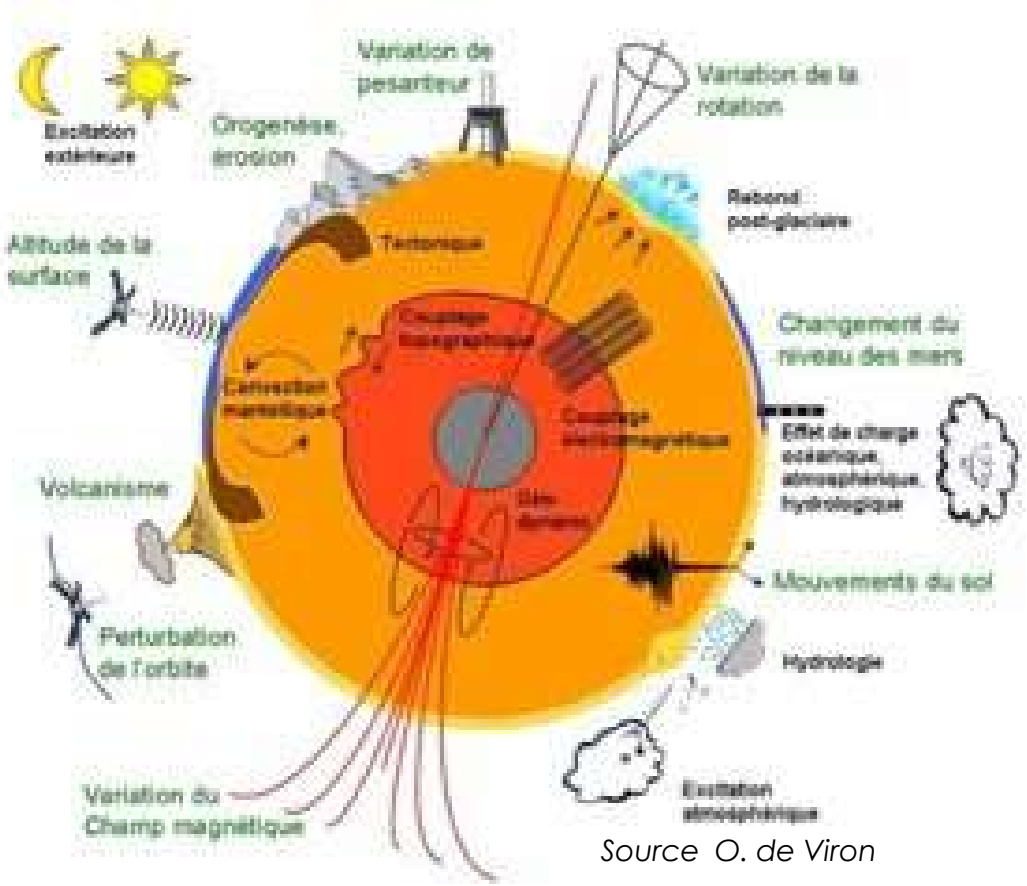
Gaia Data - Infrastructure distribuée de données et services : observation et modélisation intégrée du système Terre

Richard Moreno
Directeur Technique Infrastructure de Recherche Data Terra

Karim Ramage
Directeur Technique Adjoint Infrastructure de Recherche Data Terra

Contexte et Enjeux

- **La Terre**, a fascinating but complex system:
  - numerous geophysical, geodynamic and environmental processes,
  - very variable spatial and temporal scales,
  - many interactions, within and between its various compartments: *inner Earth, land surfaces, ocean, atmosphere and interactions with anthroposphere*



Source O. de Viron

Understanding these geophysical, geodynamic and environmental processes, demands analyzing **numerous and very large datasets (*satellite, in situ, campaigns, long term observations but also experimentation results, model outputs, …*)**

Data Terra en quelques chiffres

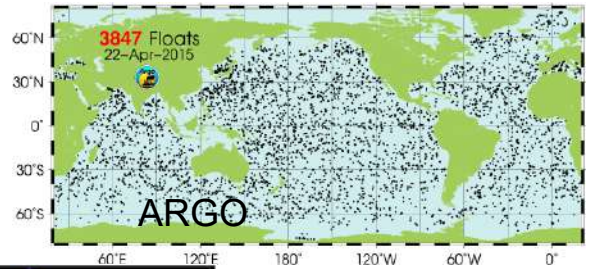
- 26 organismes et universités
- 4 pôles de données : **AERIS, FORM@TER, ODATIS et THEIA**
- 6 services (DINAMIS) et groupes de travail transversaux
- 30 Centres de Données et de Services (CDS) et Infrastructures de données spatiales (IDS)
- 25 Consortium d'Expertise Scientifique
- **170 ETPT / 400** scientifiques, ingénieurs et techniciens
- Plus de 500 produits et services, plus de 15000 utilisateurs
- 50 000 To (2018) ; 100 000 To (2022/2023) ; 150 Peta (2025)

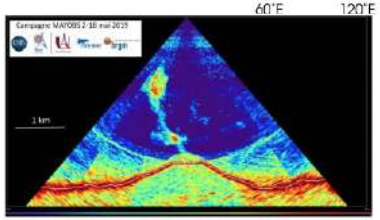
Explosion des flux et de la diversité des données du système Terre

- Systèmes d'acquisition : missions spatiales, capteurs in-situ
- Systèmes de production (Cloud, HPC) : simulations d'ensemble, amélioration des résolutions, ré-analyse de données
- Couvre l'ensemble des compartiments du Système Terre

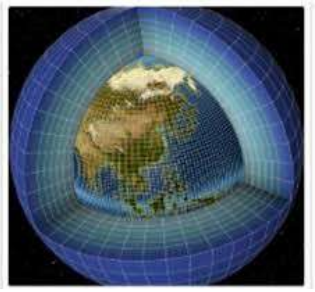
Copernicus/Sentinel

Balloon





Volcano observatories



Modélisation climatique

Le développement de nouvelles recherches repose sur notre capacité à gérer l'ensemble du cycle de la donnée :

- **Acquisition** : traitement, réduction et compression des données en continu, fourniture de données primaires
- **Archivage et gestion des données et métadonnées** : archivage à long terme, conservation (métadonnées, provenance, distribution)
- **Diffusion via FAIR data services** : services d'observatoire virtuel multi-sources,

Ces nouvelles recherches comprennent :

- **Analyse statistiques de données** : données distribuées multi-sources, ML
- **Simulations d'ensemble** : systèmes multi-physique et multi-échelle intégrant des données de forçages multithématiques
- **Inversion/assimilation** : méthodes d'inférence probabiliste à haute dimension reposant sur de très grands ensembles d'apprentissage.

Wide-area workflows (HPC/HDA) : Capacité à déployer des workflows de traitement dans un contexte **d'infrastructure multifournisseurs** en ramenant le traitement au plus proche de la donnée ou en regroupant les multiples sources de données sur les moyens de traitements.

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)

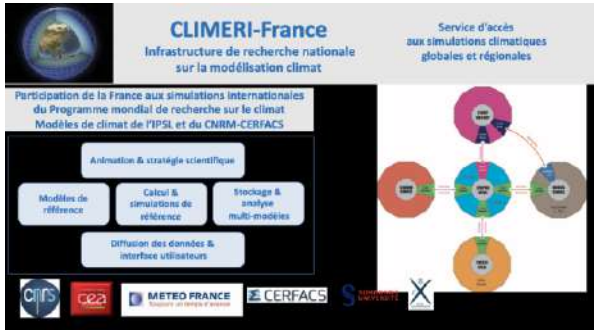
21 Partenaires : CNRS (coord.), CNES, IFREMER, IRD, BRGM, IGN, INRAE, Météo-France, MNHN, CEA, IPGP, CINES, Sorbonne Univ., Univ. Grenoble-Alpes, Univ. Lille, Univ. F. Toulouse, UNISTRA, SHOM, OCA, FRB, CERFACS

Objectif :

- Mettre en œuvre **une plateforme intégrée de données et services distribuées** soutenues par les 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 **pour les communautés scientifiques du domaine système Terre et environnement**, les acteurs **publics et socioéconomiques**

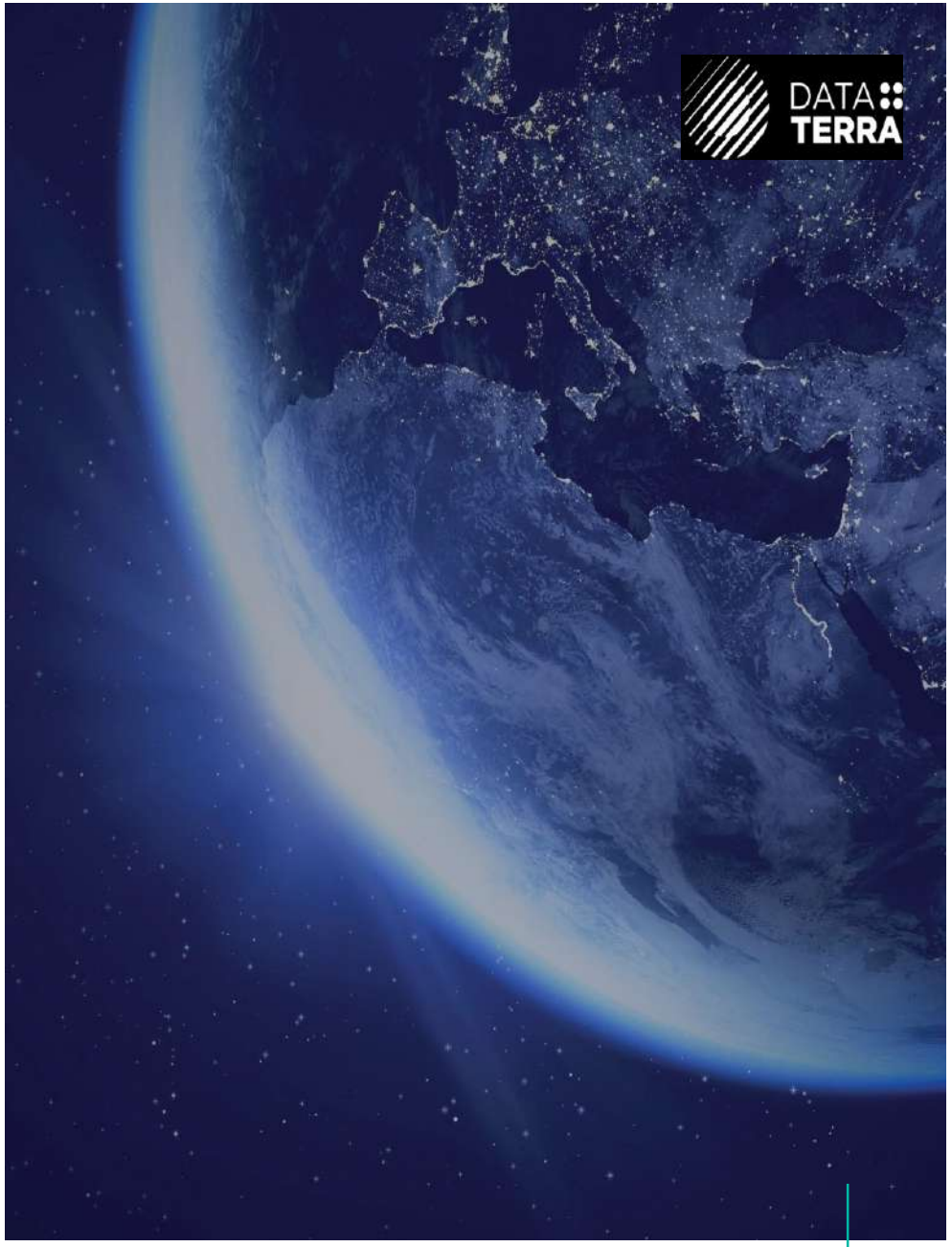
Les 3 Infrastructures de Recherche

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.

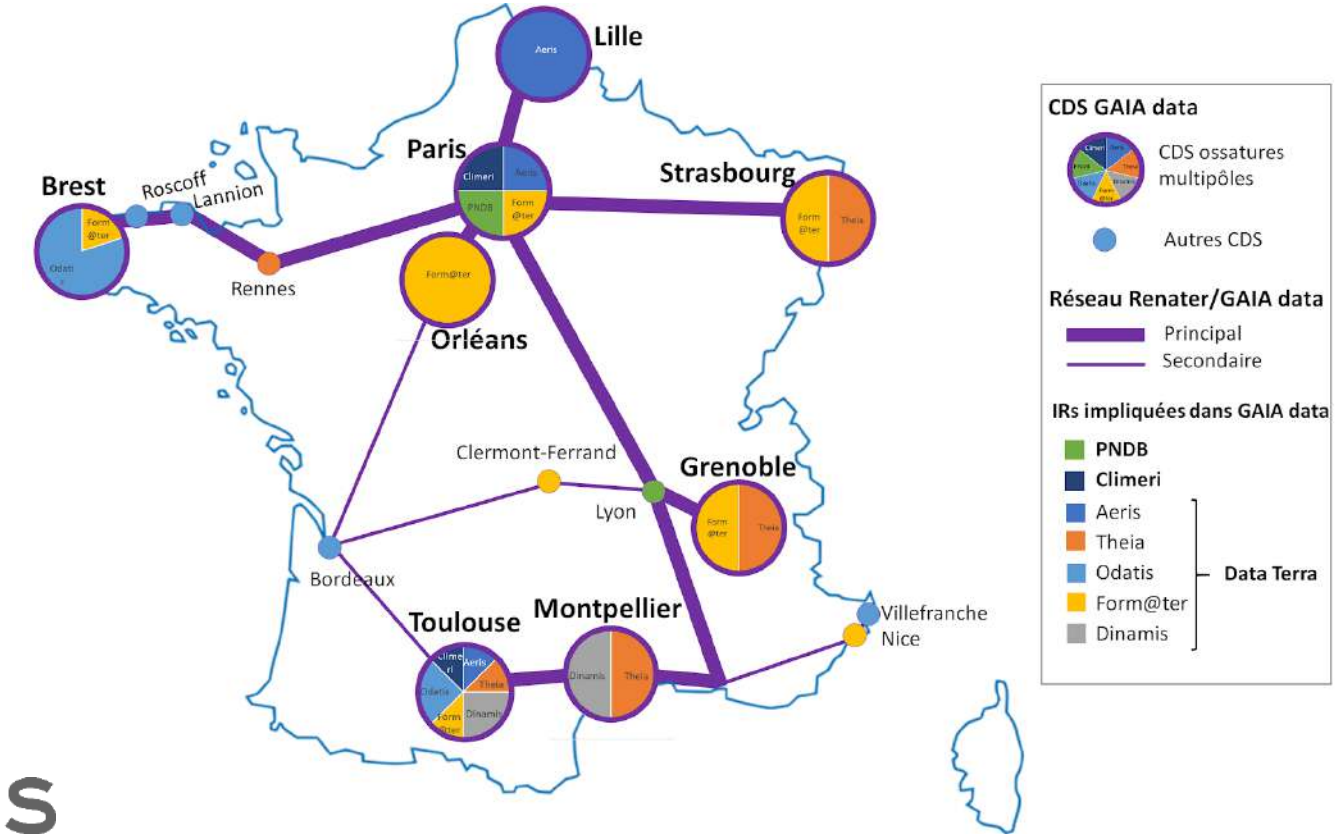
Le PNDB, le centre national de données sur la biodiversité, vise à fédérer les approches de données existantes au sein des infrastructures de recherche sur la "Terre vivante".



Vers un continuum d'infrastructures pour des usages scientifiques des données

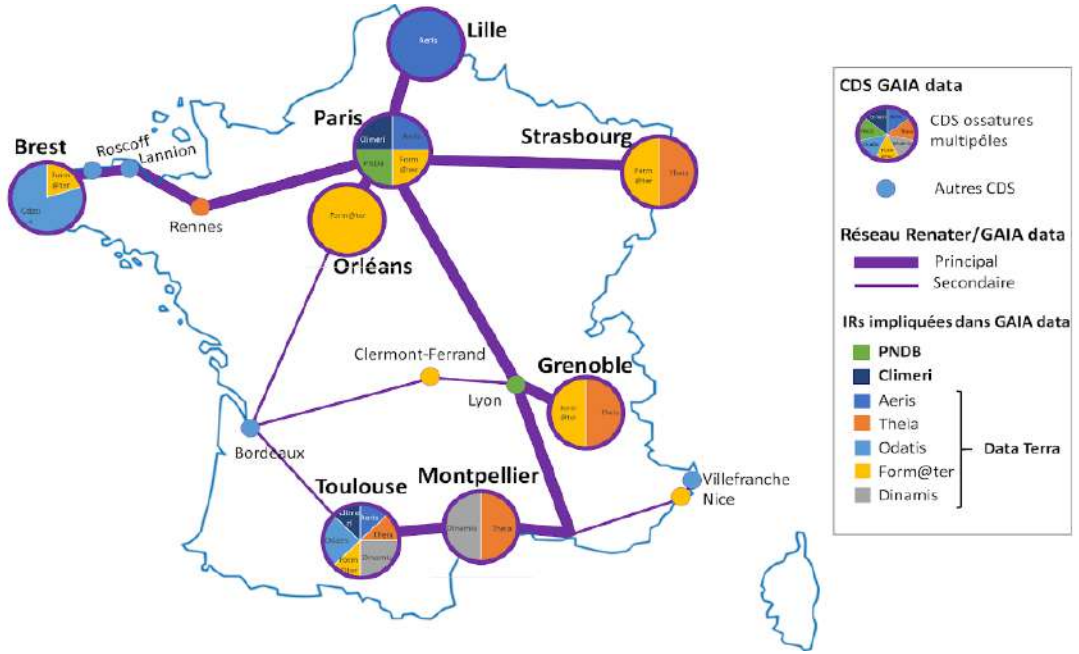
Grille de données et de services

- >30 Centres de données et de services
  - Dont 8 structurants :
    - IFREMER : Brest
    - ICARE : Lille / AERIS
    - Grenoble : OSUG
    - CNES : Toulouse
    - BRGM : Orléans
    - Strasbourg : A2S
    - Paris & Palaiseau AERIS& CLIMERI
    - GEOSUD : Montpellier
  - Data Terra 170 ETPT / 400 scientifiques, ingénieurs et techniciens
  - Plus de 300 produits et services, plus de 15000 utilisateurs
  - Une volonté de rationaliser les infrastructures informatiques : **INFRANUM**
  - Volume de données actuel
    - ~30/40 Po => 100+ Po en 2023
    - Avec IoT & 5G les données in-situ vont grossir



- **Modèle économique de la donnée : complexe**
- **Forte présence en Régions**
- **GRILLE de données et de services** iRODS
- **Cloud : continuum & interopérabilité du processing**
- **Rôle central du réseau et de la sécurité associée** RENATER CONNECTEUR DE SAVOIRS
- **Possibilité d'utiliser des moyens externes (GENCI, CC IN2P3, Clouds commerciaux)**
  - Si besoin de beaucoup de SLA ou de capacité informatique
  - Adapté aux usages « externes » tels que applications commerciales

Infrastructure Gaia Data



30 Centres de données et de services dont 8 HPC-Tier2 / (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

- Croissance des données d'observations in-situ avec 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.

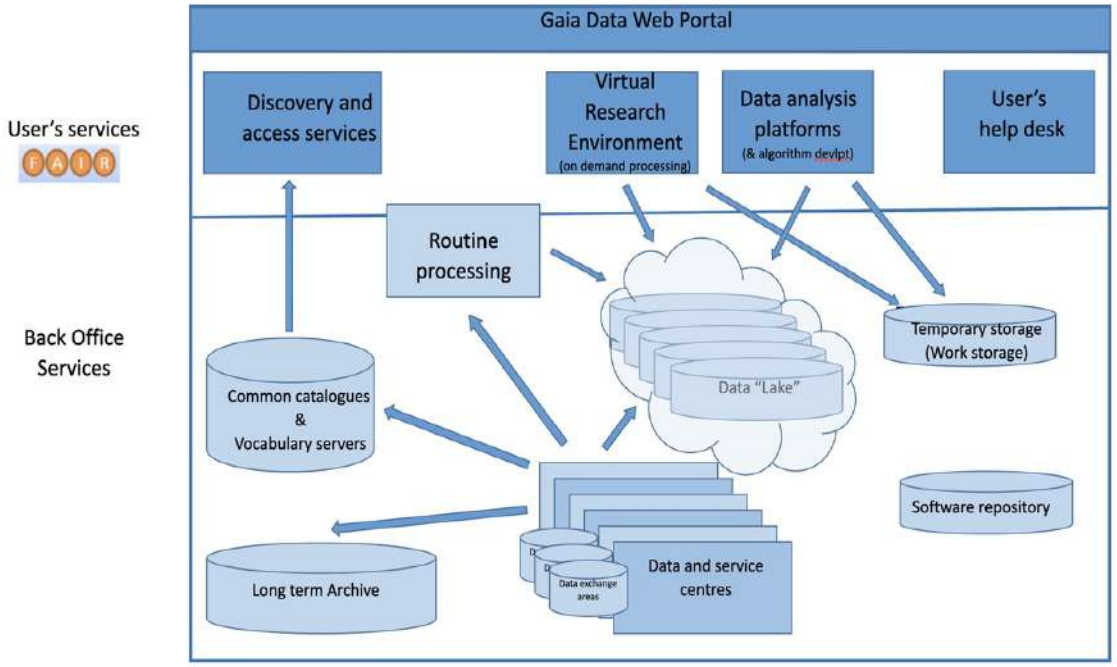
- **Interopérabilité des traitements entre les centres de Gaia Data** (Cloud Management Framework d'EGI) sur les 8 centres pour offrir des interfaces interopérables d'accès aux ressources de traitement

- **Possibilité d'utiliser des moyens externes** (GENCI, EGI, DIAS, clouds publics)

- Si besoin d'un SLA supérieur ou de débordement des capacités de traitement

- Pour les usages « externes » tels que applications commerciales

Les services Gaia Data



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

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

⇒ 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 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

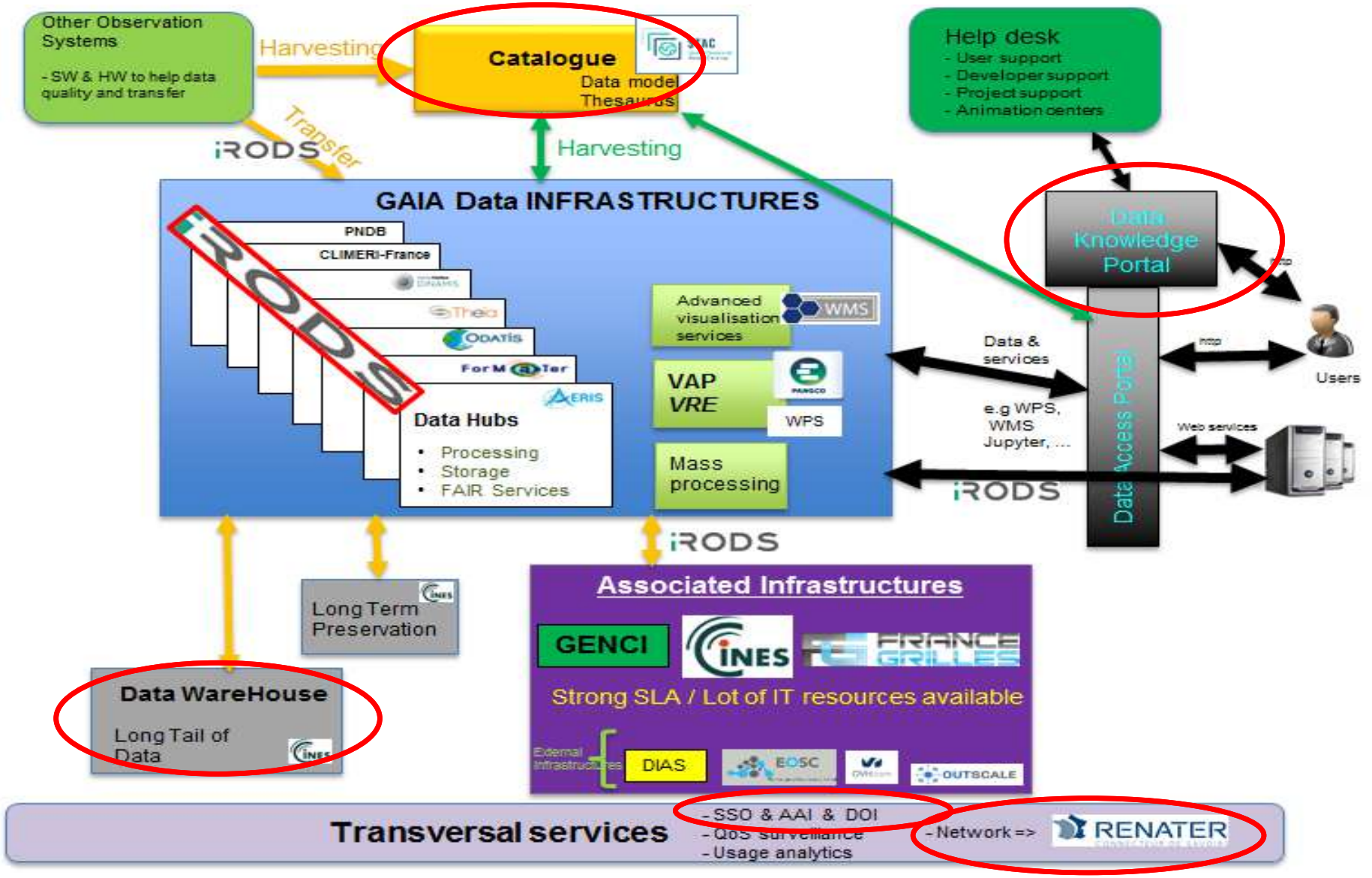
- Grille de données, cloud, portail connaissances, SSO, Métriques, support utilisateurs & formation

Services – Virtual Analysis Platform

=> Data Scientist

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

Schéma général d'architecture



contact@data-terra.org

www.data-terra.org