services de Gaia Data

- **Réseau privé dédié (L3VPN)** entre les centres partenaires de Gaia Data pour sécuriser les services distribués (ex. : vlan dédiés pour la redondance des services d'authentification)
- **Très haut-débit dédié** pour le transfert et l'accès aux données entre les centres de Gaia Data, puis avec les centres européens partenaires (Projets EOSC, IS-ENES, ...)
- **Haut-Débit vers Internet** pour la distribution de données – 10Gbps

➔ **Mise à niveau des équipements réseaux** sur chaque site en partenariat avec les gestionnaires des réseaux d'accès (boucles locales)

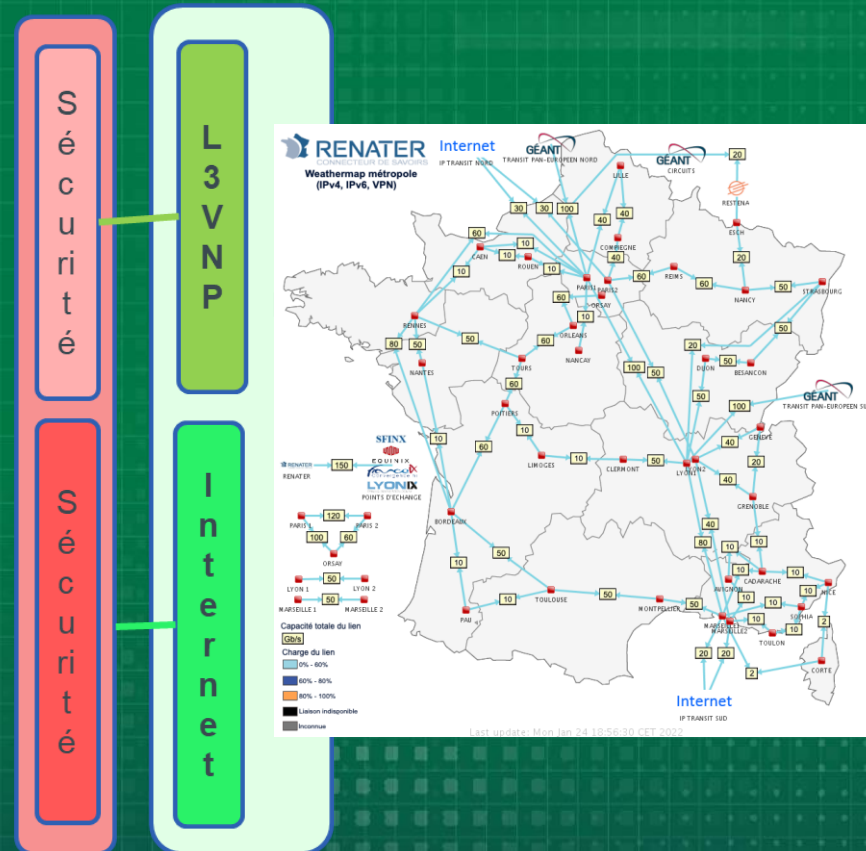
➔ **Discussions au niveau Gaia Data avec Renater** pour la mise en œuvre des interconnexions

Mise à niveau des équipements de sécurité :

- pour les centres exposant les services Gaia Data vers le public
- pour les services inter-centres Gaia Data
- ➔ Surveillance et maintenance sécurité

Evolution des infrastructures

Réseau et sécurité



Déploiement d'une Infrastructure d'Authentification et Autorisation commune interoperable pour l'accès aux données et services de Gaia Data

→ Solution basée sur les travaux menés par AERIS : **Keycloak**

Création d'un annuaire des acteurs et des utilisateurs

- Annuaire centralisé des acteurs et des utilisateurs
- Mise en place d'un mécanisme d'authentification unique
- Proposer un service de gestion des autorisations centralisé

Gestion des niveaux de confiance

- Non authentifié
- Authentification déléguée (Fédération Renater, ORCID, réseaux sociaux, ...)
- Utilisateur authentifié d'un organisme (annuaires existants)

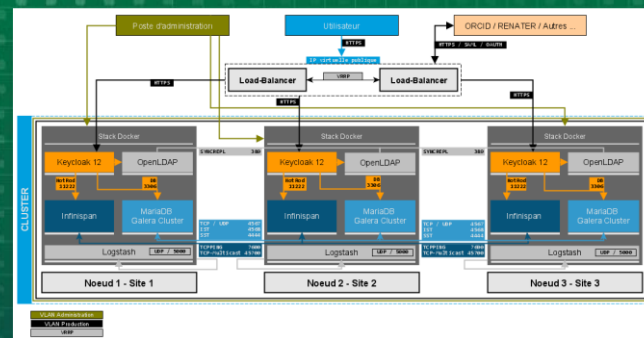
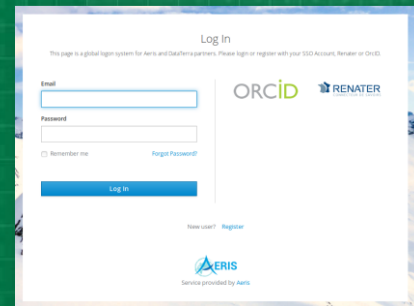
Mise en place de workflow de validation de compte

- Sécurisation fine des jeux de données des catalogues
- Statistique et traçabilité des téléchargements

Possibilité de synchroniser une fédération d'identité déjà en place dans les pôles et IR

Evolution des infrastructures Authentification / Autorisations - SSO

IAM
SSO



Développement d'un système de supervision et d'accounting mutualisé pour:

- Répondre à un niveau de service (SLA) défini,
- Faciliter l'exploitation de l'infrastructure (et le trouble shooting),
- Anticiper les besoins futurs (capacity planning),
- Connaitre les habitudes des utilisateurs, pour pouvoir mieux les servir
- Mesurer l'empreinte carbone des systèmes

→ Supervision des infrastructures systèmes, réseau et des Services

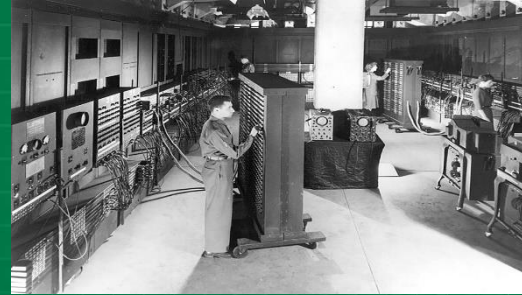
→ Système de détection et d'alerte d'indisponibilité des services, de surcharge des ressources

→ Métriques sur l'utilisation des ressources calcul, stockage

→ Métriques sur l'utilisation de la donnée

Dashboards communs Gaia Data

Evolution des infrastructures Supervision et Métriques





03

Articulation avec les Projets PIA3/PIA4

Articulation avec les Projets PIA3/PIA4 : MESONET

MESONET

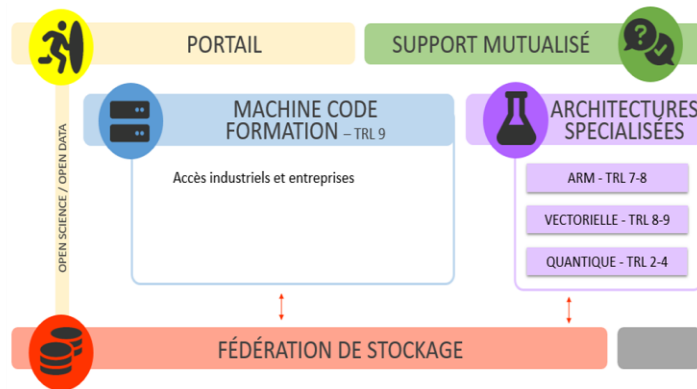
Le mésocentre des mésocentres

Infrastructure nationale distribuée de type *mésocentre*

- Renforcer la structuration de l'offre régionale
- Disposer d'infrastructures calcul / IA au meilleur niveau technologique
- Intégrer les nouvelles communautés
- Encourager les échanges Tiers1-Tiers2
- Fournir une Infrastructure agile pour le développement des codes et la formation
- S'intégrer à la vision nationale et européenne

➤ Créer une Infrastructure de Recherche

14,2 M€ financés sur un budget total de 30,4 M€
début du projet au 01/10/2021 pour une durée de 6 ans



GENCI
Le calcul intensif au service de la connaissance

Université
des Antilles

Université
de Lyon
FLMSN

GIP numérique
de Bretagne

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTE

UNIVERSITÉ
POitiers

Université
de Corse
Université
de Pau
Université
de Montpellier

UNIVERSITÉ
DE BRASSE
CHARENTAIS

Université
de Strasbourg

Université
de Lille

Université
de TOURS

université
PARIS-SACLAY

CentraleSupélec

PSL
UNIVERSITÉ PARIS

CRIANN

université
BORDEAUX

Université
Fédérale
de Toulouse

Université
de Bourgogne

Aix-Marseille
UNIVERSITÉ

UNIVERSITÉ
CÔTE D'AZUR

CENTRALE
NANTES

- Mésocentres régionaux participant aux deux projets
- Problématiques communes :
 - Fédération du stockage : 20 Po distribués iRods (workdir / staging)
 - Interconnexions réseau
 - Authentification fédérée
 - Sécurité (12 sites démarche d'audit sécurité)
- ➔ Participation croisée aux GT des différents projets
- ➔ Co-financements pour certains équipements

Articulation avec les Projets PIA3/PIA4 : CLUSSTER

CLUSSTER

Cloud Unifié Souverain de Services, de Technologies et d'infrastructures

WHAT?

Réponse à AMI de Bpifrance relatif à la stratégie d'accélération cloud : Développement et renforcement de la filière française et européenne du Cloud



Contexte

- AI For humanity (2018)
- Recherche ouverte: Jean Zay (GENCI/IDRIS)
- Recherche confidentielle : OVH, Atos, Activeeon, Qarnot, etc.
- Des expertises académiques
- Des expertises industrielles métiers
- *Aucune offre unique adressant Recherche ouverte et confidentielle et offrant des services d'expertise*



Un portail unifié et souverain

- Fédérer l'ensemble des infrastructures existantes des acteurs privés et publics et des offres de services à valeur ajoutée
- Recherche ouverte, confidentielle et activités lucratives
- Secteur Académique, Industrielle, public
- Intégration écosystème européen; GAIA-X et EOSC

- Faciliter la lisibilité de l'écosystème pour les utilisateurs et l'usage de toutes les ressources/services existantes en France
- Accompagner: Formation, Veille techno, expertise métier

Perspective

- Evolution des services : support à l'adhésion pour couvrir de nouveaux verticaux métiers
- Evolution des domaines: Extension au quantique et à la simulation numérique



Convergence possibles entre les projets :

- Gaia Data : Use Case de Clusster
- Authentification fédérée
- Portail d'accès unifié
- Interopérabilité des traitements
- Capacité de débordement
- Hébergement des services Gaia Data pour les utilisateurs du secteur aval
- Portail vers les infrastructures européennes (EOSC, Gaia-X)





RÉPUBLIQUE
FRANÇAISE
*Liberté
Égalité
Fraternité*

anr[®]
agence nationale
de la recherche



DATA
TERRA



*Ce travail a bénéficié d'une aide de l'Etat
gérée par l'Agence Nationale de la Recherche
au titre du programme Investissements
d'Avenir Equipex+.*



contact@gaia-data.org

www.gaia-data.org



Observatoire
de la CÔTE AZUR

