Equipex+ Gaia Data

Infrastructure distribuée de données et services : observation et modélisation intégrée du système Terre

Journées Climeri
03.02.2022

Le Projet Gaia Data

Porté par 3 Infrastructures de Recherche numériques du domaine
« système Terre et Environnement »
Data Terra (données observations du système Terre),
CLIMERI (données simulations climatiques),
PNDB (données biodiversité)

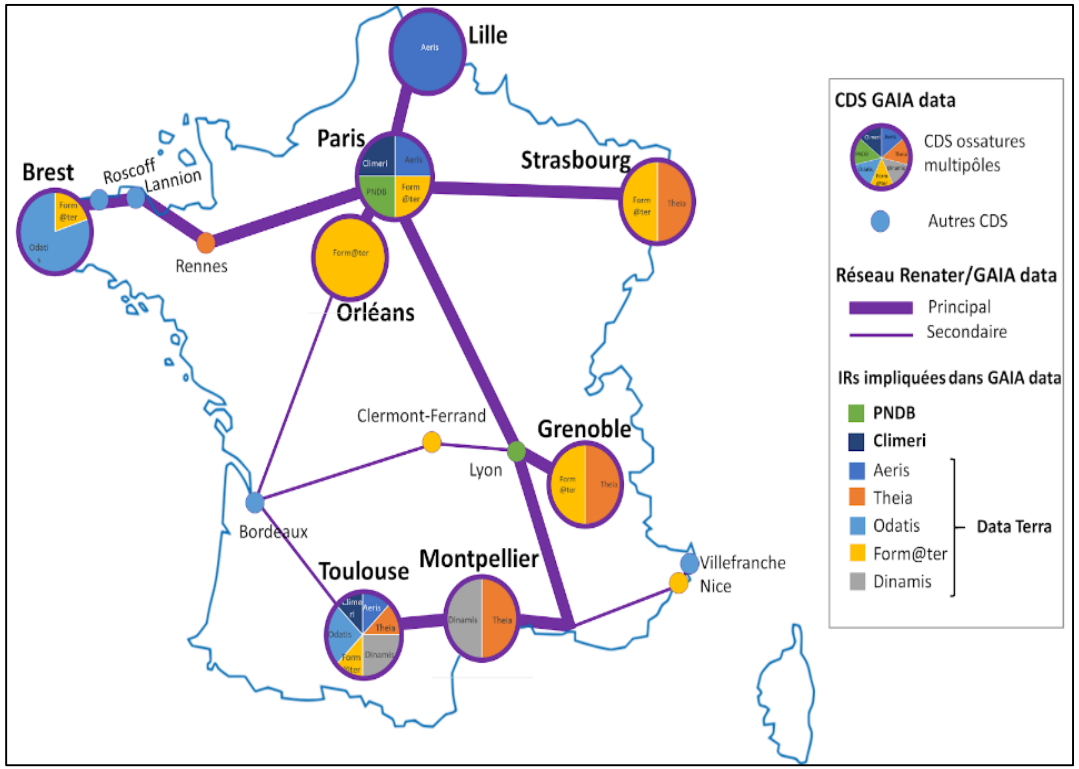
21 Partenaires : CNRS (coord.), CNES, IFREMER, IRD, BRGM, IGN, INRAE, Météo-France, MNHN, CEA, IPGP, CINES, Sorbonne Univ., Univ. Grenoble-Alpes, Univ. Lille, Univ. F. Toulouse, UNISTRA, SHOM, OCA, FRB, CERFACS

Objectif :

- **Développer et mettre en œuvre une 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** (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**

Infrastructure numérique Gaia Data



Construit autour des 8 principaux centres de calcul et de données des 3 IR :

- Centre de Calcul Nationaux (CINES, IDRIS)
- Centres de Calcul et données d'organismes (CNES, Ifremer, BRGM)
- Mésocentres Régionaux (GRICAD, UniStra, Univ Lille, Meso@LR)
- Mésocentres Thématiques (ICARE, ESPRI, IPGP-Dante)

Développement d'une Grille de données et de services :

- **Mise en place d'un réseau dédié haut-débit et sécurisé** entre les 8 centres principaux
- **Déploiement d'une grille de données (système iRODS AC) / S3** 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 8 centres de Gaia Data, avec les centres HPC en France et avec les clouds commerciaux (GAIA-X - DIAS)

Intégré dans le paysage international / européen

Destination Earth

Clouds publics (DIAS, …)

En relation avec des projets connexes

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

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

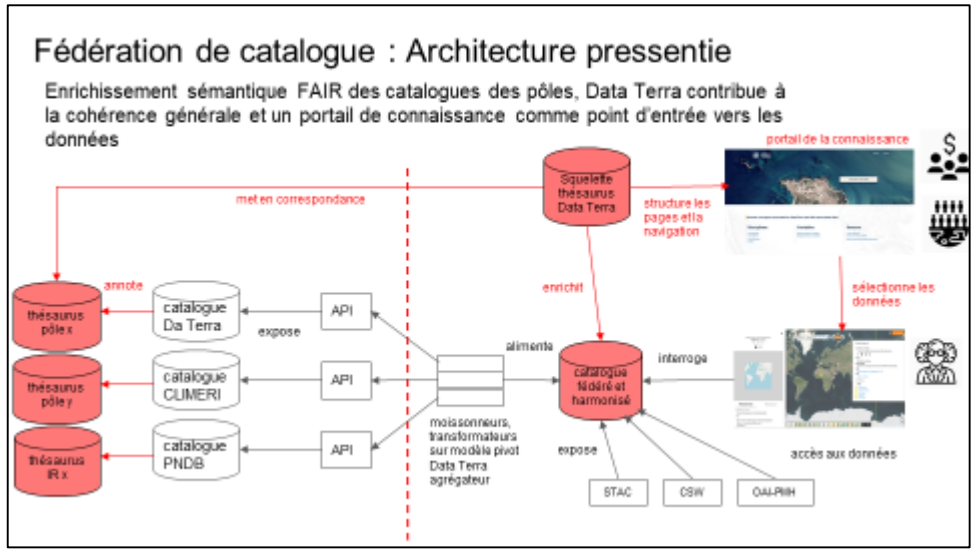
Projets H2020 – Horizon Europe
- IS-ENES
- PHIDIAS
- EOSC-Pillar
- FAIR EASE
- FAIR IMPACT

Projets CPER en région

Les services Gaia Data

Services Découverte, Accès et Gestion des données

- **Catalogue** (métadonnées, vocabulaires, ontologies), systèmes d'accès et de recherche
- Consultation et accès aux données via web services (INSPIRE, Opensearch, STAC, intake …)
- **DOI**, **Services avancés de visualisation**
- Accompagnement des communautés pour la FAIRisation
- Aide à la collecte des données des observatoires



Services transversaux => faciliter les travaux transdisciplinaires

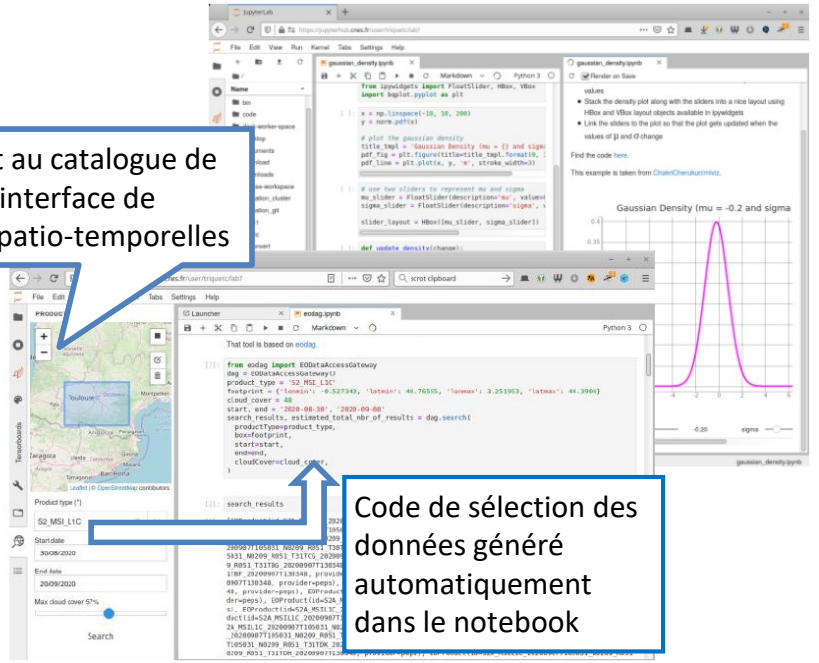
- **Grille de données**, **cloud**, **portail connaissances**, SSO, Métriques, support utilisateurs & formation – animation communautés
- Support aux campagnes
- Analysis Ready Data ⇔ Datacubes, …

Earth Analytics Lab : exploration de la donnée, bac à sable

- VAP : Virtual Analysis Platform : 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

Nécessite de coder
~Data scientist

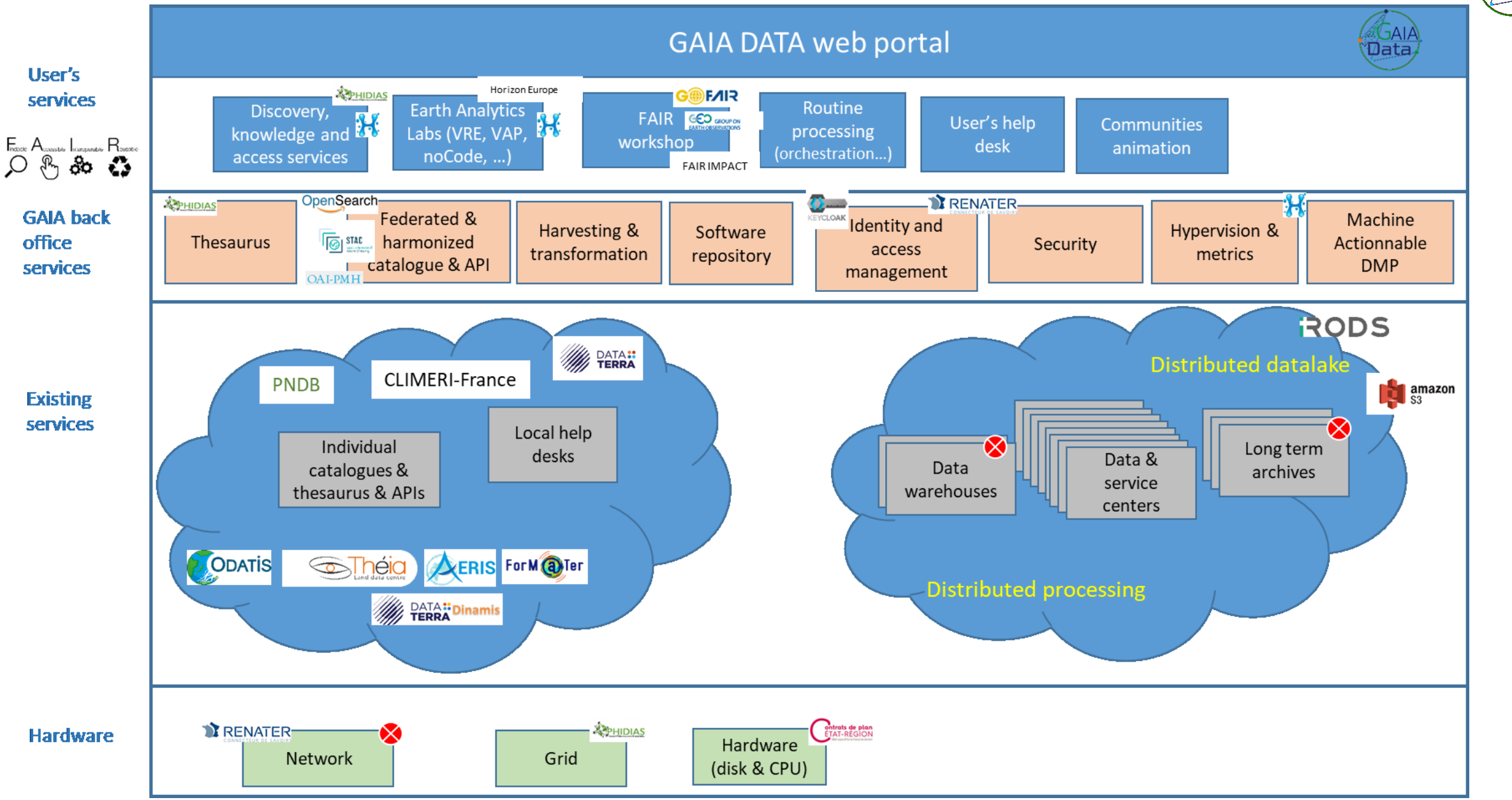
Pas besoin de coder



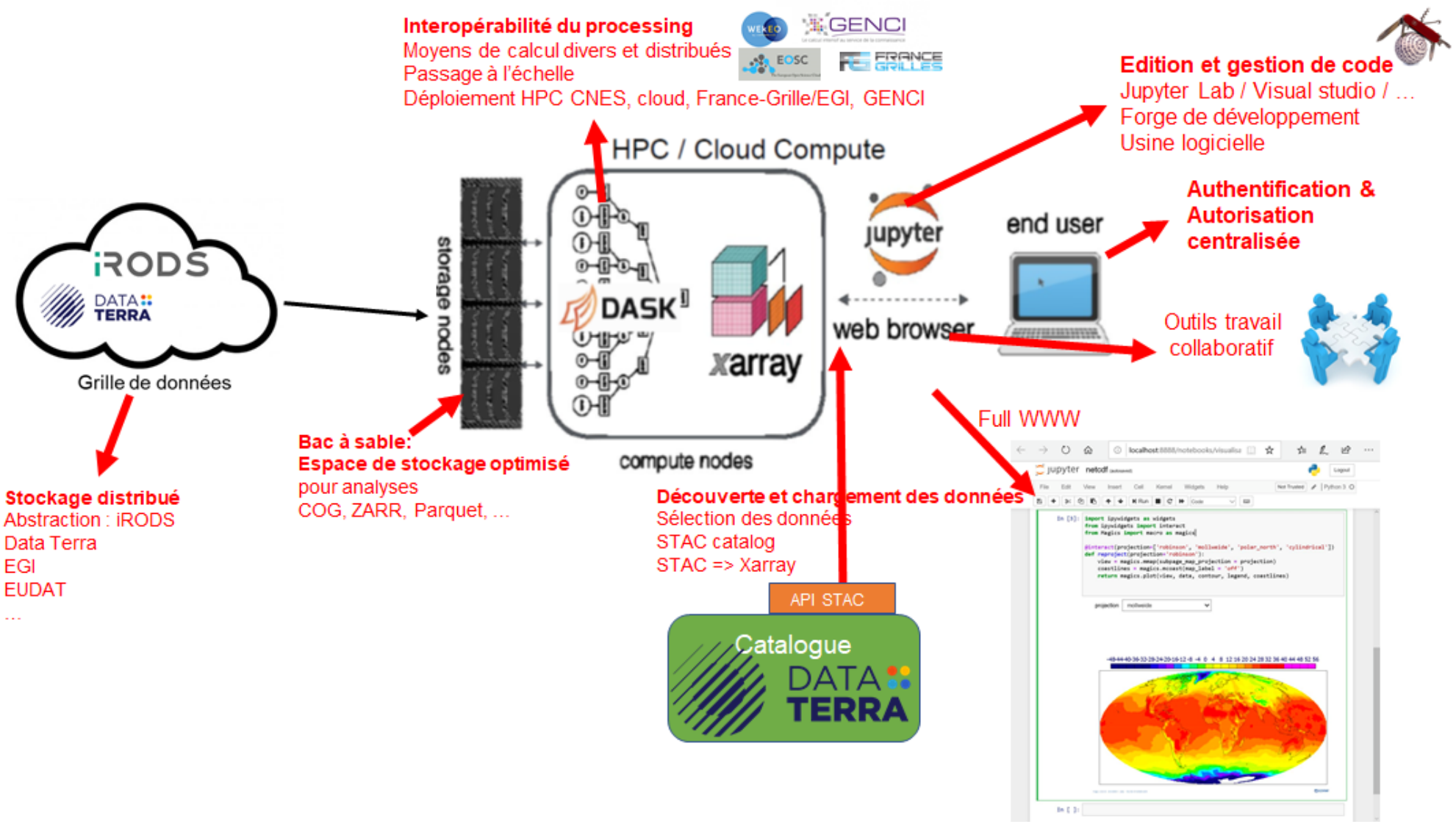
Services de production réguliers

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

Système GAIA DATA ~Puzzle A CONSOLIDER

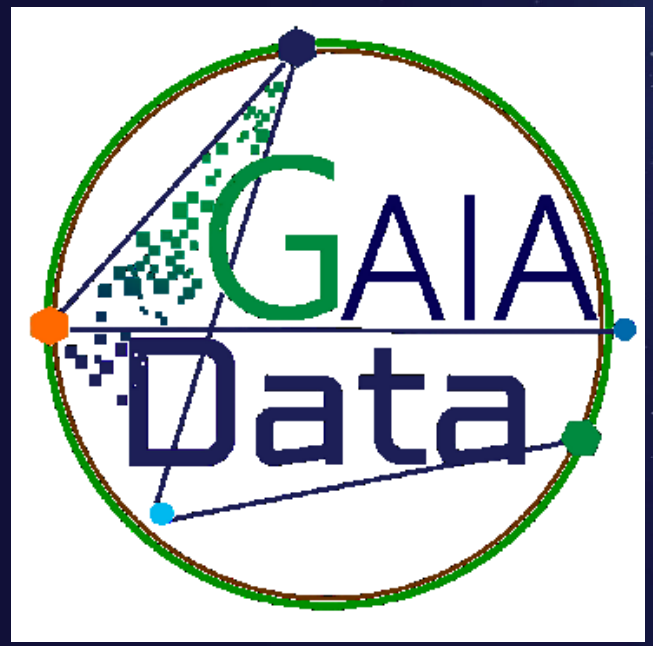


Exemple de VAP/VRE : Science Data hubIntégration Catalogue/iRODS/Pangeo/Notebooks



Source : CNES

Journées Climeri : Equipex+ Gaia Data



Le budget Gaia Data

CENTRES ou ACTIVITES	TOTAL	Réduction	Pourcentage	Reste
WP1 Management	**790 000 €**	**289 000 €**	**37%**	**501 000 €**
Fonctionnement	290 000 €	29 000 €	10%	261 000 €
CDD Projets	**500 000 €**	**260 000 €**	**52%**	**240 000 €**
WP2 RENFORCEMENT DES CENTRES OSSATURES	**12 001 724,00**	**2 613 511,05**	**22%**	**9 388 212,95**
Equipements	**11 210 124 €**	**2 432 679 €**	**22%**	**8 777 445 €**
Renforcement/Acquisition des équipements réseaux et sécurité	948 704 €	190 201 €		758 503 €
Renforcement/Acquisition des espaces de stockage d'échange iRODS	2 028 930 €	497 971 €		1 530 959 €
Renforcement des espaces de disques "rapide" pour le traitement de données	1 779 420 €	394 528 €		1 384 892 €
Renforcement des espaces "capacitifs" pour les données de référence	2 985 000 €	623 050 €		2 361 950 €
Renforcement des processeurs pour les plateformes d'analyses	1 360 070 €	290 879 €		1 069 191 €
Renforcement des processeurs graphiques pour la visualisation	644 000 €	134 810 €		509 190 €
Renforcement des plateformes de virtualisation/conteneurisation pour l'hébergement des services	1 464 000 €	301 240 €		1 162 760 €
Prestations	**791 600 €**	**180 832 €**	**23%**	**610 768 €**
Location Stockage/Calcul/Locaux sur mésocentres	376 000 €	94 272 €		281 728 €
Installation / Déploiement des équipements	415 600 €	86 560 €		329 040 €
WP3 Services Transverses	**4 809 375 €**	**266 625 €**	**6%**	**4 542 750 €**
Prestations	**2 730 000 €**	**136 500 €**		**2 593 500 €**
FAIR	1 810 000 €	90 500 €		1 719 500 €
ARCHI	920 000 €	46 000 €		874 000 €
CDD Projets	**2 079 375 €**	**130 125 €**		**1 949 250 €**
FAIR	695 625 €	33 938 €		661 687 €
ARCHI	1 383 750 €	96 188 €		1 287 563 €
WP4 CENTRES DE COLLECTE	**590 000 €**	**59 000 €**	**10%**	**531 000 €**
Fonctionnement (petit matériel)	100 000 €	10 000 €		90 000 €
Prestations	400 000 €	40 000 €		360 000 €
CDD Projets (FAIR)	90 000 €	9 000 €		81 000 €
Total sans frais de gestion	**18 191 099 €**	**3 228 136 €**		**14 962 963 €**
Total avec frais de gestion	**19 646 387 €**	**3 486 387 €**		16 160 000 €
Cible				**16 160 000 €**

Financé par AMI Equipex+ du PIA3

Classé A+

Projet sur 8 ans

- **4 ans : Phase de développement**
- **4 ans : Phase d'exploitation**
- **Implication des organismes : 38 ETP**
- **16 postes CDD (13 pérennisables)**

Budget Gaia Data : 16,1 M€

- Renforcement équipements et interconnexion des sites : 9,4 M€

- Développement des services : 5 M€

Budget Climeri dans Gaia Data

- Renforcement équipements et interconnexion des sites : 700 k€

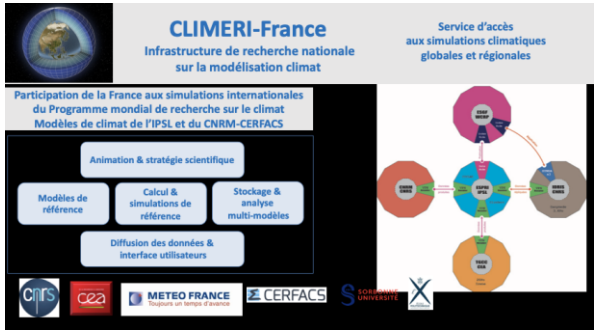
- Développement des services :

- **Personnel : 85 k€**
- **Prestation : 200 k€**

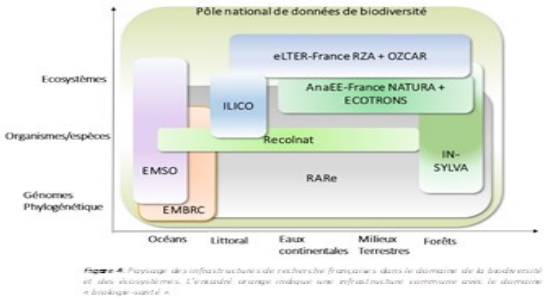
Fonctionnement : 4% des budgets alloués reversés par les organismes

Les 3 Infrastructures de Recherche

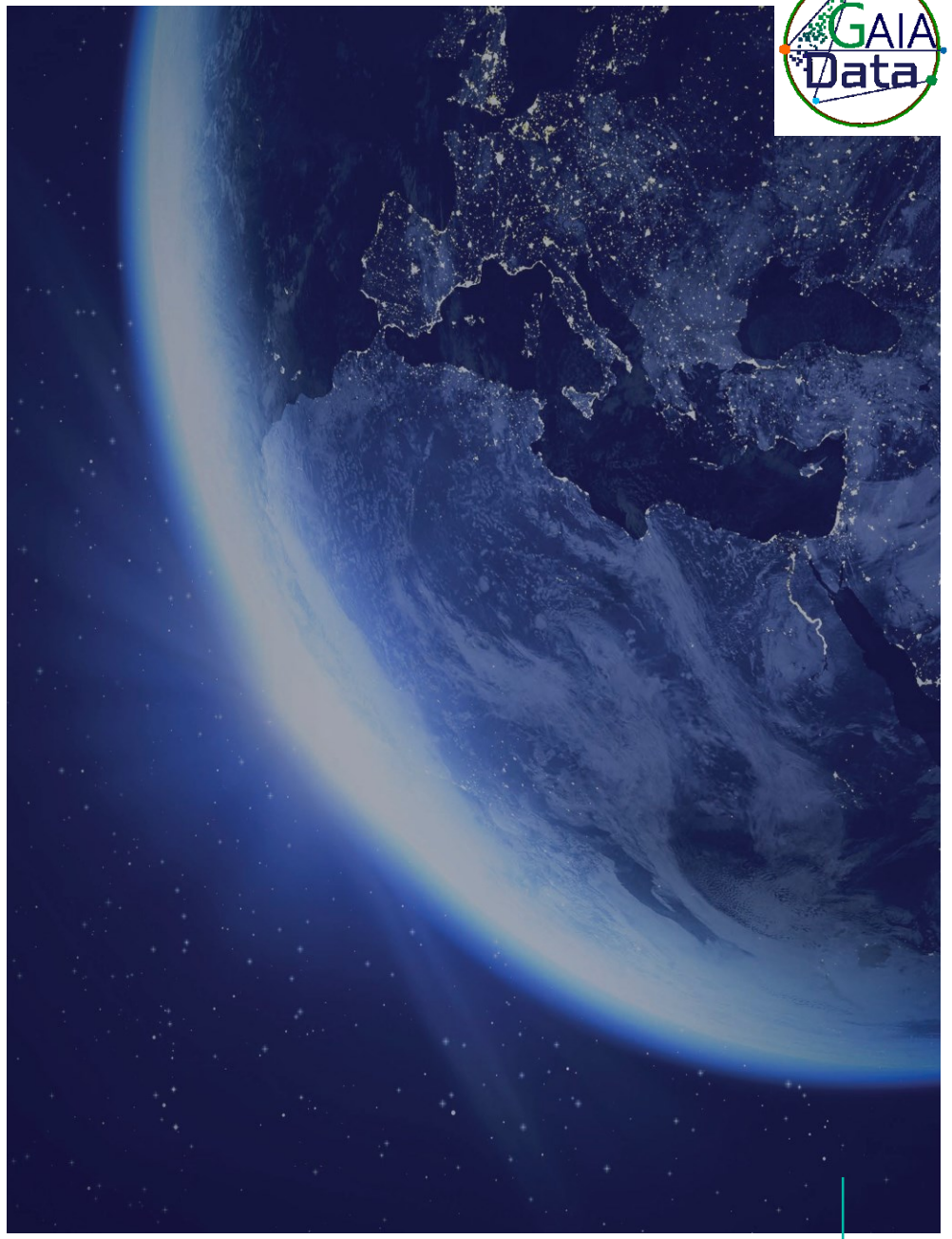
Data Terra organise l'accès intégré aux données d'observation, produits et services couvrant les différents compartiments du système terrestre et leurs interactions



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.

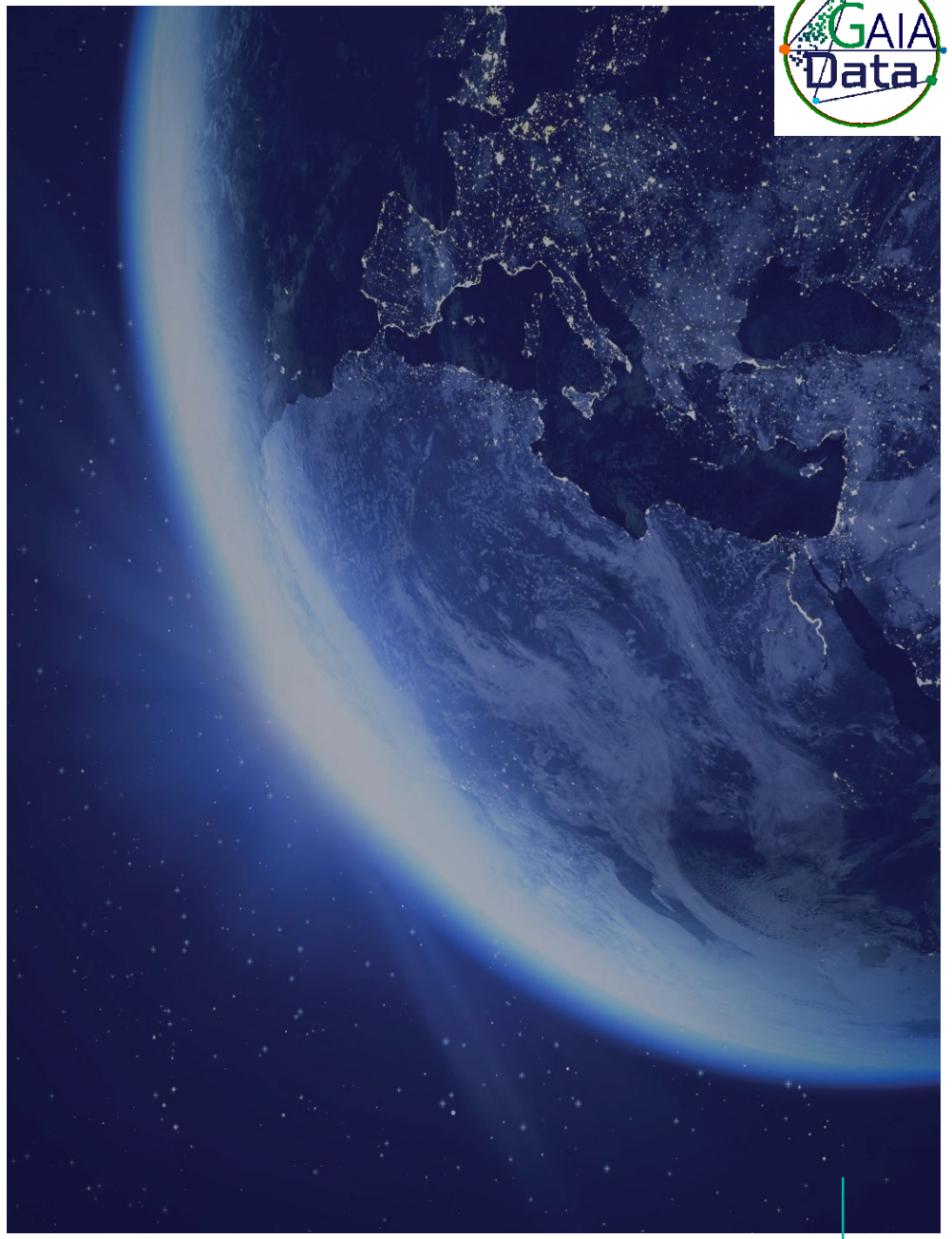


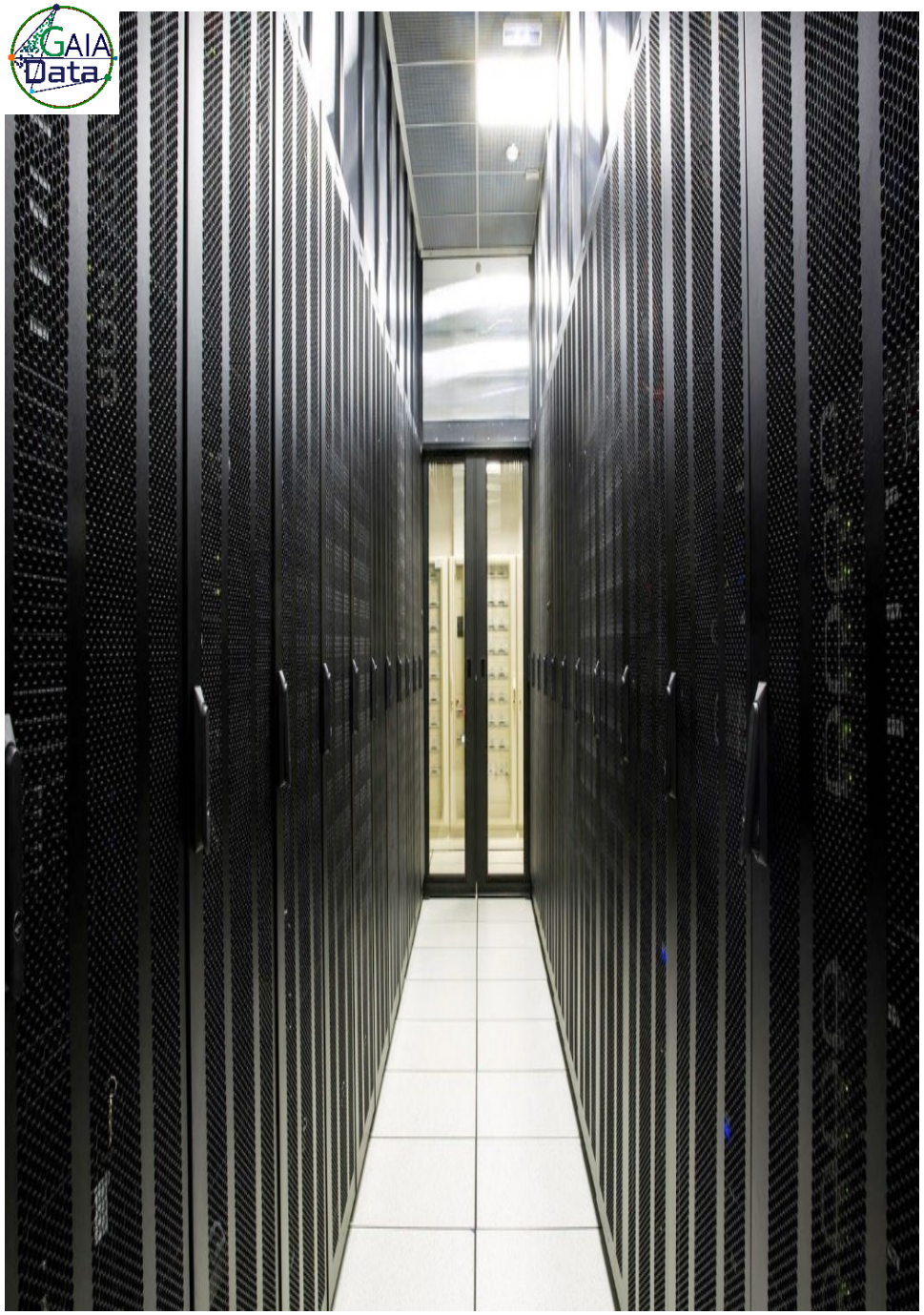
Le PNDB, le centre national de données sur la biodiversité, vise à fédérer les approches de données existantes au sein des infrastructures de recherche sur la "Terre vivante".



Réutilisation massive de solutions existantes :

- THEIA/Hysope2 : catalogue, Earth Analytics Labs
- PHIDIAS : modèles métadonnées, Connaissances
- EOSC-Pillar, FAIR EASE, PHIDIAS : Earth Analytics Labs
- EOSC-Pillar & PHIDIAS : grilles
- CNES/AI4GEO : Earth Analytics Labs
- SSO : AERIS
- Grilles : PHIDIAS, EOSC-Pillar
- PNDB : Galaxy-E
- CLIMERI-France : IS-ENES
- BRGM : animation communautés
- Projets européens à venir (FAIR EASE, FAIR IMPACT, …)





Equipements et Interconnexion des sites

Renforcer les moyens dans les Centres de Calcul et Données pour assurer les missions des pôles et IR

- Stockage des données de références
- Production et exploitation des données à valeur ajoutée
- Distribution

Développer les infrastructures pour permettre l'interopérabilité des accès aux données

- Interconnexion réseau (L3VPN, très haut débit dédié), dont CNRM-ESPRI-IDRIS
- Grille de données (iRODS : stockage virtualisé multisite fédéré, automatisation de workflow de données)
- Datalake (Analysis Ready Data + S3, openDAP, …)

Renforcer les équipements et développer les systèmes pour assurer l'interopérabilité des services entre les Centres de Calcul et Données

- Authentication and Authorization Infrastructure / Single Sign On
- Containerisation, cloud et plateforme Iaas/PaaS/IaC
- Outils de déploiements logiciels

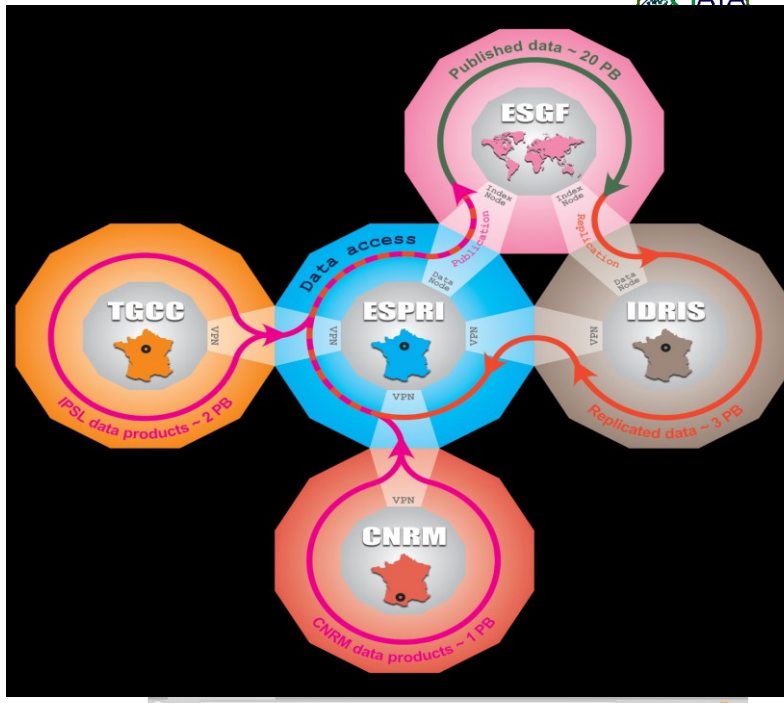
Renforcer les architectures spécialisées pour le service de la données

- Nœuds pour la visualisation, traitements à la demande (VRE, VAP)
- IA / Machine Learning

Gaia Data pour Climeri

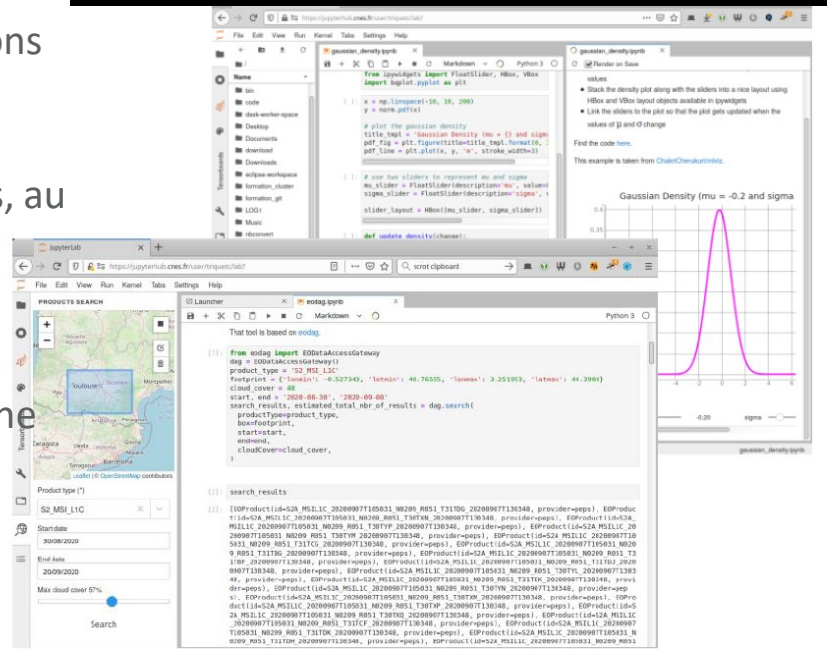
Renforcement des capacités de traitement et d'hébergement des données d'observation et de simulations climatiques

- Renouvellement et extension de l'espace de stockage Ganymède à l'IDRIS
- Extension des plateformes de virtualisation et de containerisation pour l'hébergement de services
- Extension des ressources de calcul du mésocentre ESPRI et capacité à déborder vers d'autres centres
- Accès direct aux données de simulation du CNRM-CERFACS dans la plateforme d'analyse ESPRI
- Automatisation des chaînes de traitement à la demande des données (ex. : corrections de biais)



Meilleure intégration de l'accès aux données de simulations climatiques, d'observation et de la bio-diversité

- Catalogues de données avec moteurs de recherche et thesaurus communs pour les observations et simulations des différents compartiments du système Terre et de la biodiversité
- Protocoles d'accès interopérables entre les différents centres de données
- Construction d'ensembles de données avec des formats adaptés à la fouille rapide de données, au machine-learning et à la visualisation à la demande



Extension des Environnements numériques pour la recherche et les services climatiques

- Accès transparent aux ressources d'autres centres pour le traitement et l'analyse au plus proche des données des pôles de Data Terra ou de la Biodiversité
- Enrichissement du système d'environnement virtuel de recherche type Jupyterhub avec l'intégration d'outils de découverte et d'accès aux données, de développement et d'analyse