

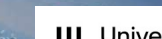


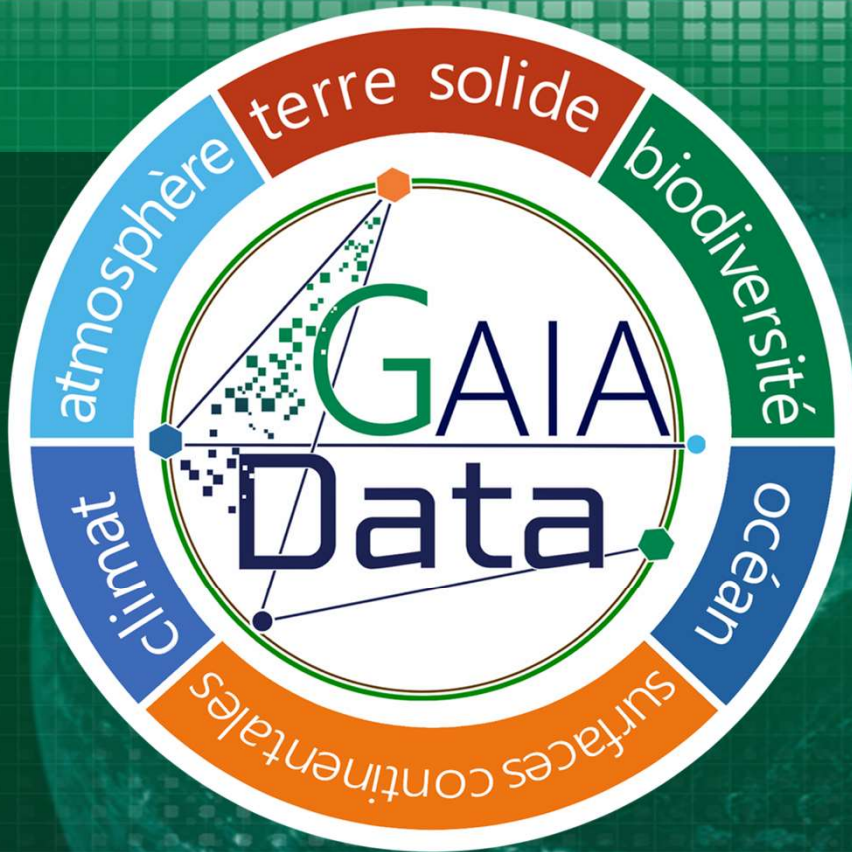
DATA
TERRA

Présentation simple de GAIA DATA

Version du 20 juin 2024

Journée scientifique du programme DATA de l'I-Site de Clermont





Porté par 2 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**

DATA TERRA, une e-Infrastructure de Recherche dédiée au système Terre / Environnement

Développer un **dispositif global d'accès et de traitement de données, produits et services** pour adresser des enjeux scientifiques et des défis sociétaux interdisciplinaires

- Producteurs de données d'observation multi-sources
- 30 Centres de Données et de Services
- 32 Centres d'Expertise scientifiques
- 250 FTE / 500 scientifiques, ingénieurs et techniciens

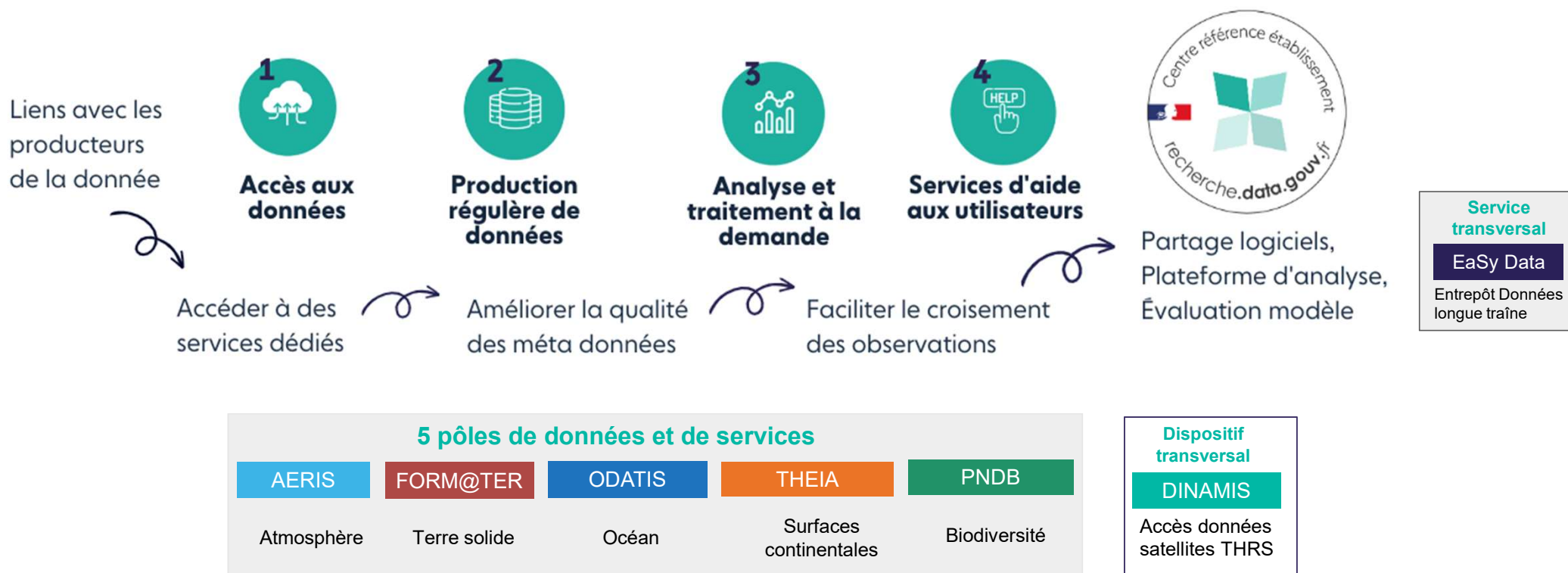
- Coordination/articulation avec les OSU, SNO
- En France
- Outre-Mer
- Etranger

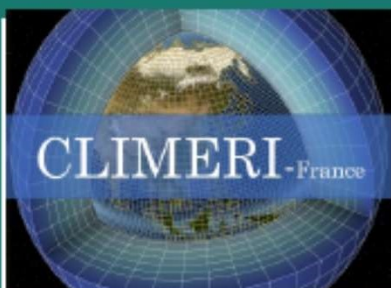


DATA TERRA, une e-Infrastructure de Recherche dédiée au système Terre



L'IR DATA TERRA propose des **services** autour des données d'observation du **système Terre interopérables et interdisciplinaires** à tous les niveaux





CLIMERI-France
Infrastructure de recherche nationale
de modélisation du climat



Gaia Data
12/04/2022

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

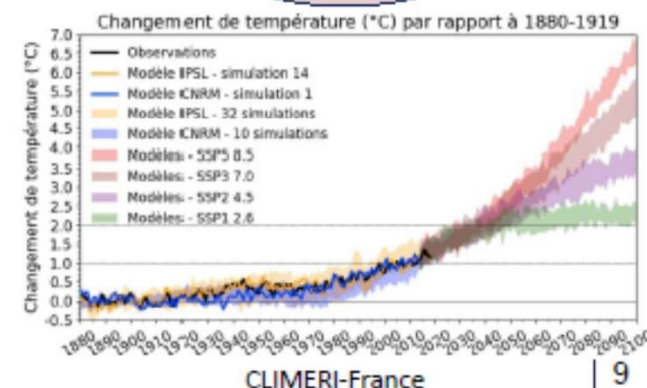
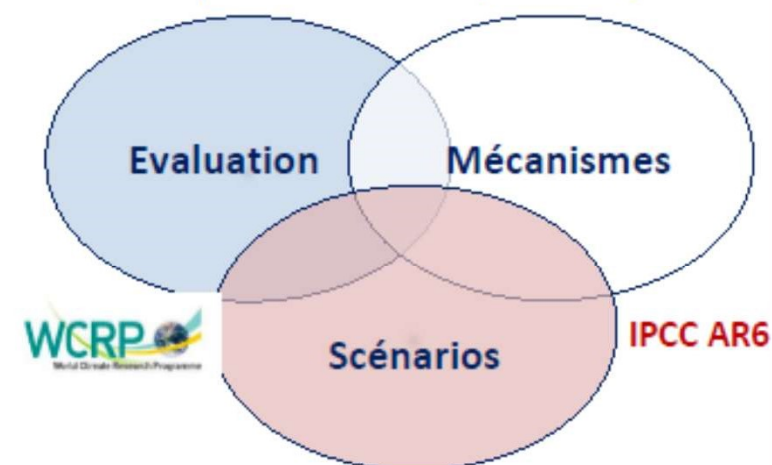
Feuille de route nationale depuis 2016

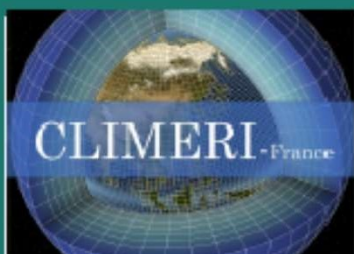


En collaboration avec
SU, IRD, Cerfacs

<https://climeri-france.fr/>

CMIP6 Coupled Model Intercomparison Project Phase 6





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)



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

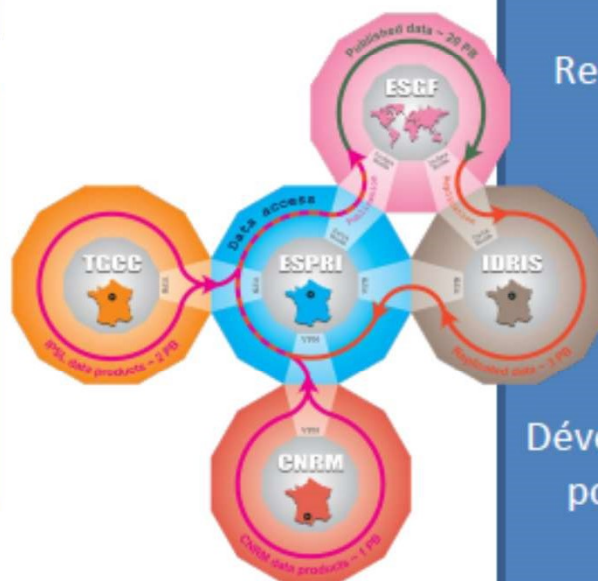
Animation & stratégie scientifique

Modèles de
référence

Calcul &
simulations de
référence

Stockage &
analyse
multi-modèles

Diffusion des données &
interface utilisateurs



Spécificités de Data Terra et de CLIMERI et objectifs de GAIA DATA

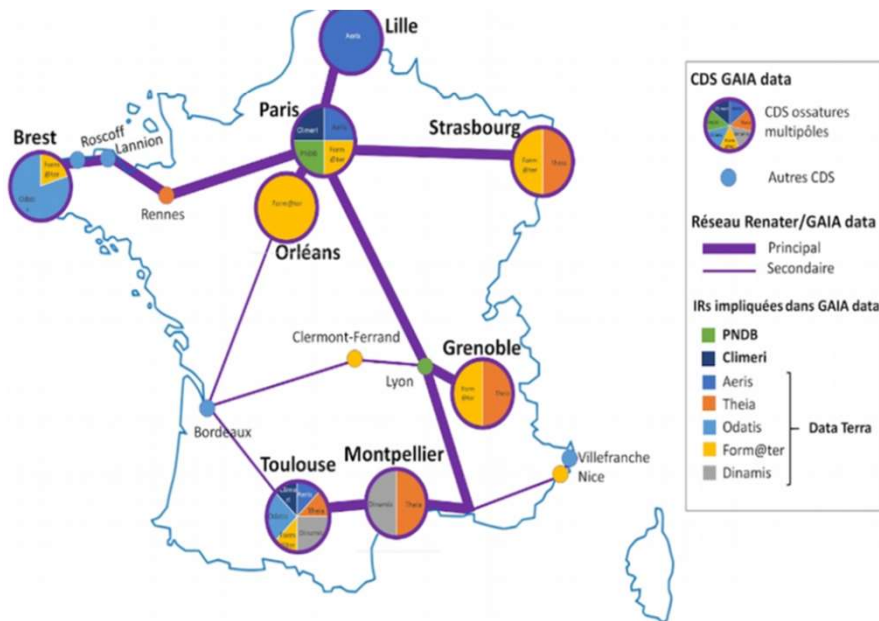


- ⇒ Besoin d'un dispositif permettant aux utilisateurs de s'y retrouver dans cette complexité.
- ⇒ Sans rajouter une couche supplémentaire qui referait ce qui est bien fait dans les pôles, dispositifs et IR
- ⇒ Mais permettant une hybridation multimodale entre toutes ces données
- ⇒ Et un continuum d'infrastructure interne et externe pour pouvoir traiter tout ça.

- Data Terra s'appuie sur 5 pôles de données et de services
 - AERIS, FORMATERRE, ODATIS, THEIA, PNDB + dispositif DINAMIS
 - Avec des thématiques très variées
 - Avec des données très diverses : satellite, in situ, avion, navires, drones, ...
 - Avec des volumétrie et des dénombrements très différents : de plusieurs Po à quelques Ko en fonction de la collection de données
 - Avec des pratiques, des vocabulaires spécifiques
 - Réparties sur environ 30 sites en France
 - Englobant des milliers de collections de données
- CLIMERI-France propose un accès aux données de simulation climatiques
 - Plus homogènes que les données d'observation
 - Très volumineuses : dizaines de Po
 - Très complémentaires des données d'observation dans des logiques d'hybridation multi-modales

INFRASTRUCTURE GAIA DATA

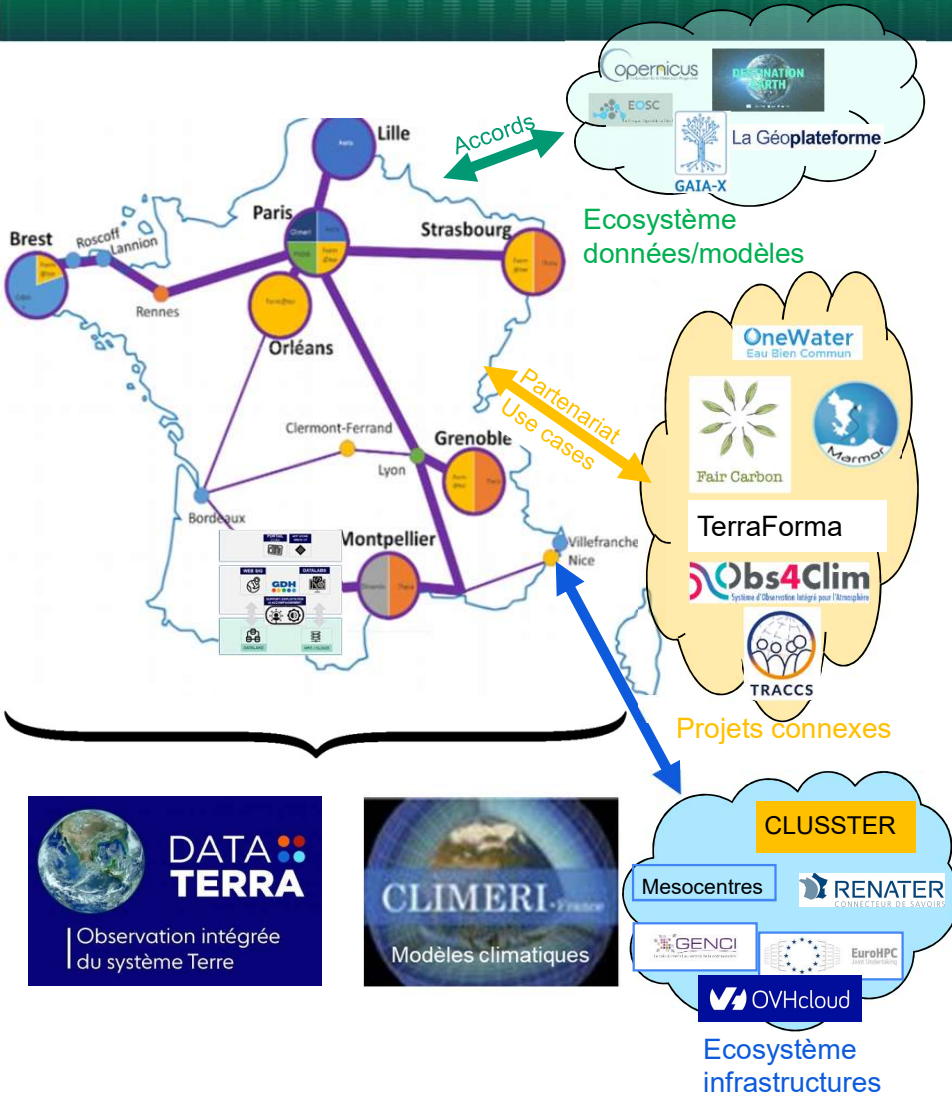
Principes de base



- La colonne vertébrale du système Distribué GAIA DATA s'appuie sur 8 centres sélectionnés d'après une liste de critères
- **Objectif : aider les utilisateurs (chercheurs, institutionnels et privés) à utiliser ces données variées** (thèmes, type de capteur) et volumineuses (des dizaines de Po)
 - Une authentification unique sur l'ensemble des dispositifs
 - Aider les utilisateurs à obtenir l'information sur les jeux de données et les ressources associées et à développer des algo / chaines de traitement
 - Faciliter l'interopérabilité et le mouvement des données et des traitements entre les sites
- **Méthodologie : simple, agile et en phase avec les besoins des utilisateurs**
 - Allergie des communautés aux métacatalogues, oneStopShop, guichets uniques qui fleurissent partout => incompréhensible pour les utilisateurs
 - Dispositif simple et agile :
 - Faibles coûts de maintien en condition opérationnelle
 - Facile à faire évoluer dans un contexte numériques très dynamique
 - En phase avec les pratiques des utilisateurs (Design Thinking, devOPS)

INFRASTRUCTURE GAIA DATA

Intégration dans l'écosystème



En raison de son ampleur, de la diversité de ses domaines d'intervention et de son expertise, l'IR Data Terra dispose d'un atout majeur pour assumer un rôle de premier plan sur la scène nationale, européenne et internationale.

Lien vers l'écosystème des données et modèles. Exemple

- Fédération WekEO, Lien vers la core Platform et le Datalake destination Earth
- Portage et participation à nombreux projets européens (EOSC, ...)

Continuum
Recherche-
Innovation

Intégré dans l'écosystème de la recherche française

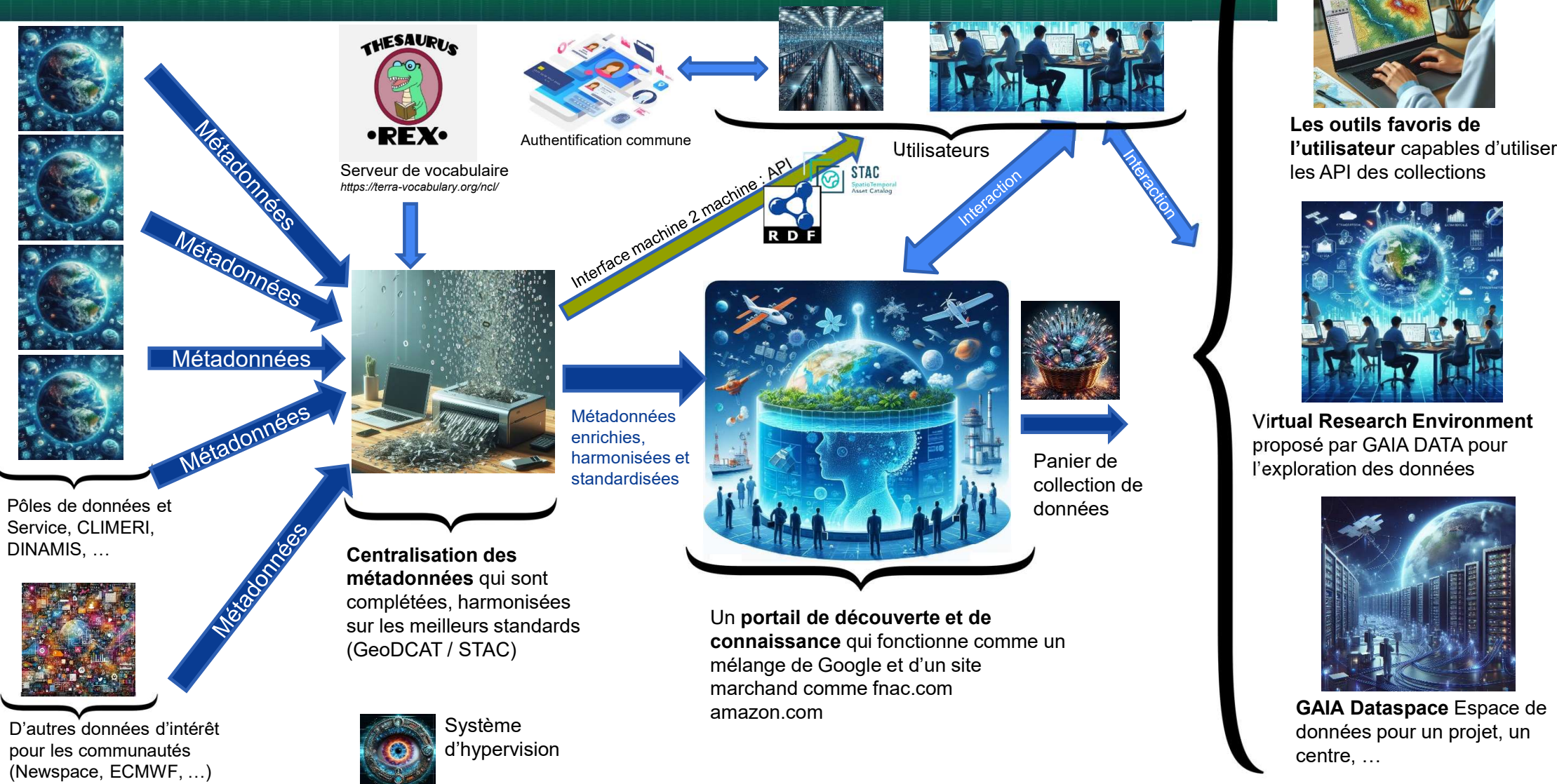
- PEPR -
- Autres projets du type de GAIA DATA
- ...

Intégré dans l'écosystème des infrastructures ⇔ **continuum d'infrastructures**

- Partenariat avec GENCI, Renater, CLUSTER
- CLUSTER : cloud scientifique souverain français
 - Lien OVH, GENCI, EuroHPC, GAIA-X

INFRASTRUCTURE GAIA DATA

Architecture de haut niveau



INFRASTRUCTURE GAIA DATA

Portail de découverte



Image générée par Dall-E 22/04/2024

- Data Terra et CLIMERI hébergent des **milliers de collections de données, très variées**
 - modèles climatiques, Océan, Biodiversité, surfaces continentales, atmosphère,
 - Données images satellite très haute résolution
- Enjeux pour les utilisateurs de **ne pas se perdre dans cet ensemble de données**
 - Ils ne sont pas experts de tous les thèmes
 - Besoin de comprendre ces données, leur potentiel et les ressources existantes associées (documentation, code, algorithmes, logiciels)
- => **portail de découverte des données et de la connaissance**, inspiré de :
 - Google : recherche simple
 - FNAC.com, Amazon.com : recherche avec critères et informations associées
 - ChatGPT, Meta.AI, Bard : IA conversationnelle

INFRASTRUCTURE GAIA DATA

Virtual Research Environment



Image générée par Dall-E 22/04/2024

- Pour exploiter les paniers de données, constitués grâce au portail de découverte, l'utilisateur peut utiliser ses outils favoris ou les outils (VRE) proposés par GAIA DATA
 - Un panier est une liste de collection de données, partagées et constituées avec une communauté
 - Pour chaque collection, on connaît la localisation et l'adresse (API) des données
- Architecture souple et légère
 - Facile à maintenir
 - S'adapte aux changements rapides des technologies d'exploitation de données
- Une solution de base proposée : VRE
 - Basée sur les solutions les plus utilisées du moment : les notebooks Jupyter et Galaxy-E
 - L'utilisateur pourra lancer interactivement des traitements, visualiser les données et résultats et ajuster ses algorithmes en fonction des résultats visualisés interactivement

INFRASTRUCTURE GAIA DATA

Alignement et recueil des métadonnées => MTEP



Image générée par Dall-E 22/04/2024

- Pour nourrir le portail de découverte, besoin de métadonnées sur les collections de données
 - A jour
 - Harmonisées
 - Permettant des recherches sur les collections de données et une utilisation des données associées
- Ces métadonnées implémentent les meilleurs standards et permettent de représenter toute la diversité des données.
 - Thématique
 - Type de données : satellitaires, in situ, avion, navire, drone, bouées, ...

INFRASTRUCTURE GAIA DATA

Catalogue léger intégré dans VRE

EODAG – JupyterLab extension

The screenshot illustrates the EODAG JupyterLab extension interface. On the left, a sidebar titled 'PRODUCTS SEARCH' allows users to filter data by product type (e.g., S2_MSI_L1C), date range (01/08/2022 to 10/08/2022), and maximum cloud cover (75%). The main panel shows a map of Europe with a search box and a table of search results. A red arrow points from the 'Generate code' button in the sidebar to the JupyterLab code editor, which contains Python code for searching and downloading data using the EODAG API.

```
from eodag import EODAGAccessGateway, setup_logging
setup_logging(1) # 0: nothing, 1: only progress bars, 2: INFO, 3: DEBUG
dag = EODAGAccessGateway()
geometry = "POLYGON ((13.854888 47.832835, -2.548829 43.573240, 9.439463 48.642958, 5.722895 44.883376, 1.954888 47.832835))"
search_results, total_count = dag.search(
    productType="S2_MSI_L1C",
    geometry=geometry,
    start="2022-08-01",
    end="2022-08-30",
    cloudCover=75,
    "titleIdentifier": "S3TC2",
)
```

Atelier technique GAIA DATA

EODAG (Earth Observation Data Access Grid) est un catalogue léger qui facilite l'accès et l'analyse de données d'observation de la Terre. Il permet aux utilisateurs de se connecter à une multitude d'API interopérables et de sélectionner des données pertinentes pour leurs besoins.

EODAG : Un catalogue léger pour l'exploration et l'analyse de données

Exploration de données : EODAG offre une interface intuitive pour explorer les différentes sources de données disponibles, en fonction de critères tels que la date, la localisation, le type de capteur et les paramètres mesurés.

Sélection de données : Une fois les données souhaitées identifiées, EODAG permet de les sélectionner et de les télécharger facilement.

Intégration avec Jupyter Notebook : EODAG peut être intégré dans un notebook Jupyter, ce qui facilite l'analyse et la visualisation des données téléchargées.

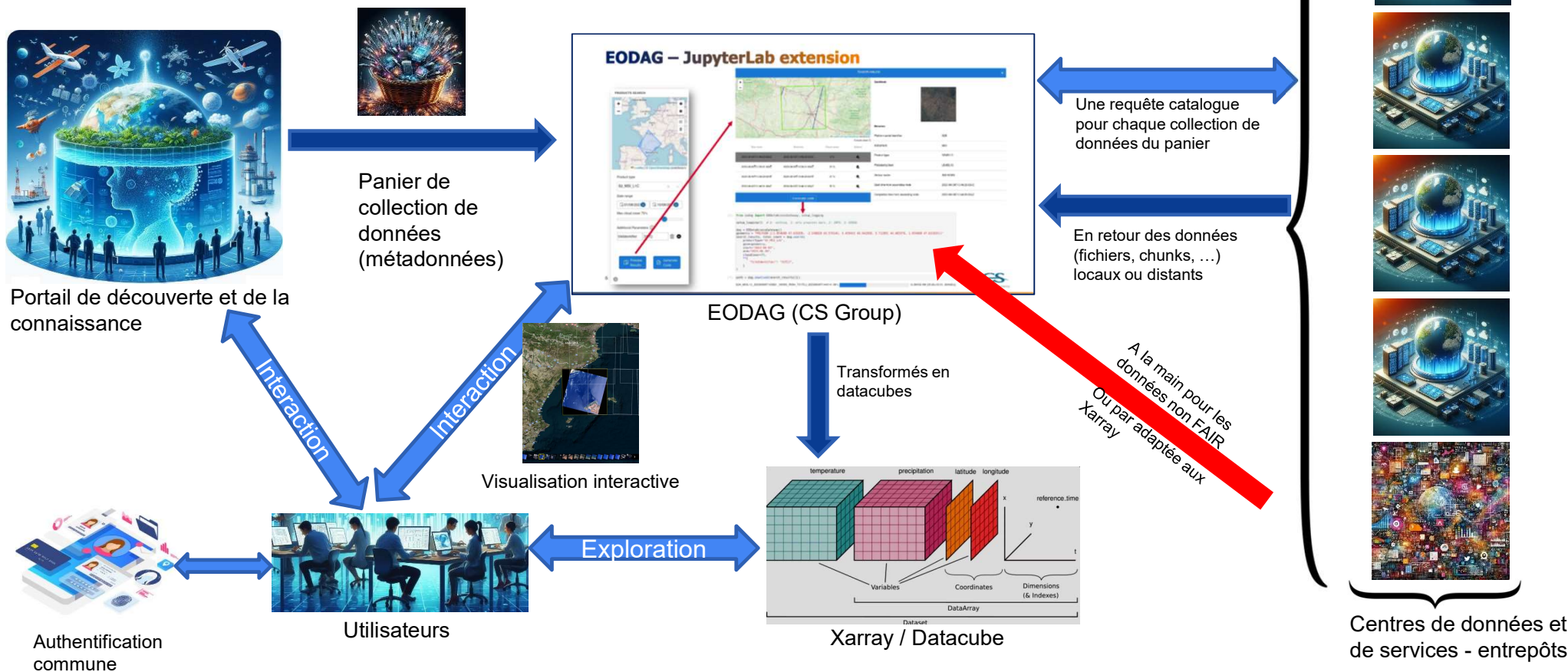
Création de DataCubes Xarray : EODAG permet de remplir des tableaux Xarray avec les données sélectionnées, créant ainsi des DataCubes prêts à être analysés à l'aide d'outils d'analyse scientifique avancés.

INFRASTRUCTURE GAIA DATA

Catalogue léger intégré dans VRE



Système d'hypervision

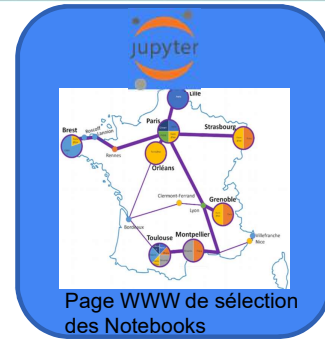


INFRASTRUCTURE GAIA DATA

Gestion des JupyterLab/Hub dans les centres GAIA DATA (et autres)



Système d'hypervision



Page WWW de sélection des Notebooks



Portail de découverte et de la connaissance



Panier de collection de données (métadonnées)

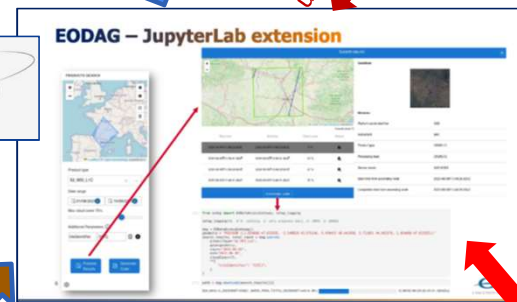
Page permettant de sélectionner le notebook (variante, site, type de machine)



JupyterHub du site choisi



Fichier de conf local EODAG
- Quelles données en local)
- Hiérarchie des sources de données si plusieurs existent pour chaque JDD



JupyterLab sur site choisi



Utilisateurs

Données des utilisateurs BYOD

Peut exécuter des traitements à distance (OGC Process API si nécessaire)

Une requête catalogue pour chaque collection de données du panier

En retour des données (fichiers, chunks, ...) locaux ou distants

A la main pour les données non FAIR
Ou par adaptée aux Xarray

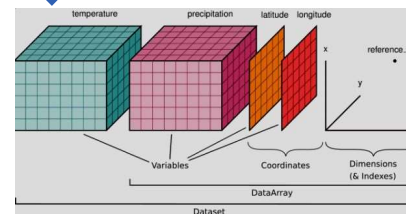
L'utilisateur colle du code relatif au panier dans le JupyterLab
Interaction
Et sélectionne les données jeux par jeux

Exploration



Utilisateurs

Transformés en datacubes



Xarray / Datacube



Authentification commune



Centres de données et de services - entrepôts


INFRASTRUCTURE GAIA DATA

EODAG Le bout de code à insérer



File Edit View Run Kernel Git Diagram Tabs Settings Help

PRODUCTS SEARCH



Product type (*)

S2_MSI_L1C

Date range

Start: 28/04/2024

End: 02/05/2024

Max cloud cover 100%

Additional Parameters

Add search parameter

Search

```
[1]: from eodag import EODataAccessGateway, setup_logging

      setup_logging(1) # 0: nothing, 1: only progress bars, 2: INFO, 3: DEBUG

      dag = EODataAccessGateway()
      search_result, total_count = dag.search(
          productType="S2_MSI_L1C",
          start="2024-04-28",
          end="2024-05-02",
          cloudCover=100,
      )
```

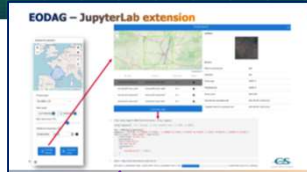
Identifiant du JDD : nickname

INFRASTRUCTURE GAIA DATA

Workflow sur un/plusieurs site(s)



Système d'hypervision



Définition de certains éléments du WF avec JupyterLab et des traitements distribués (OGC API process)



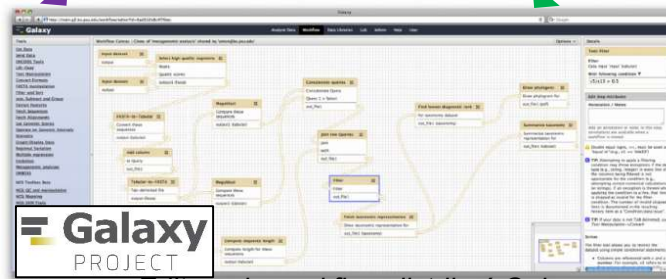
Définition de certains éléments du WF avec OpenEO



Portail de découverte et de la connaissance



Panier de collection de données (métadonnées)



Editeur de workflow distribué Galaxy

Exécution du workflow selon plusieurs protocoles - OpenEO

Exécution du workflow selon plusieurs protocoles

- En local
- JupyterHub
- OGC API Process
- ...



Centres de données et de services - entrepôts



Authentification commune



Utilisateurs

Interaction

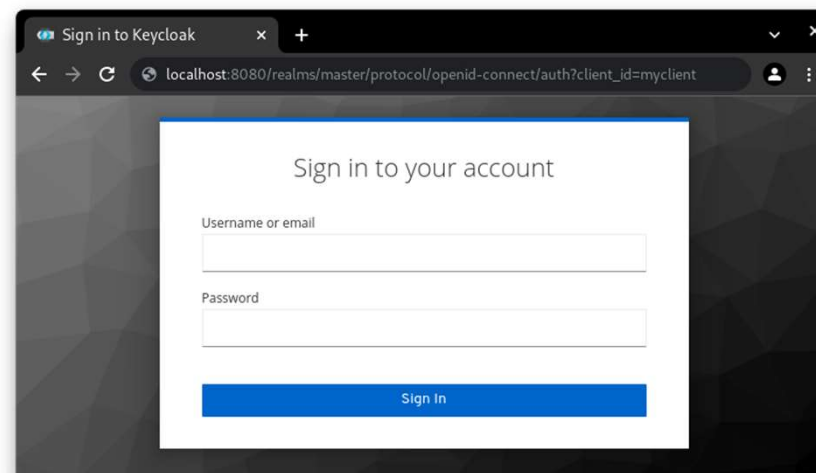
L'utilisateur définit son workflow distribué dans Galaxy

INFRASTRUCTURE GAIA DATA

Le service SSO de Gaia Data fournit un système d'authentification centralisé pour l'ensemble des applications de la plateforme. Il assure les fonctions de :

- Connexion vers des fournisseurs d'identité externe (EduGain, ORCID)
- Gestion du cycle de vie des comptes utilisateurs (création, authentification, modération, autorisations, expiration, désactivation, suppression)
- Gestion des groupes organisationnels d'utilisateurs (ex: groupe au niveau des organismes, laboratoires, collectivités territoriales, ...)
- Attribution de rôles applicatifs pour les utilisateurs (gestion des droits d'accès aux ressources au sein des applications)
- Single Sign On : L'utilisateur se logue une fois et accède à l'ensemble des applications Gaia Data en fonction de ses droits

Authentification utilisateur IAM / SSO Gaia Data



INFRASTRUCTURE GAIA DATA

IAM / SSO Gaia Data Implémentation



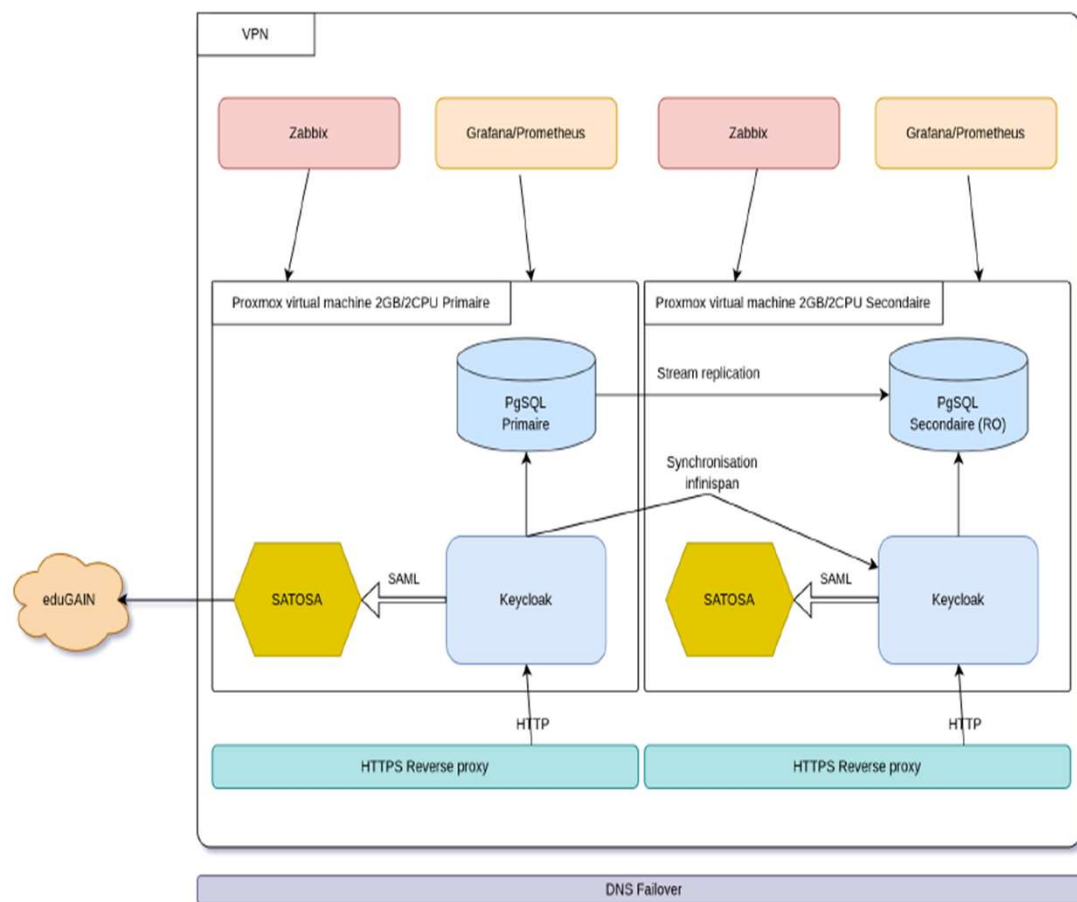
Le système IAM Gaia Data se décompose en deux parties :

- Keycloak SSO :

- Connexion vers les IdP
- Base de données users et groups
- Centralisation des attributs des utilisateurs
- Base de données des applications et rôles (pour les applications qui le souhaitent)
- Services d'authentification (OIDC, SAML, Oauth2)
- Single Sign On

- Application (UI & API) de gestion des droits

- Attribution des rôles des utilisateurs par les respons. applicatifs
- Modération des demandes d'accès
- Création de compte et groupes utilisateurs



Architecture de déploiement du Keycloak Gaia Data

INFRASTRUCTURE GAIA DATA

IAM / SSO Gaia Data Modération des demandes d'accès



Accès à un jeu à accès restreint

Une fois authentifié, l'utilisateur peut demander l'accès au PI

Si le modérateur valide : ajout des droits au compte de l'utilisateur (keycloak)
Il a maintenant accès aux données

Appel à l'API pour enregistrer la demande

Le modérateur a accès à la demande dans l'application de gestion

Cas d'usages	Je suis...	Avec iRODS, je veux...
	Propriétaire/ Gestionnaire de données et produits de données	Partager publiquement des ensembles de données / produits de données
		<ul style="list-style-type: none"> → Rendre mes données/produits de données accessibles à toute la communauté GAIA-DATA → Garder la pleine gestion de mes données/produits de données → Pouvoir mettre à jour et publier une nouvelle version d'une donnée / d'un produit de donnée → Pouvoir définir un centre de données préférentiel d'archivage physique (selon contraintes de calcul, de spécificités des données - volume, type -, etc) → Téléverser et partager automatiquement un produit de données (issu d'un service à la demande GAIA-DATA)
	Un utilisateur des services GAIA-DATA (chercheur, etc)	Partager des données au sein d'un groupe restreint de personnes (groupe de travail, projet, pôles, ...)
		<ul style="list-style-type: none"> → Gérer les accès en lecture/écriture sur mes jeux de données (sous embargo, non qualifié, sensible,...) → Téléverser et partager des produits de données (issus d'un service régulier GAIA-DATA)
	Gestionnaires de la plateforme GAIA- DATA	Télécharger des ensembles de données et produits de données publiés
		<ul style="list-style-type: none"> → Télécharger des données et produits de données FAIRisés
		Traiter des données et produits de données (par des services à la demande et des VRE)
		<ul style="list-style-type: none"> → Téléverser des données privées dans mon espace utilisateur (pour les rendre accessible à un VRE) → Accéder à mes données privées depuis un VRE quel que soit son lieu d'exécution → Disposer d'un espace de stockage disponible au plus près du back-end (HPC) du service à la demande et ou du VRE d'intérêt → Partager des données privées lors d'une collaborations sur un VRE
		Fédérer la grille de données GAIA-DATA avec des infrastructures externes
		<ul style="list-style-type: none"> Intégrer des données d'entrepôts nationaux et internationaux externes à GAIA-DATA → ingestion de grandes quantités de données
		Administrer l'infrastructure logicielle iRODS
		<ul style="list-style-type: none"> → déploiement des composants logiciels → monitoring et comptabilité → maintien en condition opérationnelle

INFRASTRUCTURE GAIA DATA

Grilles de données iRODS pour quoi faire ?

RODS

/GaiaDataZone/catalog/<datahub>/<CDOS>/<theme>/<category>/...

Données du catalogue, stockage à long terme. L'accès en lecture peut être publique ou restreint, l'écriture est réservée pour chaque type de donnée à son producteur identifié.

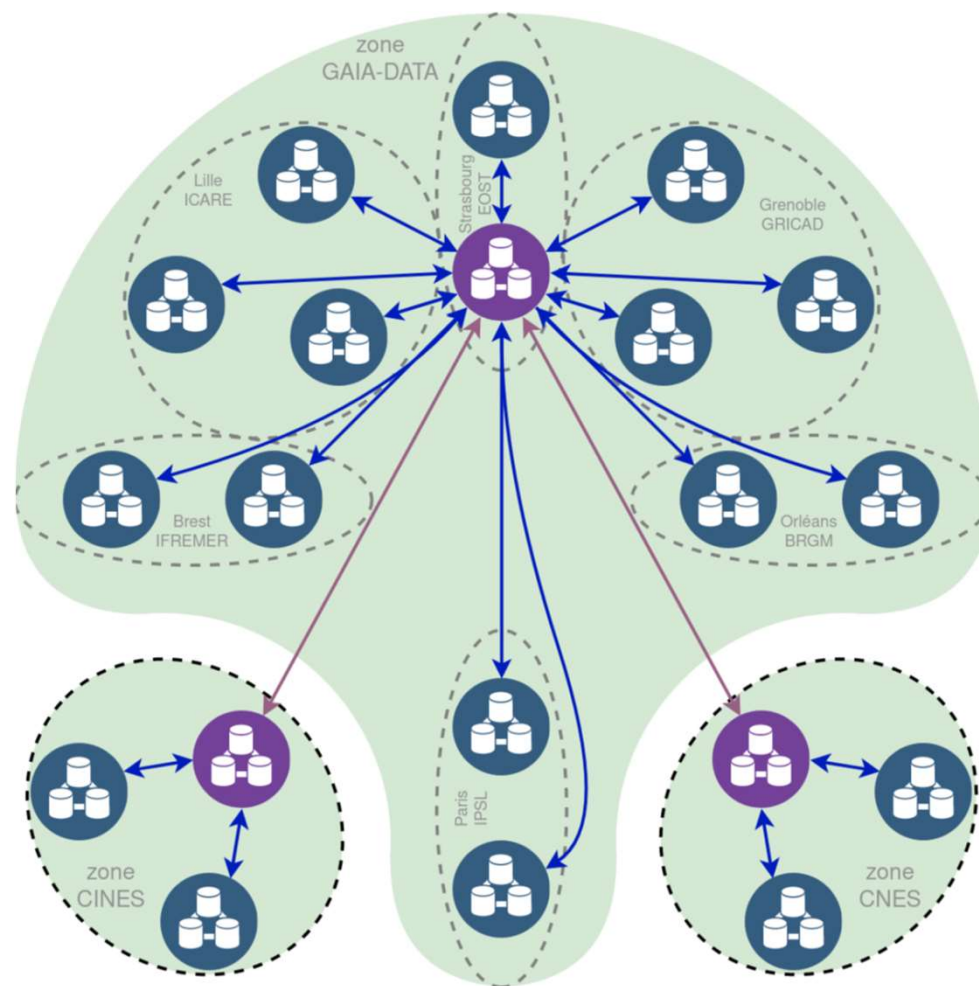
/GaiaDataZone/home/<username>/...

Données utilisateur avec un espace de stockage privé court/moyen terme, selon quota d'espace disque et de durée à définir, propre à chaque utilisateur.

Exemple : analyses via VRE.

/GaiaDataZone/project/<projectname>/...

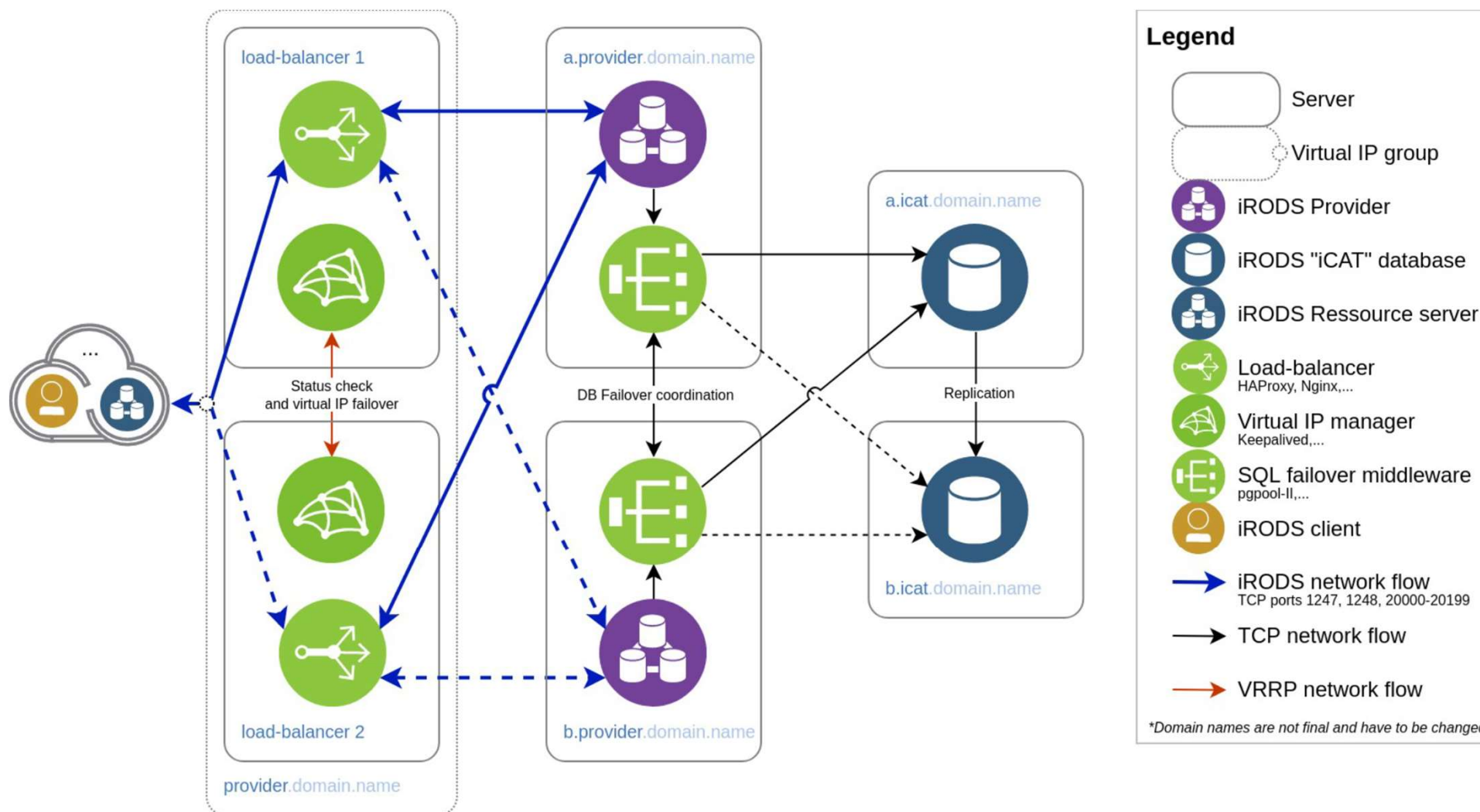
Stockage dédié à un groupe d'utilisateurs en mode projet collaboratif, selon quota d'espace disque et durée à définir, ACLs définis au cas par cas.



INFRASTRUCTURE GAIA DATA

Grilles de données Infrastructure cible

RODS



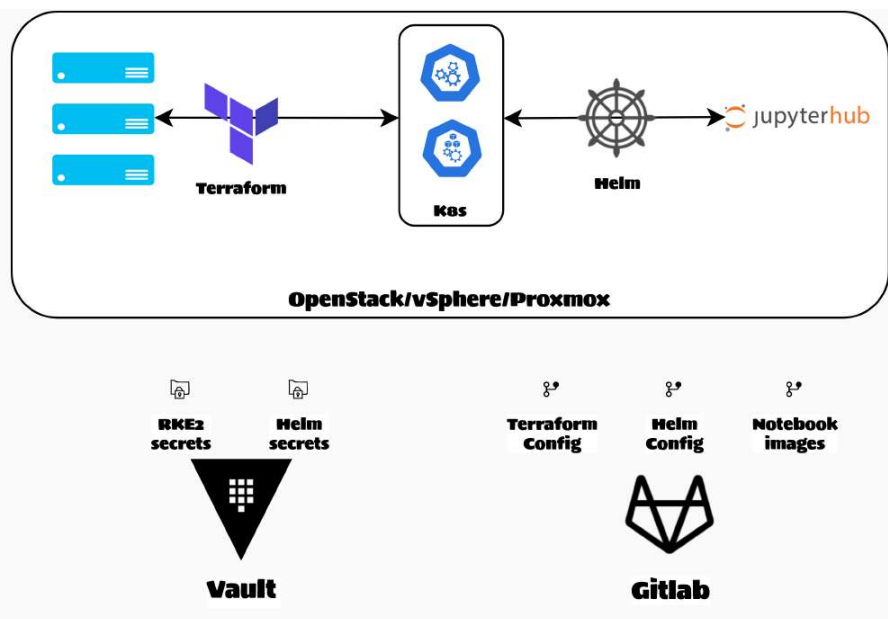
INFRASTRUCTURE GAIA DATA

Déploiement Infrastructures & Applications Usine Logicielle Gaia Data

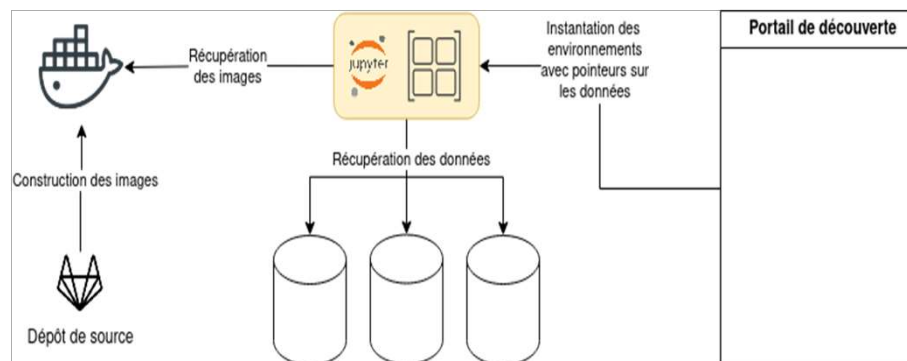
Objectifs :

- Simplifier et homogénéiser les déploiements, la validation des solutions et les tests de sécurité entre partenaires de Gaia Data
- Favoriser les architectures en micro-services (containerisation) pour faciliter le passage à l'échelle
- Utiliser des méthodes « cloud ready » (Infrastructure As Code) pour faciliter le débordement vers d'autres infrastructures numériques (GENCI, EOSC, France-Grille, cloud commerciaux)

Déploiements d'Infrastructures



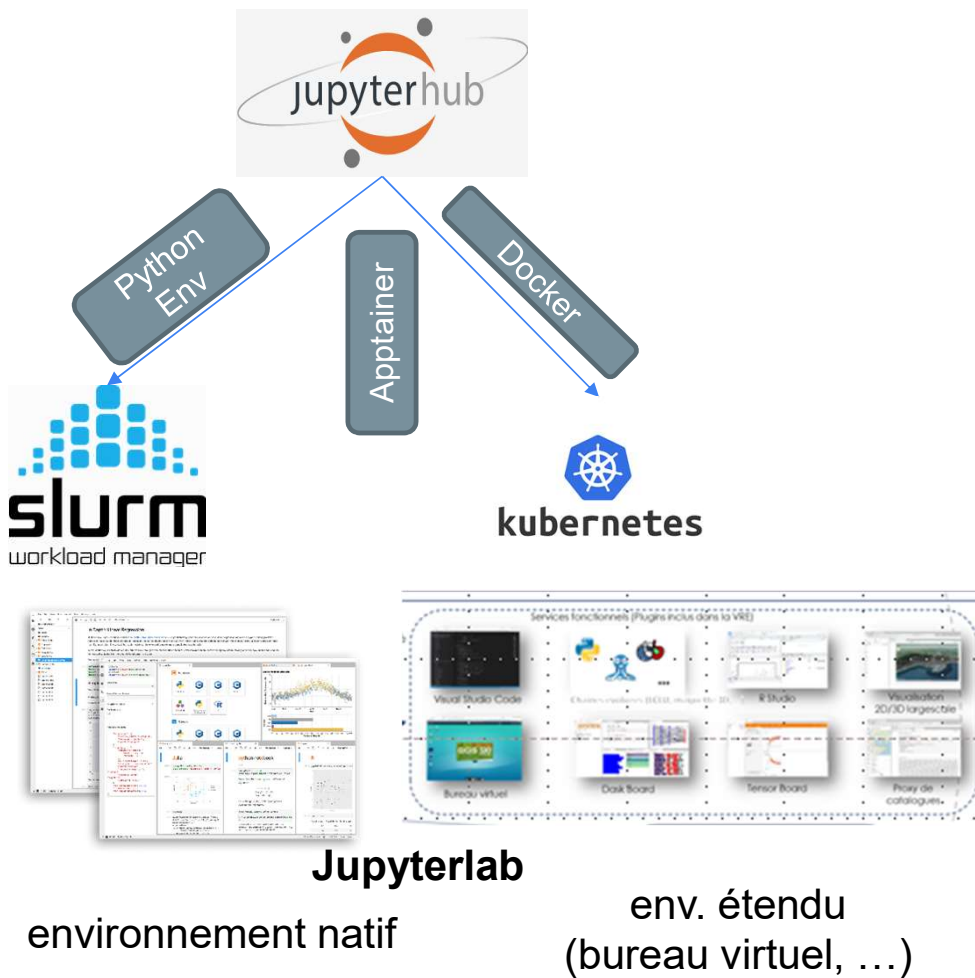
Déploiements des environnements applicatifs



Flux de construction des images VRE

- Spécification des environnements CONDA, PIP → Gitlab
- Construction des images docker → Gitlab CI/CD
- Scan de vulnérabilité des images docker → Trivy
- Stockage des images docker sur un dépôt → Nexus / Harbor
- Si besoin, conversion des images docker en aptainer

INFRASTRUCTURE GAIA DATA



Déploiement Infrastructures & Applications

Instanciation Jupyterlab

Exemple d'instanciation VRE « Observation de la Terre » au CNES

