Investissements d'Avenir

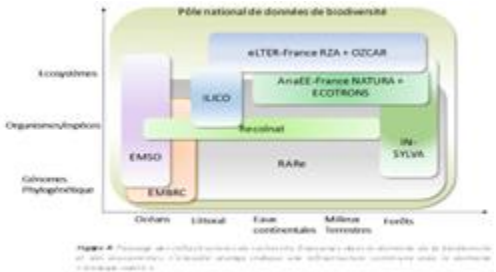
Appel à Manifestations d'Intérêt

EQUIPEMENTS STRUCTURANTS POUR LA RECHERCHE / EQUIPEX+

Projet GAIA Data
Infrastructure distribuée de données et services pour l'observation, la modélisation et la compréhension du système Terre, de la biodiversité et de l'environnement

Projet GAIA DATA

Les 3 Infrastructures de Recherche



Data Terra organise l'accès et les traitements intégrés de données d'observation, produits et services couvrant les différents compartiments du système Terre et leurs interactions

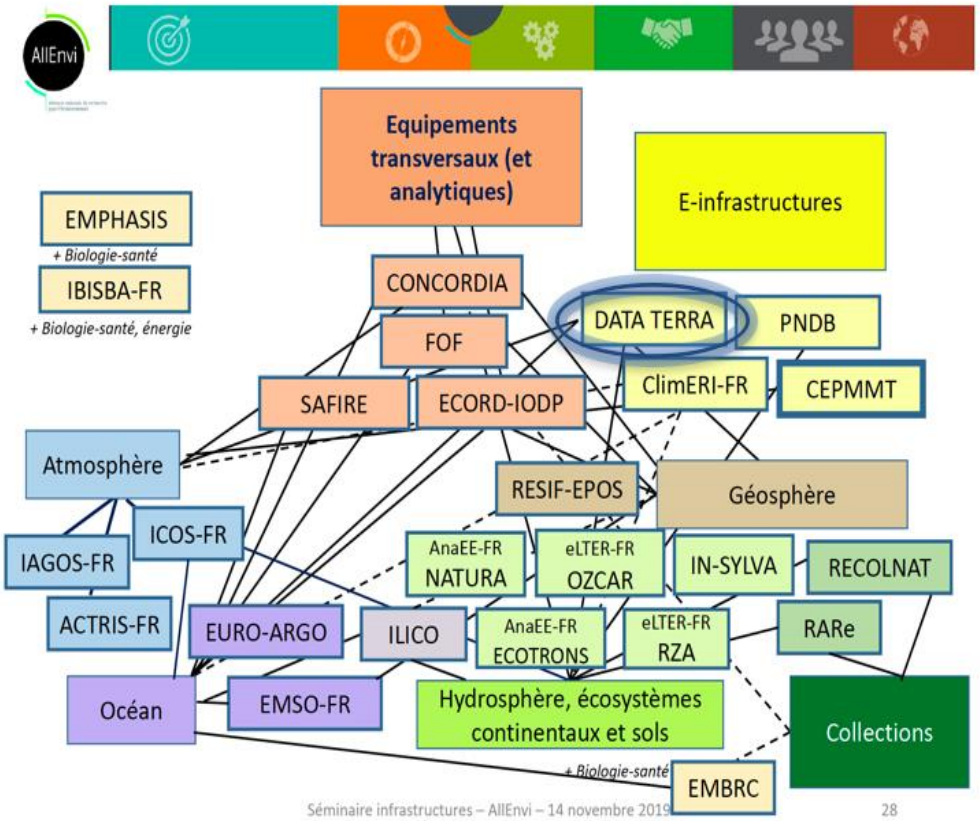
CLIMERI-France est l'infrastructure nationale de modélisation du climat, sa mission est de produire des simulations numériques internationales pour le PMRC et de mettre leurs résultats à la disposition de divers utilisateurs en France et à l'étranger.

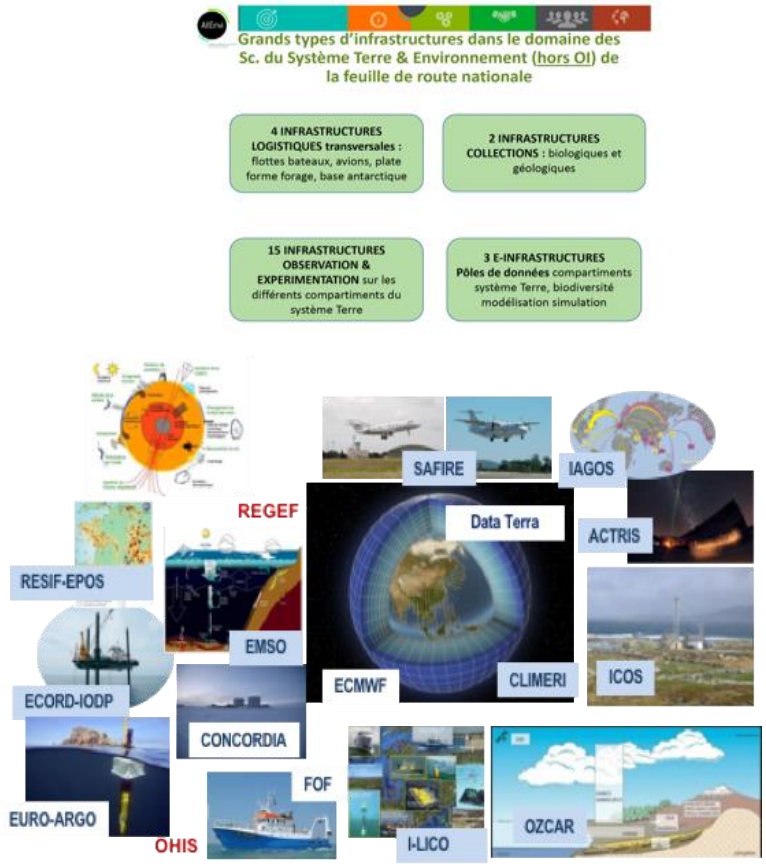
Le PNDB, le centre national de données sur la biodiversité, vise à fédérer les approches de données existantes au sein des infrastructures de recherche sur la "Terre vivante".

Projet GAIA DATA

Les 3 E-Infrastructures de Recherche

Infrastructures de recherche du domaine Système Terre et Environnement





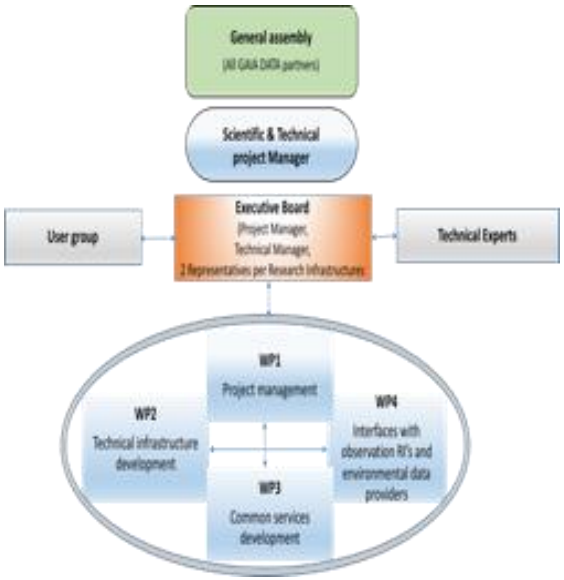
Projet GAIA Data : résumé

Porté par 3 Infrastructures de Recherche / E-Infra du domaine « système Terre et Environnement »
Data Terra, CLIMERI, PNDB

21 Partenaires *: CNRS (coord.), CNES, IFREMER, IRD, BRGM, IGN, INRAE, Météo-France, MNHN, CEA, IPGP, CINES, Sorbonne Univ., Univ. Grenoble-Alpes, Univ. Lille, Univ. F. Toulouse, UNISTRA, SHOM, OCA, FRB, CERFACS*

Objectif : Développer et mettre en œuvre **une infrastructure/plate-forme intégrée de données FAIR et de services distribuées 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 (observation, modélisation),** 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**



Budget : 65 M€ *(coûts complets partiels car principalement RH permanents)* - **Demande ANR-EQUIPEX+**PIA3 : 19,6 M€ demandé => **16,2 M€ obtenu**

RH : **339 ETP** (soit 4066 p.m.) en personnels permanents mobilisés ; 59 ETP (711 p.m.) cdd (1 pour 5,6 permanents)

Co-financements (organismes, projets Européens, ..) : plus de 45 M€ (apports RH consortium), 5 M€ (EOSC-Pillar, Phidias, Blue Cloud, …)

Projet GAIA DATA : contexte et enjeux

- **Constat** : Des systèmes d'information **existent** mis en place par les **pôles de données : IR** DATA TERRA, PNDB (pôle national de données de biodiversité) et CLIMERI-France pour les données de simulations climatiques. **Une organisation par domaine**, voire par source de données, avec des standards et des outils différents, avec une grande **diversité** et un large spectre de **volumes de données**
- **Verrous**
  - Intégrer des données hétérogènes, complexes, multidisciplinaires, multi-sites
  - S'adapter aux pratiques interdisciplinaires d'utilisation de données
  - Gestion « à la demande » de gros volumes de données en particulier spatiales, services IA
  - Prendre en compte la diversité de plate-formes et infrastructures réparties sur le territoire et opérées par de nombreux acteurs

Enjeux

- **Mettre en œuvre une plateforme intégrée de données et services distribuées soutenues par des centres d'expertise scientifique du domaine**
- **Développer des services** accessibles, via des portails permettant des recherches et traitements **inter et transdisciplinaires** à partir **de données multi-source acquises par satellites, navires, avions, drones, submersibles, ballons, dispositifs in situ, inventaires, observatoires et expérimentation, ainsi que, sur des données issues de simulations de référence**
- Co-construire, organiser et adapter les services avec et p**our les communautés scientifiques du domaine système Terre et environnement,** les acteurs **publics et socio-économiques**

Projet GAIA DATA – Ambitions et Objectifs

Etre à l'avant garde en faisant évoluer les E-infrastructures de données thématiques vers un continium d'infrastructures distribuées de données, services et de connaissances du système Terre et de l'environnement

- **Mettre en œuvre au plan national, européen et internationale une infrastructure distribuée de services innovante du domaine système terre et environnement**
- Travailler, en étroite relation avec l'IR* GENCI, avec les centres nationaux, HPC nationaux (CINES, IDRIS, CCIN2P3) et centres de données régionaux labelisés/Meso-Centres
- Renforcer les synergies et collaborations avec les IR et IR* d'observation (Terre Solide, Atmosphère, Océan, Surfaces continentales, biodiversité, …) et IR Numérique
- Contribuer à la souveraineté des données et connaissances scientifiques et technologiques (préservation des connaissances ; maîtrise de la chaine de valeur-ajoutée : données – informations – connaissances)
- Contribuer aux initiatives nationales (science ouverte, Infranum,…), européennes (EOSC, Copernicus, DTE, …) et internationale (GEO, GoFAIR, …)
- Participer à la mise en œuvre des jumeaux numériques du système Terre dans le cadre Destination Earth

Projet GAIA DATA
Une infrastructure distribuée fortement mutualisée

Caractéristiques de la plateforme de services

- Basée sur des équipements, ressources et infrastructures **existants interconnectés** et **renforcés**
- Des **services distribués** aux **capacités et fonctionnalités nouvelles** facilitant le **croisement** et **l'exploitation transparente et continu** du **dispositif national**, mais aussi européen et international
- Facilite l'exploitation de **gros volumes de données** et la **génération à la demande** d'informations **combinant des données et des produits d'origines multiples**, des **satellites** aux **observations** sol, d'expériences de laboratoire ou de terrain aux **sorties de modèles**

Avancées majeures : mise en œuvre d'un **continuum de services ouverts**, **interopérables, FAIR et distribués** permettant, de manière **transparente** et **continue**, de mobiliser et d'exploiter un **continuum de ressources avec des outils adaptés** aux besoins des communautés scientifiques

S'appuie sur des **architectures de type Cloud hybrides distribuées**, **flexibles** et **optimisées énergétiquement**

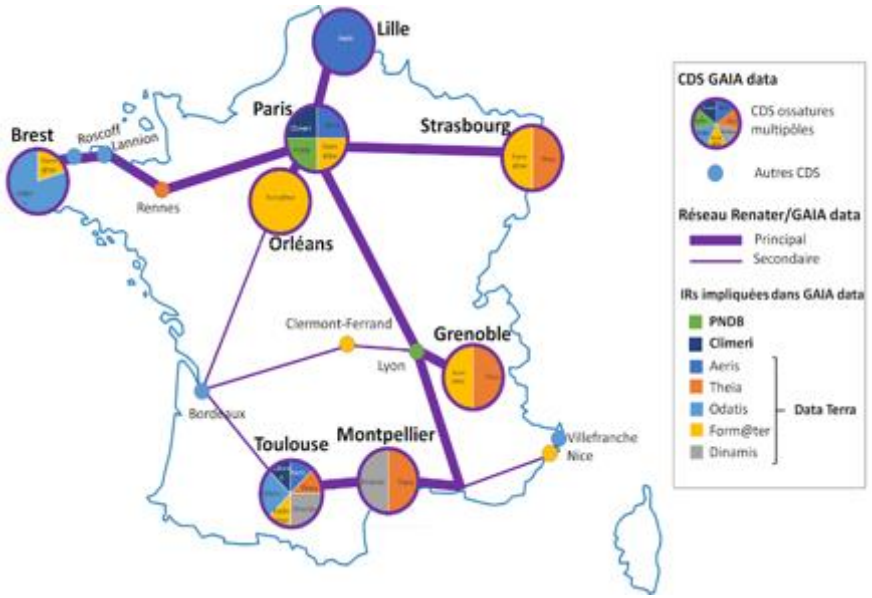
=> Ce projet dotera la France d'une capacité inédite qui confortera ainsi son positionnement européen et international (EOSC, Copernicus, Destination Earth, GEO, …)

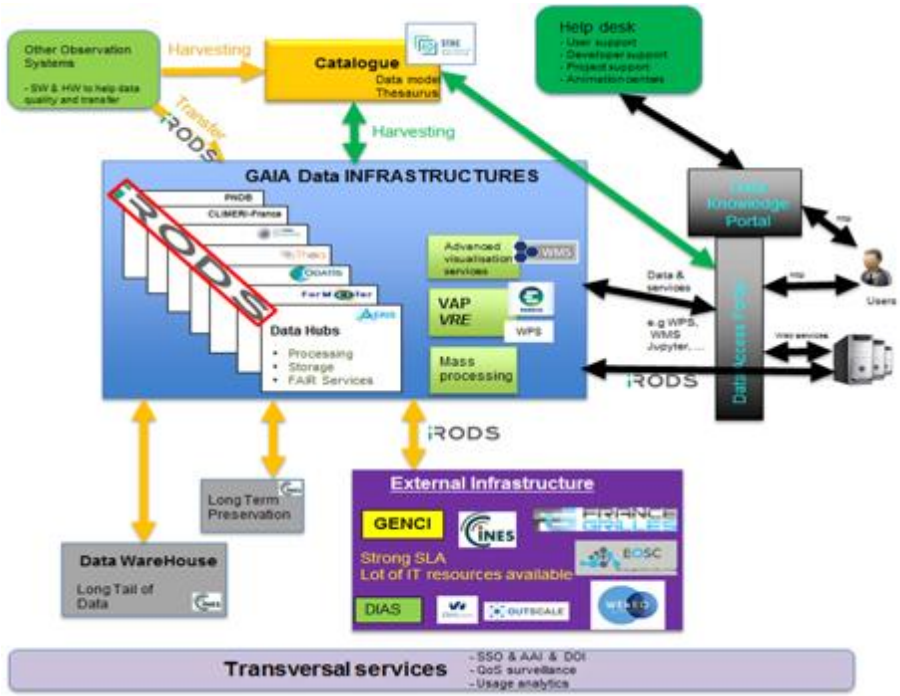
GAIA DATA Caractéristiques du projet

Infrastructure distribuée de services

8 sites principaux

30 sites existants

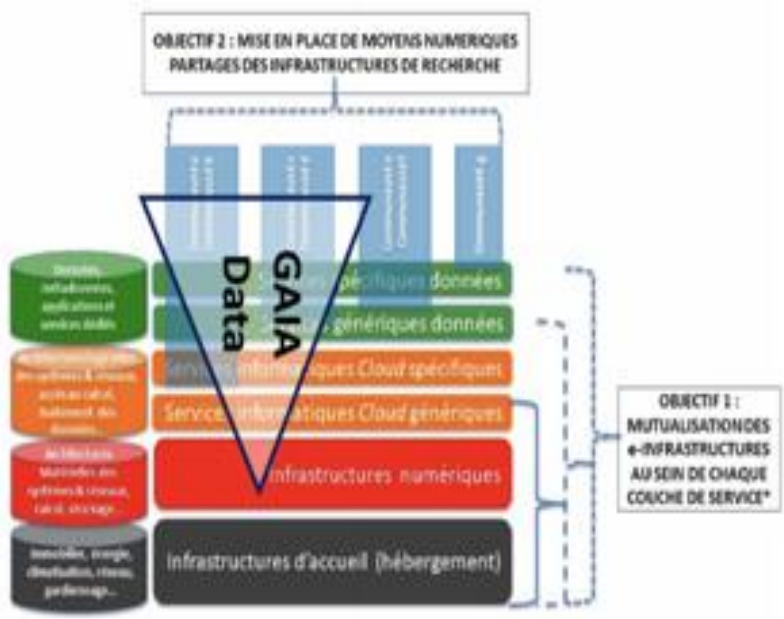




Projet GAIA DATA

Plateforme de services distribuée interopérable avec les centres nationaux, régionaux, Européens et Internationaux

- **Articulation** avec les **infrastructures nationales** et régionales : **faciliter un accès intégré à un continuum de services distribués** de **stockage** adaptés à la diversité et aux volumes de ces données, et de **calcul**
  - ⇒ *accélérer les chaînes de traitement, d'analyse, de modélisation et de visualisation de ces données multi-source, ainsi que leur logistique tout au long de ces chaînes, au travers du continuum de ressources.*
- Collaboration avec la TGIR GENCI, via le **CINES**, et l'IDRIS, avec le CC-IN2P3 : contribuer à **l'évolution du modèle des infrastructures de calcul et de données**, et à la **convergence** entre **calcul** et **l'analyse de données haute performance**,
- Activités **co-construites** et co-implémentées en **coordination avec les politiques nationales** (science ouverte, **InfraNum**, …), régionales, Européennes (**EOSC**, ESFRI, Copernicus, …) et internationales (**GEO**, ONU Env., PMRC, GBIF,…).



Projet GAIA DATA : Services proposés

Services Découverte, Accès et Gestion des données

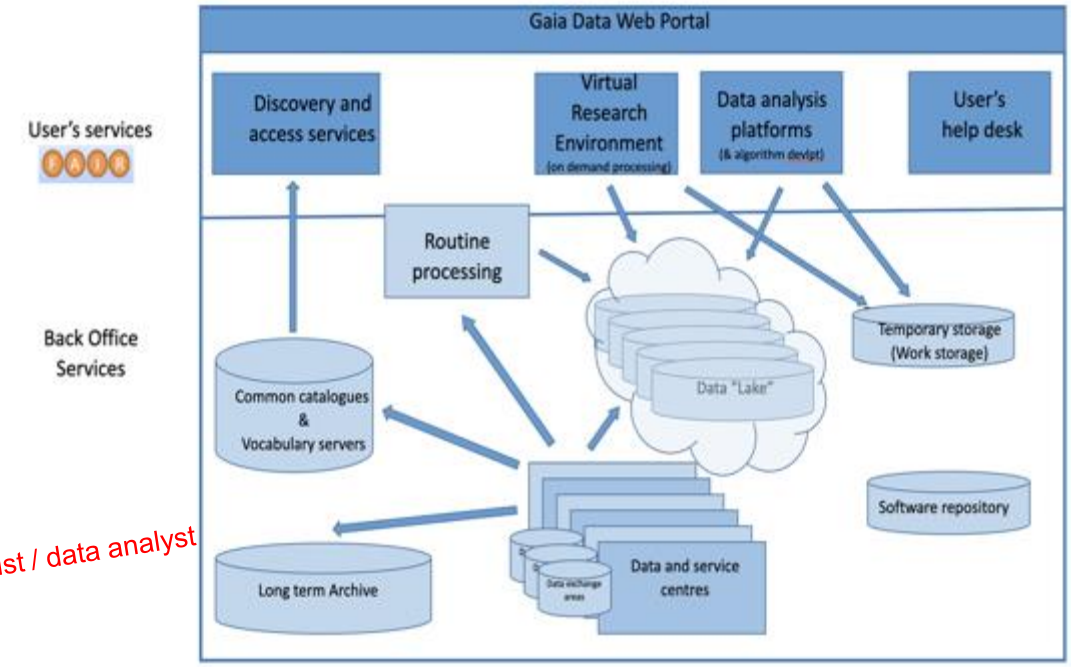
- **Catalogue** (métadonnées, vocabulaires, ontologies), systèmes d'accès et de recherche
- Archive long terme, **entrepôts, DOI, Services avancés de visualisation**
- Aide à la collecte des données des observatoires

Services transversaux => faciliter les travaux transdisciplinaires

- Grille de données, cloud, standardisation de la production, portail connaissances, support utilisateurs & formation

Services – Virtual Analysis Platform => Data Scientist / data analyst

- Travail collaboratif, bac à sable, développement et exécution d'algorithmes
- Ecosystème PANGEO/STAC/Intake



⇒ Thématiciens
⇒ Scientifiques non informaticiens

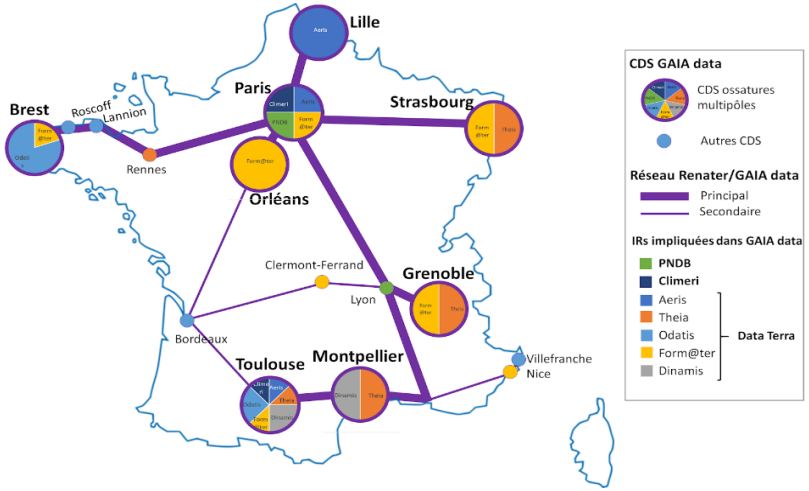
Services de production réguliers

- Optimisation des traitements et formats de données (Zarr, CoG, Dask distributed computing, …)
- Supporté sur un continuum d'infrastructures partagées

Services d'analyse des données à la demande & Virtual Research environnement & AI

- Interface interactive
- Exécution par les utilisateurs
- VRE : définition et exécution de workflows de traitements spécifiques des domaines

Infrastructure Gaia Data



30 Centres de données et de services dont 8 HPC-Tier2 => science (big)data centers

- 400 scientifiques, ingénieurs et techniciens (experts des données, thématiciens)
- Plus de 300 produits et services, plus de 15000 utilisateurs

Infrastructure actuelle :

~50 Po de données => 100 Po en 2023, 150 Po en 2025

- Croissance des données d'observation spatiales, in-situ, modélisation, ainsi que IoT & 5G (capteurs personnels, science participative, …)

- 50 000 cœurs de calcul cumulés

Grille de données et de services en mettant en réseau les 8 principaux centres

- **Mise en place d'un réseau dédié haut-débit et sécurisé** entre les 8 centres principaux

- **Déploiement d'une grille de données** (système iRODS) sur les 8 centres pour permettre un accès distant aux données et le transfert rapide et automatique de grands ensembles de données d'un centre vers un autre.

- **Interopérabilité des traitements entre les centres de Gaia Data** (Cloud Management Framework d'EGI) sur les 8 centres pour offrir des interfaces interopérables d'accès aux ressources de traitement

- **Possibilité d'utiliser des moyens externes** (GENCI, EGI, DIAS, clouds public)
- - Si besoin d'un SLA supérieur ou de débordement des capacités de traitement
- - Pour les usages « externes » tels que applications commerciales

Projet GAIA Data : négociation ANR/MESRI

Orientations stratégiques et techniques négociés avec l'ANR :

- Maintenir les composantes développements et mise en œuvre des services aux données et en particulier les **RH et l'expertise**
- Réduire les investissements « Equipements » de manière différentiées pour disposer d'une plate-forme distribuée adaptée aux besoins de la communauté environnement
- Maintenir l'organisation et articulation à partir de 8 sites principaux tout en sélectionnant les investissements en fonction de synergies possibles avec d'autres projets PIA3/Equipex+, projets CPER en lien avec les politiques nationales de labélisation des data centre et la présence de Meso-centres partenaires,
- Réorganiser le projet en renforçant la phase de développement par rapport à la phase d'exploitation

Projet GAIA Data

Engagements significatifs des organismes et Universités

- Renforcement du projet par des **engagements supplémentaires des organismes** partenaires permettant d'absorber une partie de la réduction budgétaire

  => + **25 postes supplémentaires** (CDI, mobilités/recrutements, …) sur 5 ans

- Positionner le projet GAIA Data comme **nœud France des services** / hub de données et de services EOSC pour l'environnement ou initiative française contribuant à l'initiative **Destination Terre** (Destination Earth) et aux développements de digital twins earth, en vue de participer et obtenir des moyens et cofinancements dans le cadre des projets Horizon Europe 2022-2027.

Projet GAIA Data

Collaborations fortes avec des projets PIA3-EQUIPEX+

OBS4CLIM : porté par les IR Atmosphère (ACTRIS, ..)

Le volet données est assuré par le pôle AERIS - besoins pris en compte dans le projet GAIA Data

TERRA FORMA : porté par l'IR OZCAR et RZA

Liens via le pôle THEIA au travers de THEIA - OZCAR In situ pour la partie système d'information

Lac de données / service de collectes : composante à mutualiser ; Plateforme d'annotation IA (sciences participatives) ; gestion des observations (et observateurs smartphone) / dimensions science participative

MARMOR : Projet porté par les IR Terre Solide

ANVOL : porté par l'IR SAFIRE

FITS : Projet porté par le CNRS IN2P3 et INS2I : IR CCIN2P3, IDRIS

MESO-NET : porté par GENCI et Meso-Centre

Projet GAIA Data

Synergies avec des projets CPER

Occitanie

- **GEO Data Terra Occitanie** : développement d'une infrastructure de données spatiales distribuées : stockage gros volumes, données spatiales FAIR, Data Lake pour le traitement des données spatiales. Coord. CNES ; partenaires de l'IR data Terra ; Renforcement du dispositif du centre de données et services DINAMIS/THEIA (données THRS) pour des services territoriaux.
  Coord : CNES, INRAE, IRD, CIRAD, CNRS, IGN 14 M€ dont 6,725 M€ demandé (5M€ TLS ; 1,725M€ Mpl)
  => forte réduction budgétaire attendue

Bretagne

- Projet **AIDA** : Infrastructure Science des Données et IA de Bretagne Océane. Coord. Ifremer ; SHOM, UBO, ENSTA, CNRS, IRD, SB Roscoff, INSERM, IMT, Ecole Navale, CEREMA, AFB. Budget : 12M€, 375 hm – 4,1 M€ demandé
  Apport Etat/Ifremer/Shom : 2,8 M€ => accepté mais réduction à 6,55 M€

Haut de France

- Projet **CLIMENSE** : changements CLimatiques : Impacts ENvironnementaux sur la Santé des organismes et des Ecosystèmes - CPER (2021-2026)
- Projet **Cornelius** : démonstrateurs d'utilisation de l'IA pour le processing de Big Data

Iles de France

- Centre de données Energy4Climate (Energy4Climate datahub), coord Institut Polytechnique/LMD, AERIS/ESPRI. 1,5 M€ => pas d'information

Rhone-Alpes

- Projet CINAuRA 30 M€ (50% construction de datacenter + réseau; 50% calcul-stockage)

⇒ pas d'information

En attente des montants arbitrés définitivement

Projet GAIA Data

Conclusions

- L'organisation du projet **garantit sa mise en œuvre en insistant sur les objectifs à fort impact pour la communauté et l'écosystème de la donnée environnementale**
  - Priorité aux développements et à la mise en œuvre des services aux données
  - Priorités aux renforcements des compétences (RH), à l'expertise et à la cohérence avec les projets des centres de données régionaux, de Meso-Net et de GENCI
  - Engagements significatifs des organismes, en plus des engagements initiaux

- Recherche d'une temporalité de l'investissement permettant via la montée en compétence de la phase initiale :
  - L'attractivité pour les communautés
  - L'implications dans des projets européens (EOSC, jumeaux numériques, …)
  - La recherche active de cofinancements dans Horizon Europe
  - Le cofinancement sur la durée par les partenaires

- GAIA Data permettra à la France une contribution significative à EOSC (Environmental FAIR data and services), Copernicus et Destination Earth au travers des digital twins…