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

JOURNÉE DE LANCEMENT INSTITUTIONNELLE DU PROJET GAIA Data

12 AVRIL 2022

PROGRAMME

PROGRAMME

8h30-9h00 **Café d'accueil**

9h00-10h00 **Ouverture officielle -** Animée par **Fabienne Chauvière**, Journaliste-productrice Radio France

Mots de bienvenue de **Bruno David**, Président, MNHN

Discours d'ouverture par **Cyril Moulin,** Adjoint à la Directrice générale de la recherche et de l'innovation, Chef du SSRI, MESRI

Interventions de :

Nicolas Arnaud, Directeur INSU, CNRS
François Houllier, Président-Directeur général, Ifremer, Président d'AllEnvi
Lionel Suchet, Directeur général délégué, CNES
Virginie Schwarz, Présidente-Directrice générale, Météo-France
Philippe Mauguin, Président-Directeur général, INRAE *
Valérie Verdier, Présidente-Directrice générale, IRD
François Jacq, Administrateur général, CEA
Christophe Poinssot, Directeur général délégué, BRGM
Magali Stoll, Directrice générale adjointe, IGN
Gauthier Hulot, Directeur adjoint, IPGP

10h00-10h45 **Présentation du projet GAIA Data**
Frédéric Huynh, Responsable scientifique et technique du projet GAIA Data, Directeur de l'IR DATA TERRA, Pôles de données et services pour le système Terre
Sylvie Joussaume, Directrice de l'IR CLIMERI-France de modélisation du climat
Jean-Denis Vigne, Coordinateur de l'IR PNDB, Programme National de Données Biodiversité

PAUSE

11h15-11h40 **Présentation générale de l'infrastructure technique intégrée de données et de services**
Richard Moreno (CNES), coordinateur technique du projet GAIA Data

11h40-12h30 **Table ronde : infrastructures distribuées : synergies, partenariats et mutualisation**
GAIA Data, Meso-Net, FITs, GENCI, CINES, IDRIS
Modérateurs : **Laurent Crouzet** (MESRI), **Sylvie Joussaume** (CNRS, CLIMERI)
Intervenants : **Stéphane Requéna** (GENCI), **Jean Pierre Vilotte** (INSU-CNRS), **Boris Dintrans** (CINES), **Pierre François Lavallée** (IDRIS), **Arnaud Renard** (U. Reims, projet Mesonet) et **Pierre Etienne Macchi** (IN2P3, projet FITS)

14h00-14h35 **Présentation de l'infrastructure technique intégrée de données et de services : renforcement des équipements et interconnections des sites**
Karim Ramage (CNRS, IPSL), coordinateur technique adjoint projet GAIA Data

14h35-15h15 **Table Ronde : Science ouverte, EOSC, Copernicus, Destination Earth**
Modérateurs : **Isabelle Benezeth** (MESRI), **Caroline Blanke** (IR Data Terra)
Intervenants : **Volker Beckmann** (MESRI/DGRI, Chargé de Mission EOSC), **Alessandro Rizzo** (IRD), **Isabelle Blanc** (MESRI/DGRI-DGSIP), **Alain Arnaud** (Mercator International), **Richard Moreno** (CNES)

15h15-15h45 **Table Ronde : Interfaces avec les Infrastructures de recherche d'observation : dispositifs de collecte de données**
Modérateurs : **Gilbert Maudire** (Ifremer), **Jean-Denis Vigne** (MNHN/CNRS)
Intervenants : **Isabelle Braud** (INRAE), **Paolo Laj** (CNRS), **Michael Chelle** (INRAE), **Lucie Cocquempot** (IFREMER), **Raphael Pik** (CNRS)
Et avec le concours des coordinateurs, directrices et directeurs des pôles de données thématiques

PAUSE

16h15-16h45 **Table ronde : Partenariats entre les projets EQUIPEX+ du domaine système Terre et environnement** : GAIA Data, e-COL+, Marmor, OBS4CLIM, Terra Forma
Modérateurs : **Anne Puissant** (UNISTRA), **Michel Diament** (IPGP)
Intervenants : **Isabelle Braud** (INRAE, Terra Forma), **Pierre-Yves Gagnier** (MNHN, e-COL+), **Hélène Leau** (Marmor), **Paolo Laj** (CNRS, OBS4CLIM)
Et avec le concours des coordinateurs, directrices et directeurs des pôles de données thématiques

16h45-17h00 **Bilan et synthèse**
Sylvie Joussaume, Directrice de l'IR CLIMERI-France de modélisation du climat
Jean-Denis Vigne, Coordinateur de l'IR PNDB, Programme National de Données Biodiversité
Frédéric Huynh, Responsable scientifique et technique du projet GAIA Data, Directeur de l'IR DATA TERRA, Pôles de données et services pour le système Terre

SOMMAIRE

01

E-Infrastructures de Recherche :

- IR PNDB
- IR CLIMERI-France
- IR DATA TERRA

TROIS E-INFRASTRUCTURES DE RECHERCHE DU DOMAINE SYSTÈME TERRE ET ENVIRONNEMENT

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 produit des simulations numériques internationales pour le Programme Mondial de Recherche pour le Climat et met leurs résultats à la disposition de divers utilisateurs en France et à l'étranger.

PNDB propose des outils & services pour accompagner et faciliter la compréhension, le partage et l'utilisation des données de biodiversité produites pour et par les communautés de recherche.

POLE NATIONAL DE DONNÉE ET BIODIVERSITE

Créé en 2018, renouvelé sur feuille de route MESRI 2022

Etablissement porteur de l'infrastructure :

- Muséum national d'Histoire naturelle (UMS PatriNat)

18 partenaires de l'infrastructure :

8 Organismes

10 Universités

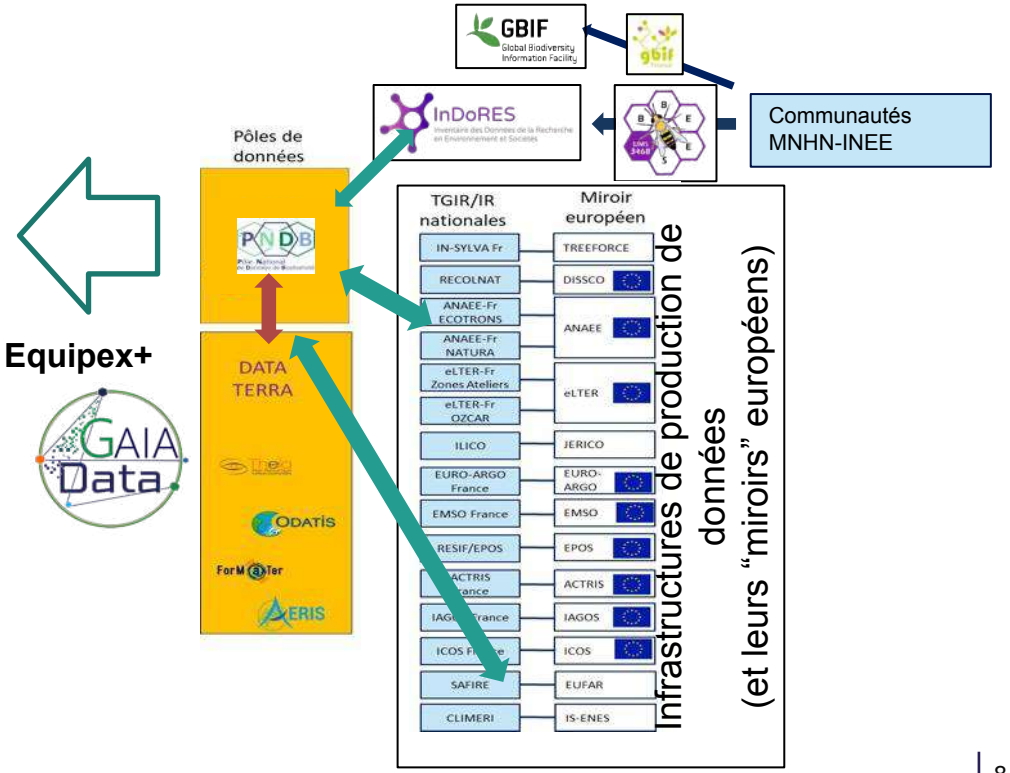
1 Fondation pour la Recherche sur la Biodiversité

POSITIONNEMENT ET FONCTIONNEMENT DU PNDB

Services à la recherche :

- Coordination entre les entrepôts de données (interopérabilité, métadonnées)
- Accès aux (méta)données de biodiversité ouvertes
- Outils de FAIRisation des (méta)données
- Outils de gestion de croisement et d'analyse des données
- Accompagnement et animation des communautés

Cadre scientifique intégratif :
temps long, tous milieux, de la molécule aux anthroposystèmes ; pressions anthropiques (lien avec les politiques publiques à travers le SIB)



CLIMERI-France
Infrastructure de recherche nationale
de modélisation du climat

Missions :

- **Réalisation des simulations internationales du WCRP avec les deux modèles de climat français: IPSL et CNRM-Cerfacs**

  Comprendre / Evaluer / Prévoir - Global (CMIP) & Régional (CORDEX)

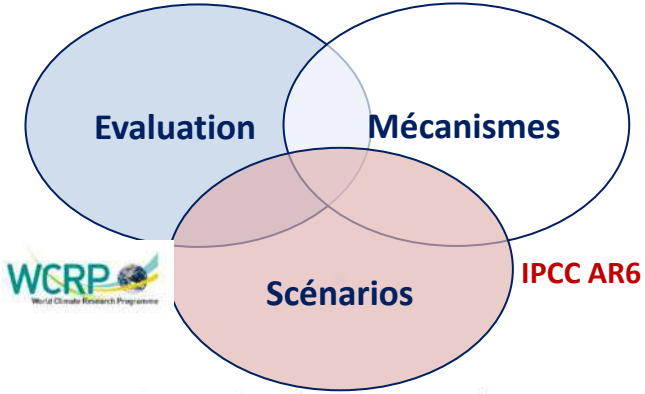
- **Réalisation des simulations de référence sur la France**
- **Mise à disposition des résultats pour diverses communautés**

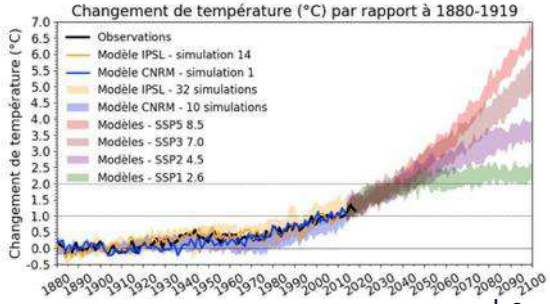
  Sciences du climat, Impacts, Services climatiques, Copernicus C3S,

  Rapports du GIEC

Feuille de route nationale depuis 2016

CMIP6 Coupled Model Intercomparison Project Phase 6





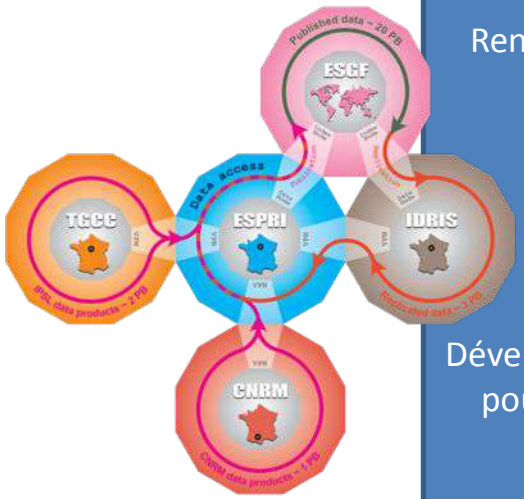
En collaboration avec
SU, IRD, Cerfacs

https://climeri-france.fr/

De la production des simulations à la diffusion des données des simulations de référence National / Europe / International

ESGF
> 15 000 utilisateurs
30 Po de données
Nœuds EU:
530 TB/mois (2021)

- Animation & stratégie scientifique
- Modèles de référence
- Calcul & simulations de référence
- Stockage & analyse multi-modèles
- Diffusion des données & interface utilisateurs



Projet Gaia-Data:

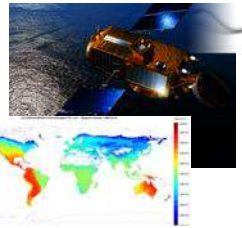
Renforcer les capacités de traitement des données de simulations

Mieux intégrer l'accès aux simulations climatiques et aux observations

Développer des environnements virtuels pour faciliter les traitements croisés de données

PÔLES DE DONNÉES ET SERVICES POUR LE SYSTÈME TERRE

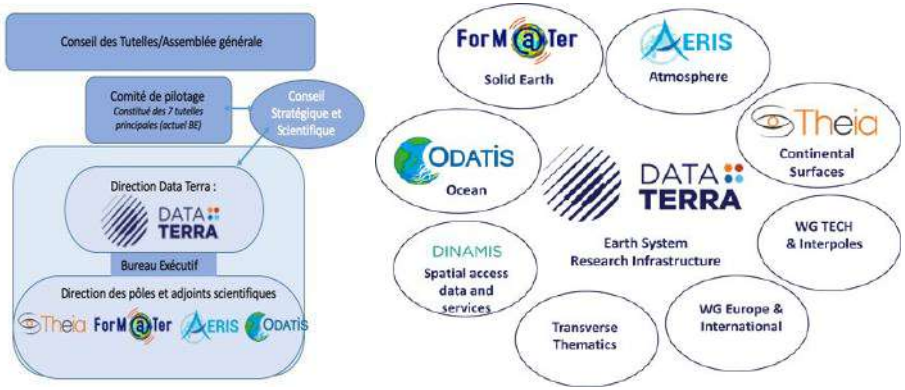
Développer et mettre en œuvre **une infrastructure/plate-forme intégrée de données et de services distribuées pour l'observation et la compréhension du système terre et de l'environnement sur l'ensemble du cycle de la donnée**, de son **acquisition** (spatiale, sols, in-situ) jusqu'à ses **multi-usages**



- **Faciliter l'accès et l'utilisation** des **données et produits** de qualité sur l'ensemble des **compartiments du système Terre (Données spatiales, aéroportées, sols, in-situ)**
- **Développer des services de visualisation et de traitements adaptés aux besoins, à l'accroissement de la volumétrie et aux avancées technologiques**
- **Favoriser la mutualisation, interopérabilité,** émergence d'**approches multi- et inter-disciplinaires**
- **Servir** les communautés **scientifiques**, les acteurs de l'**action publique et de l'innovation**
- **Mettre en œuvre une stratégie nationale,** européenne et internationale

PÔLES DE DONNÉES ET SERVICES POUR LE SYSTÈME TERRE

Créé en 2016, feuille de route IR MESRI 2018, 2021



Data Terra est fondée sur quatre pôles correspondant à chacun des grands compartiments du Système Terre : surfaces continentales, atmosphère, océans et terre solide, complétés par des services transverses.

- 26 organismes et universités
- 4 pôles de données
- 6 services (DINAMIS) et groupes de travail transversaux
- 30 Centres de Données et de Services (CDS) et Infrastructures de données spatiales (IDS)
- 25 Consortium d'Expertise Scientifique
- 200 ETPT / 450 scientifiques, ingénieurs et techniciens
- 33 M€ (2016), 39 M€ (2017), > 40 M€ (2019, 2020)
- Plus de 500 produits et services, plus de 15000 utilisateurs réguliers
- 50 (2018) ; 10[illegible]0 To (2023) ; 150 Peta (2025)

Positionnement des E-Infrastructures de recherche domaine Système Terre et environnement

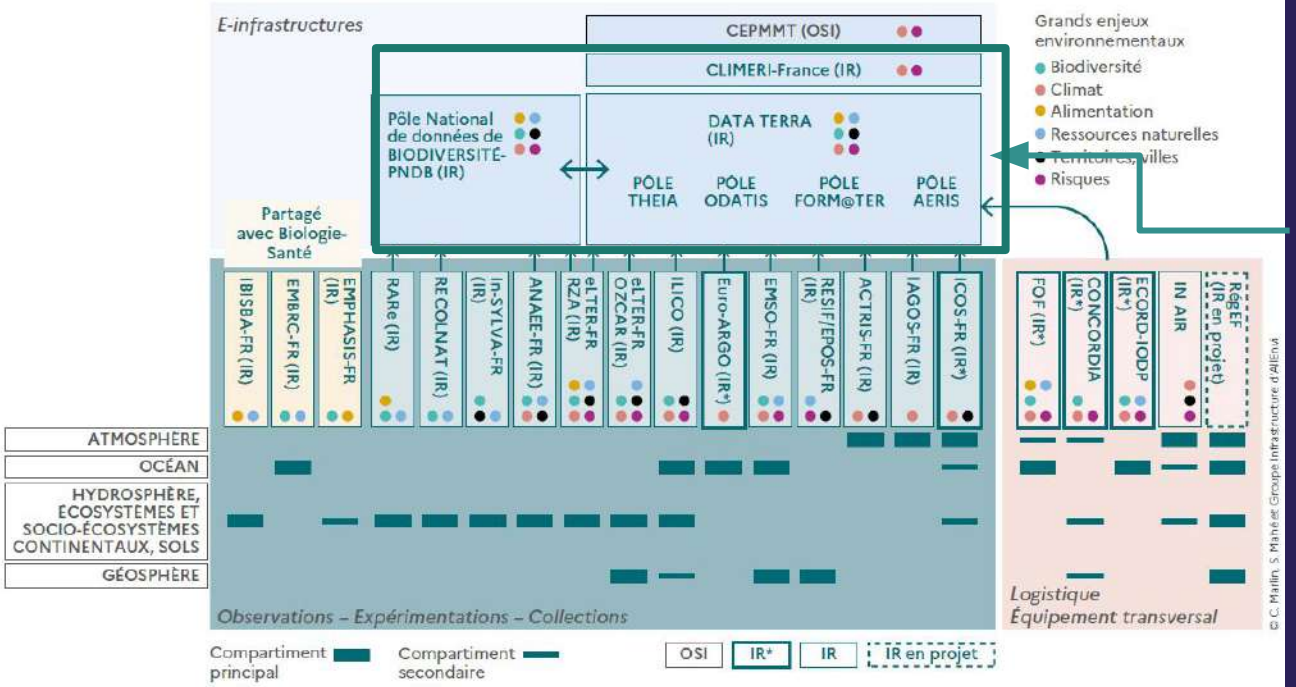


Figure 1 : OSI/IR*/IR du domaine SST & ENV par grands types (observations – expérimentaux – collections, logistiques et e-infrastructures), par grands compartiments du système Terre (atmosphère, océan, hydrosphère–écosystèmes – socio-écosystèmes continentaux – sols et géosphère) et par grands enjeux environnementaux (biodiversité, climat, alimentation, ressources naturelles, territoires-villes, risques). En jaune figurent trois infrastructures partagées avec le domaine Biologie – Santé.

02

MISSIONS, OBJECTIFS ET ENJEUX

GAIA Data | EQUIPEX+ /PIA3

Porté par 3 Infrastructures de Recherche

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,** 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**

Equipement Structurant pour la Recherche / EQUIPEX+

AXE NUMERIQUE: 168 M€ **(29 %)** — **12 projets (23%)**
AXE GENERIQUE: 404 M€ **(71 %)** — **40 projets (77%)**
=> 572 M€ (**52 projets A+/A)**

GAIA Data, classé A + ; avis très positif du jury international

« holistic view of the Earth system requires such databases and high-speed data exchange » ; « great impact on Earth science and all its disciplines » « data potential for both the scientific and industrial sectors » « massive advantage in developing a data management infrastructure of the magnitude » ; « potential to generate major benefits for the scientific and technical community and with the excellent outreach ambitions expected, economic and socioeconomic benefits »

Budget total : 62 M€ (coûts complets)
Demande ANR-EQUIPEX+PIA3 : **16,2 M€**
Forte contribution RH : **339 ETP** (soit 4066 p.m.) personnels permanents + **59 ETP** (711 p.m.) cdd

Apports additionnels des organismes : **25 postes** (recrutements cdi, postes permanents, mobilités)

21 Partenaires CNRS, 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

CONTEXTE ET ENJEUX

Des systèmes d'information existent : IR **DATA TERRA** pour les données d'environnement, **PNDB** pour la 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**
- ➔ Concilier / s'intégrer / influencer / contribuer aux dynamiques régionales, nationales, européennes et internationales

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, expérimentation et des données issues de simulations de référence
- ➔ **Co-construire**, organiser et adapter les services **avec et pour les communautés scientifiques** du domaine système Terre et environnement, les **acteurs publics et socio-économiques**

AMBITION ET OBJECTIFS

- Travailler, en étroite relation avec l'IR* GENCI, les centres nationaux, CINES, IDRIS, CCIN2P3 et centres de données régionaux labelisés/Meso-Centres

- Renforcer les synergies et collaborations avec les IR/IR* d'observation (Terre Solide, Atmosphère, Océan, Surfaces continentales, biodiversité, …) et IR Numérique

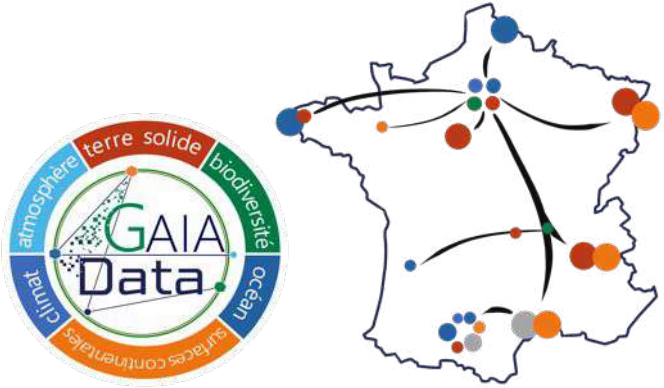
- Contribuer à la souveraineté des données et des connaissances scientifiques et technologiques (préservation des connaissances ; maîtrise de la chaine de valeur-ajoutée : données – informations – connaissances)

- Contribuer aux initiatives nationales (science ouverte, Infranum,…), européennes (EOSC, Copernicus, DTE, …) et internationale (GEO, GoFAIR, ONU, coopération Pays du Sud…)

- Participer à la mise en œuvre des jumeaux numériques du système Terre dans le cadre Destination Earth

- Mettre en œuvre des modèles de partenariats permettant d'associer les acteurs publiques, de l'innovation, du secteur privé et de la société

Infrastructure distribuée de services



Mettre en œuvre au plan national, européen et internationale une infrastructure distribuée de services innovante du domaine système terre et environnement

COLLABORATIONS FORTES AVEC DES PROJETS EQUIPEX+/PIA3 et PEPR

Domaine système Terre

OBS4CLIM : Atmosphère (ACTRIS, ..)

TERRA FORMA : Surfaces continentales l'IR OZCAR et RZA

MARMOR : Terre Solide

E-COL+ : Collections

Domaine numérique

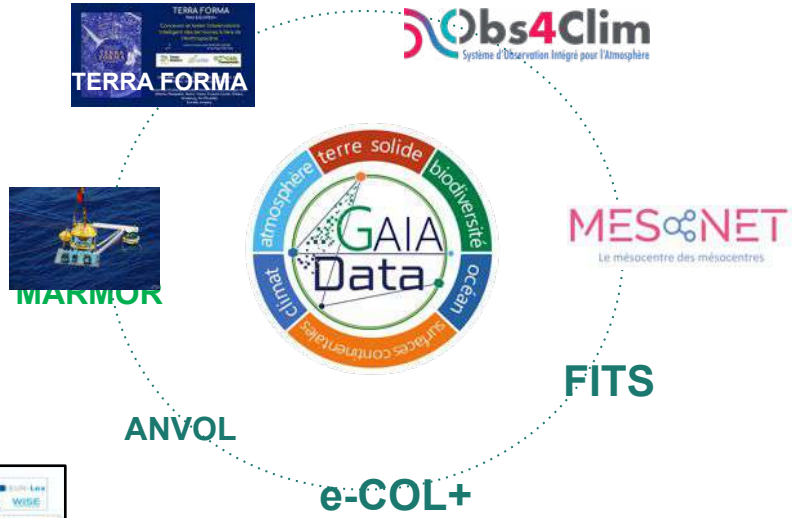
MESO-NET : porté par GENCI et Meso-Centre

FITS : CNRS (IN2P3 et INS2I)

PPR et PEPR

OneWater ...

Consolider les relations entre les IRs



03

CARACTÉRISTIQUES DU PROJET

UNE INFRASTRUCTURE FORTEMENT MUTUALISÉE

INTERCONNEXION

Basée sur des équipements, ressources et infrastructures **existants interconnectés** et **renforcés**

NOUVELLES FONCTIONNALITÉS

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

BIG DATA "5V" : Volume, Vitesse, Variété, Véracité et Valeur

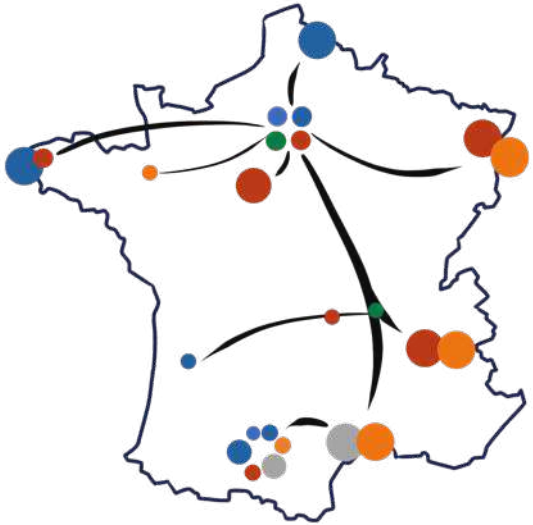
Facilite l'exploitation de **gros volumes de données** et la **génération à la demande** d'informations **combinant des données et des produits d'origines multiples**, des **satellites** aux **observations** sol, d'expériences de laboratoire ou de terrain aux **sorties de modèles**

AVANCÉE MAJEURE : 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, des acteurs publiques et de l'innovation

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

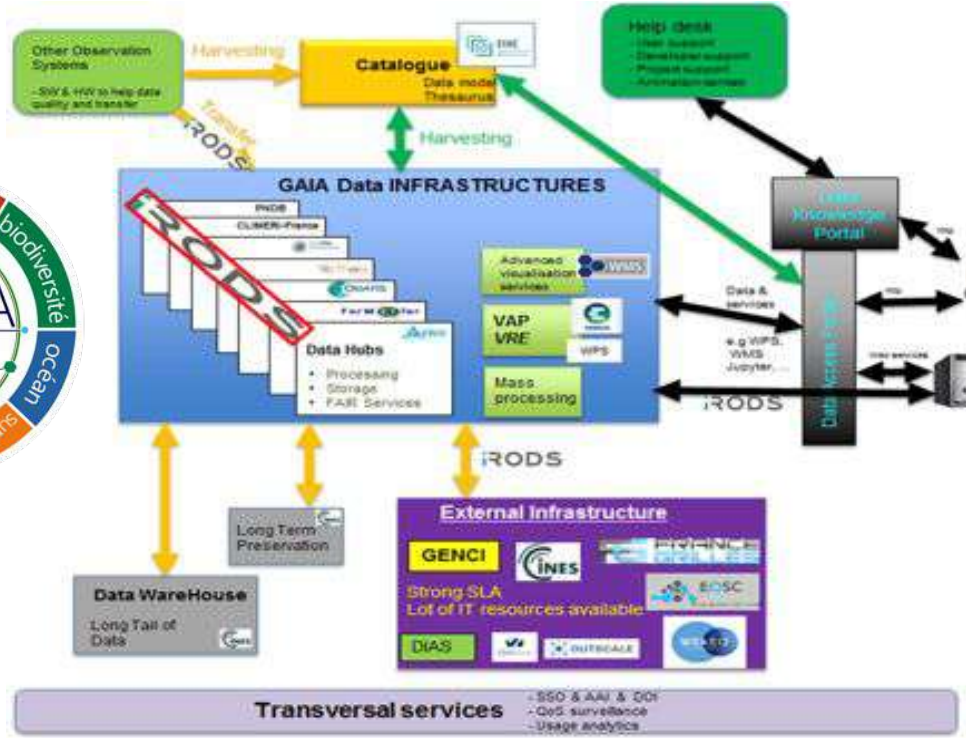
Ce projet dotera la France d'une capacité inédite qui confortera ainsi son positionnement européen et international (EOSC, Copernicus, Destination Earth, GEO, …)

UNE INFRASTRUCTURE DISTRIBUÉE DE SERVICES



8 sites principaux

30 sites existants



INFRASTRUCTURE GAIA DATA

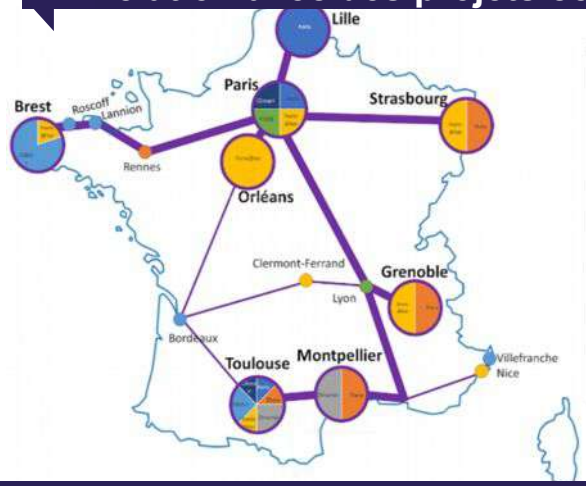
Grille de données et de services : 8 principaux centres en réseau

Projets Equipex+ ou PIA4 infra
- FITS
- MesoNet
- Clusster

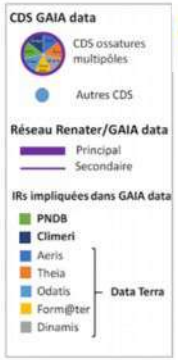
Projets Equipex+ ou PEPR thématiques
- Obs4Clim
- TerraForma
- Marmor
- OneWater

Projets H2020 – Horizon Europe
- IS-ENES
- PHIDIAS
- EOSC-Pillar
- FAIR EASE
- FAIR IMPACT

Projets CPER en région



relation avec des projets connexes



Intégré dans le paysage international / Européen

- **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
- **Interopérabilité des traitements** entre les 8 centres de Gaia Data, avec les centres HPC en France et avec les clouds commerciaux (DIAS – OVH – Orange, …)

SERVICES GAIA DATA

Services découverte, Accès et Gestion de données

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

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

DOI, **Services avancés de visualisation**
Accompagnement des communautés pour la FAIRisation

Services transversaux
pour faciliter les travaux transdisciplinaires

Grille de données, **cloud**, **portail connaissances**, SSO, Métriques, support utilisateurs & formation – animation communautés

Support aux campagnes

Analysis Ready Data Datacubes, ...

Earth Analytics Lab
exploration de la donnée, bac à sable

Virtual Analysis Platform - VAP : écosystème Notebook/PANGEO/STAC

Capacité à se connecter directement sur les centres via ssh ou autre Datacubes

Traitements à la demande (WPS)

NoCode : Galaxy-E, FG/VIP, ~Matlab/Simulink

Services de production réguliers

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

Supporté sur un continuum d'infrastructures partagées

LES SERVICES GAIA DATA : réutilisation, intégration, couplage

User Services

Findable Accessible Interoperable Reusable

Discovery, Knowledge & Services — PHIDIAS, EOSC

Earth Analytics Lab — HORIZON EUROPE

FAIR Workshop — FAIR IMPACT, GEO

Rooting Processing

User helpdesk

Communities animation

Back Office

Thesaurus

Federated and harmonized catalogues & API

Harvesting Transformation

Software repository

Identity & Access management

Secu-rity

Hypervision metrics

Machine Actionable DMP

Existing services

Individual catalogues, thesaurus and APIs — *PNDB*, *DATA TERRA*

Local helpdesks — *CLIMERI-France*

Data Warehouse

Data & Service Centers — iRODS

Long-term archives

Distributed processing

Distributed datalake

Hardware

NETWORK — RENATER

GRID — PHIDIAS

HARDWARE — Contrats de plan ÉTAT-RÉGION

04

GOUVERNANCE ET ORGANISATION

EQUIPEX+ /PIA3

RÉPARTITION BUDGÉTAIRE DE LA SUBVENTION

WP1 : Project management
501 K€

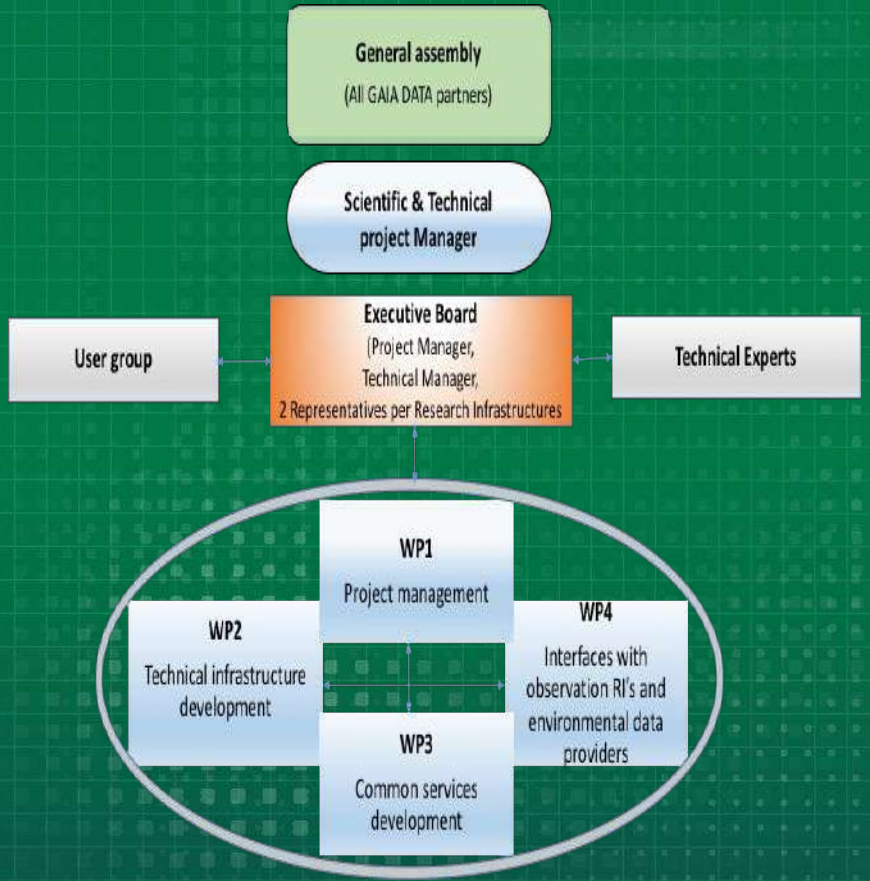
WP2 : Infrastructures techniques et développement
9,4 M€

16,16 M€

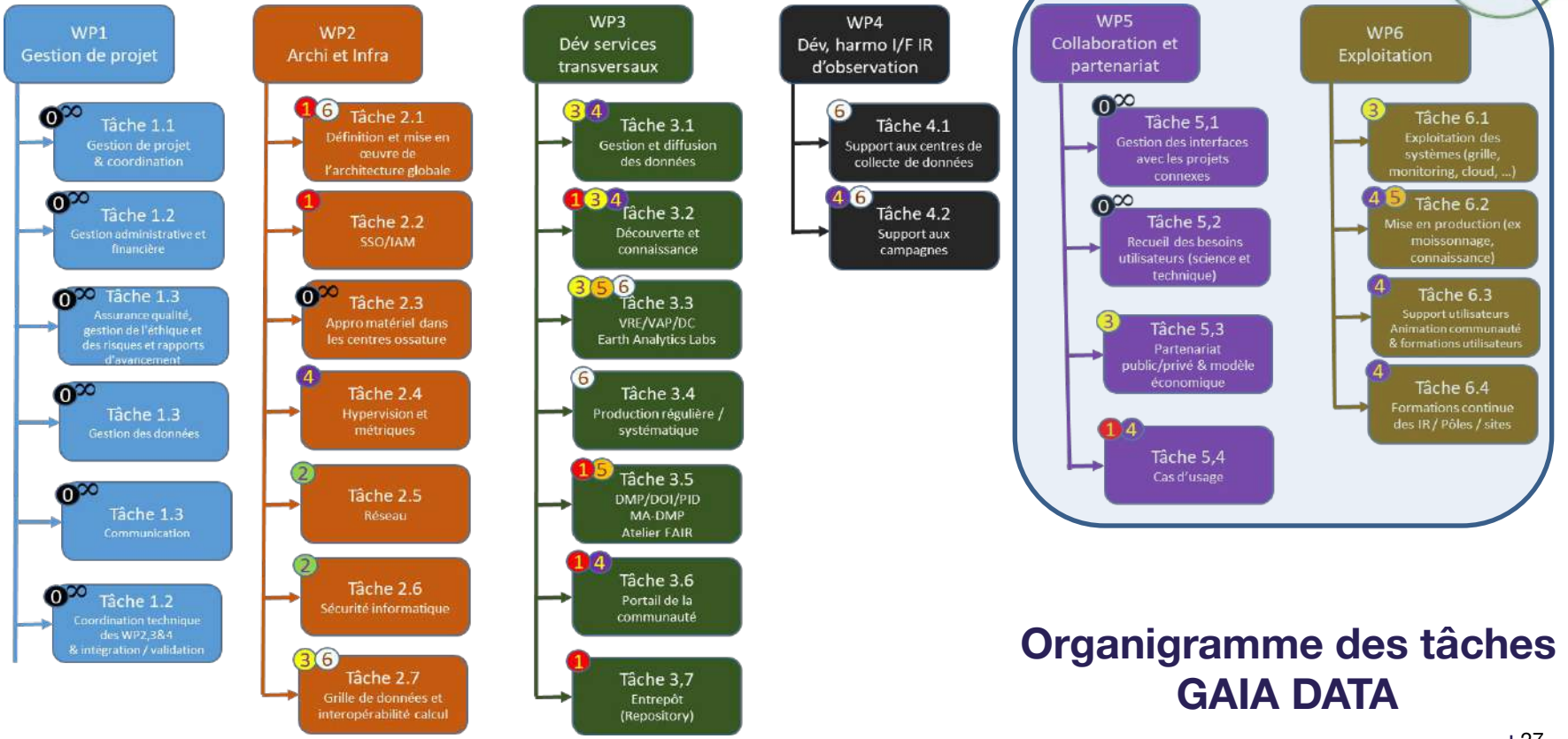
WP3 : Développement des services communs
4,5 M€

WP4 : Interface avec les IRs et producteurs de données
531 K€

GOUVERNANCE

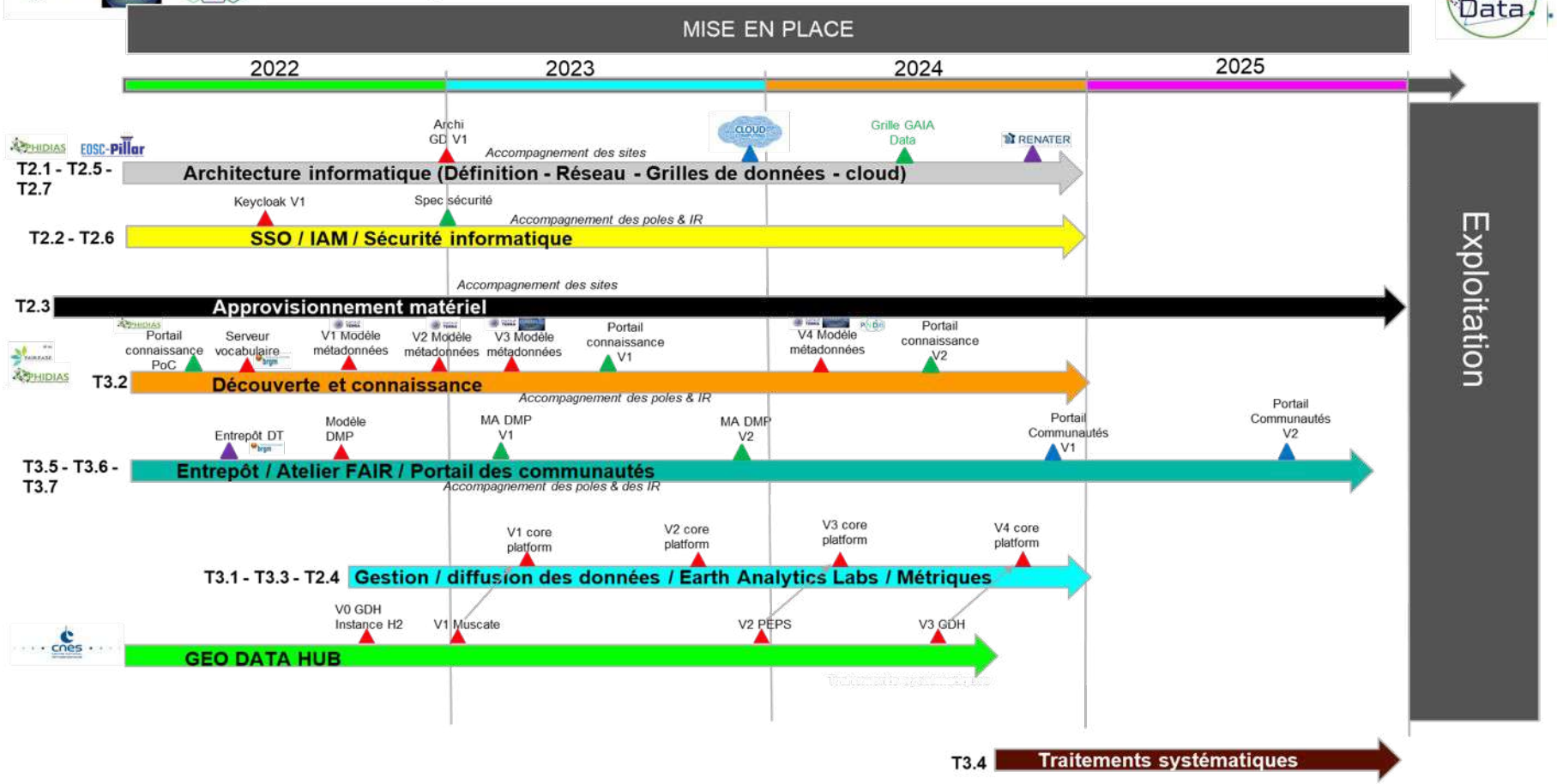


Organigramme des tâches GAIA DATA



Organigramme des tâches GAIA DATA

Planning GAIA DATA



Bilan et perspectives

- Convention ANR – CNRS signée le 12 décembre 2021
- 10 conventions de reversement de fonds signées entre l'établissement coordinateur (CNRS) et les 12 partenaires => fin avril 2022
- Accord de consortium bien avancé pour diffusion prochaine (avril/mai)
- Démarrage des premières dépenses et recrutements des CDD, CDI
- Réunion de l'AG de GAIA Data le 21 mars 202
- Réunion de lancement ANR du projet le 21 mars 2022
- Affectation de nouvelles compétences (IRD, CNES, …)
- Discussions avec tous les partenaires pour analyser engagements et attentes
- Analyse et identification de projets pilotes eet use-cases pour tester, dimensionner et répondre aux besoins

Le projet est en ordre de marche

Bénéficie des engagements de tous

Recherche de projets européens pour cofinancer certaines composantes du projet

Analyse modèles économiques






contact@gaia-data.org

www.gaia-data.org

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+.