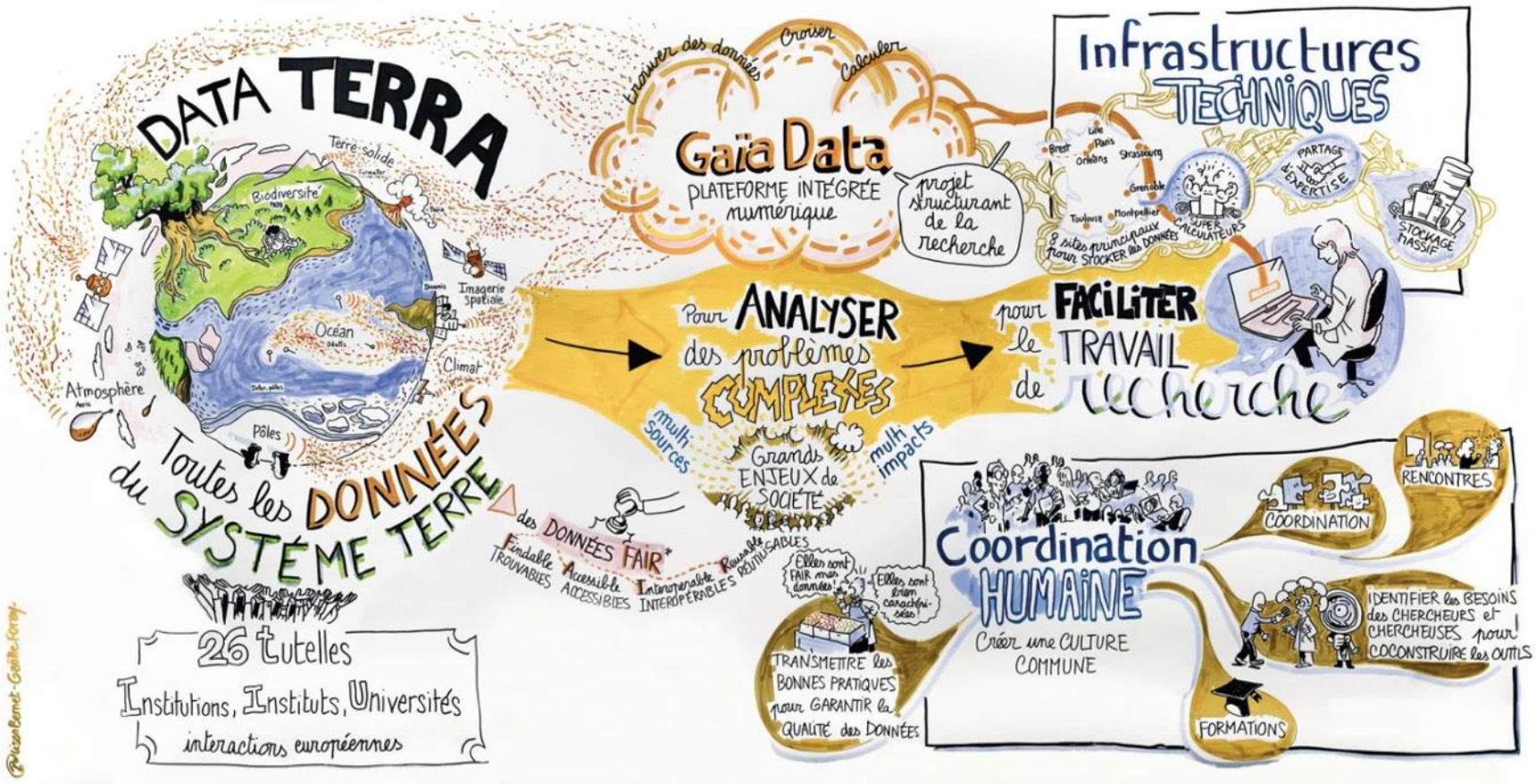
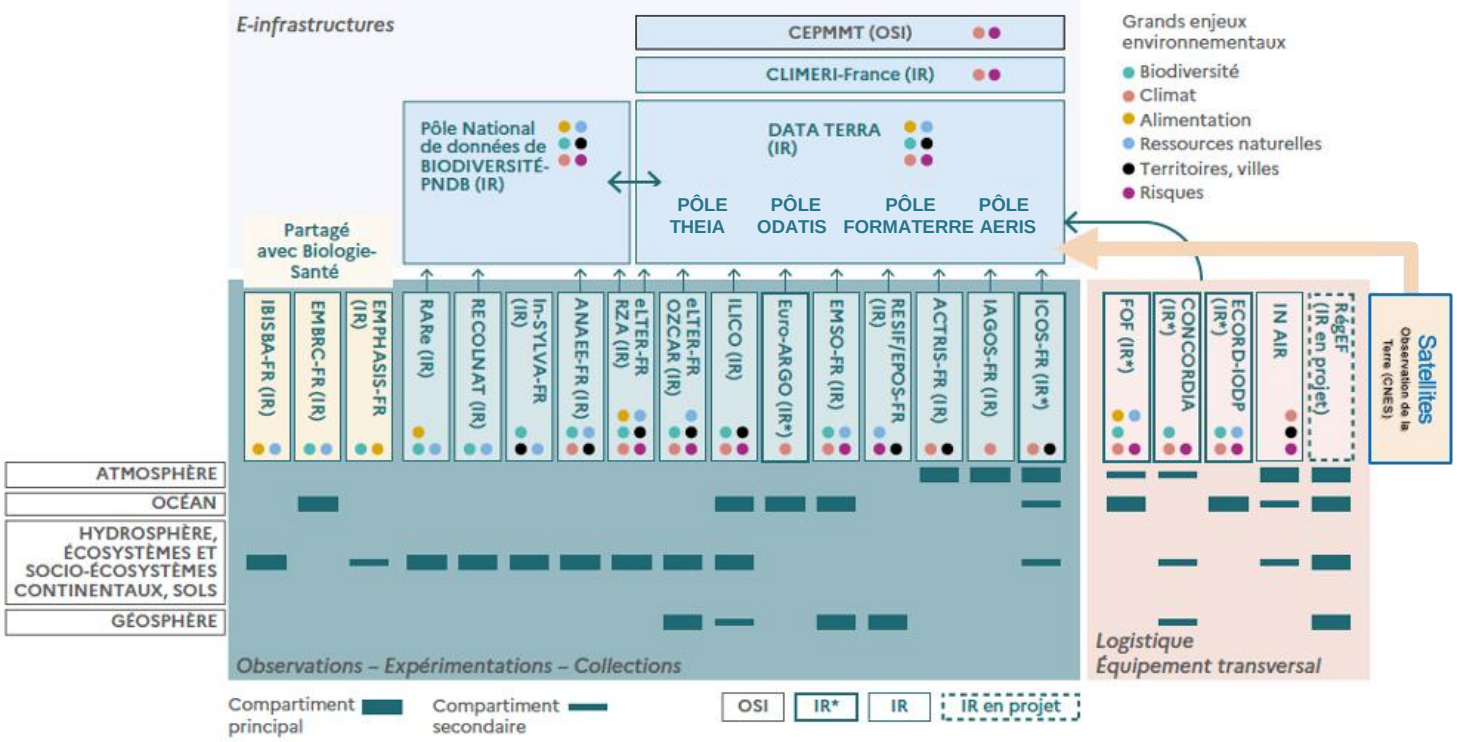
La préservation des données au sein de Data Terra / Gaïa Data

Joël Sudre *(CNRS - joel.sudre@data-terra.org), Coordinateur technique* ***UAR Data Terra et du Projet GAIA DATA***

R.I.P. Data 2024 – 01/10/2024

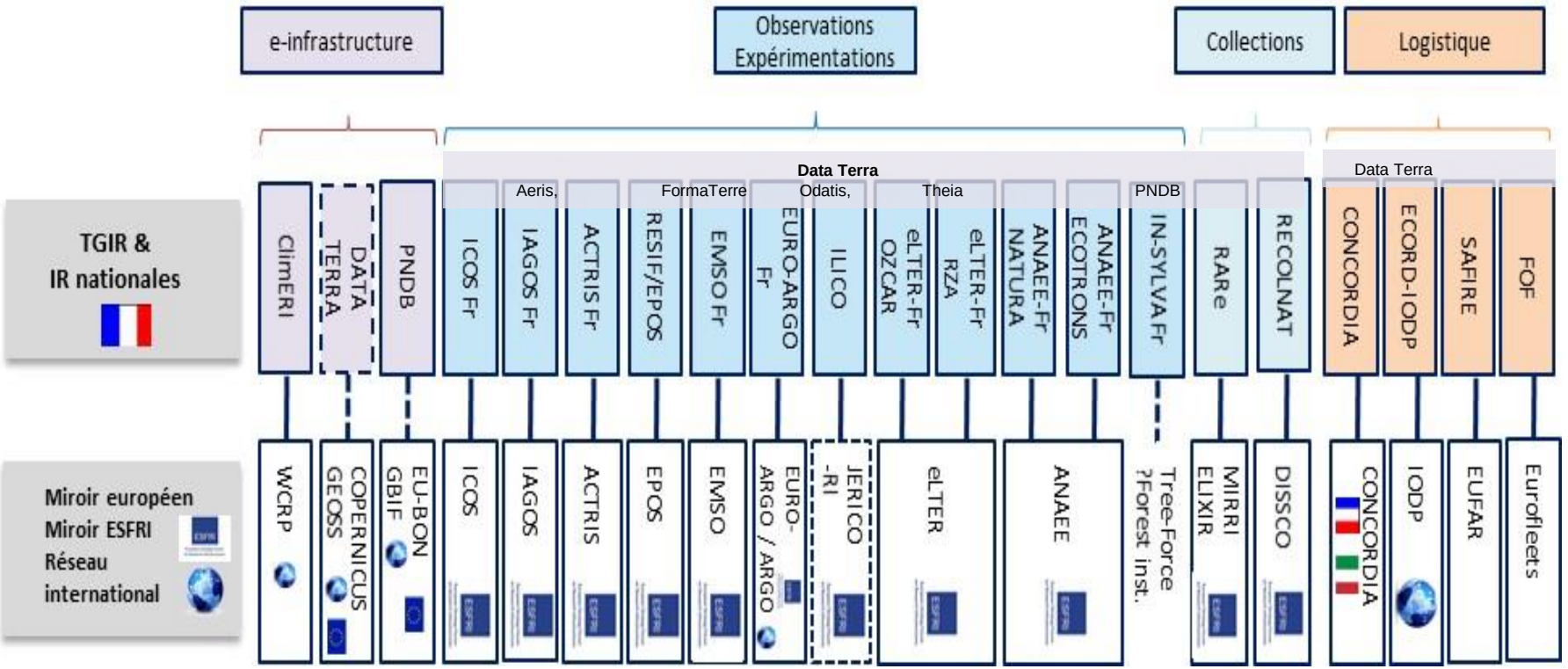

DATA TERRA
Toutes les DONNÉES du SYSTÈME TERRE
Terre solide
Biodiversité
Océan
Imagerie spatiale
Climat
Atmosphère
Pôles
Explorer des données
Croiser
Calculer
Gaïa Data
PLATEFORME INTÉGRÉE numérique
projet structurant de la recherche
Infrastructures TECHNIQUES
Brest
Lille
Paris
Orléans
Strasbourg
Grenoble
Toulouse
Montpellier
SUPER CALCULATEURS
2 sites principaux pour STOCKER les DONNÉES
PARTAGE d'EXPERTISE
STOCKAGE PASSIF
Pour ANALYSER des problèmes COMPLEXES
multi sources
multi impacts
Grands ENJEUX de SOCIÉTÉ
pour FACILITER le TRAVAIL de recherche
des DONNÉES FAIR*
Findable TROUVABLES
Accessible ACCESSIBLES
Interoperable INTEROPÉRABLES
Reusable RÉUTILISABLES
26 tutelles
Institutions, Instituts, Universités
interactions européennes
Coordination HUMAINE
Créer une CULTURE COMMUNE
Elles sont FAIR mais données !
Elles sont bien caractérisées !
TRANSMETTRE les BONNES PRATIQUES pour GARANTIR la QUALITÉ des DONNÉES
COORDINATION
RENCONTRES
IDENTIFIER les BESOINS des CHERCHEURS et CHERCHEUSES pour COCONSTRUIRE les OUTILS
FORMATIONS

Positionnement des Infrastructures de Recherche du domaine système Terre et environnement - France



MESRI - FRN 2022-2025 - HC-IR, mai 2021

Positionnement des Infrastructures de Recherche du domaine système Terre et environnement - Europe



MESRI - FRN 2022-2025 - HC-IR, mai 2021

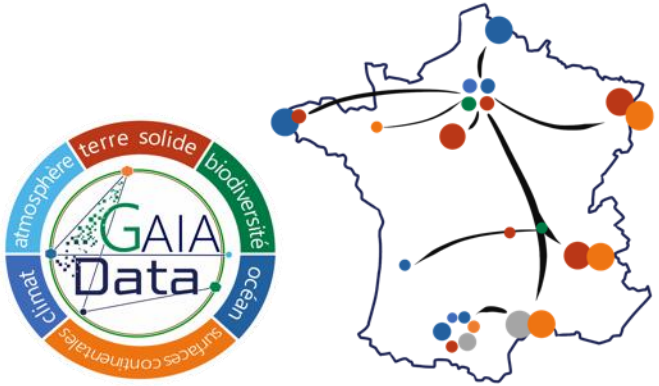
PROJET PORTÉ PAR TROIS E-INFRASTRUCTURES DE RECHERCHE DU DOMAINE SYSTÈME TERRE ET ENVIRONNEMENT

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

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.
> Depuis 2024 5e pôle de DATA TERRA

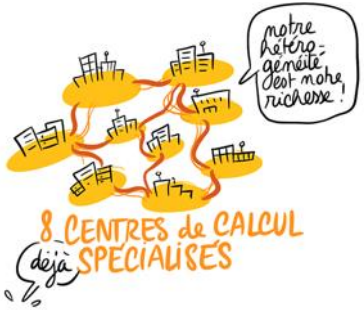
CLIMERI-France produit des simulations numériques internationales pour le Programme Mondial de Recherche pour le Climat et met leurs résultats à la disposition de divers utilisateurs en France et à l'étranger.

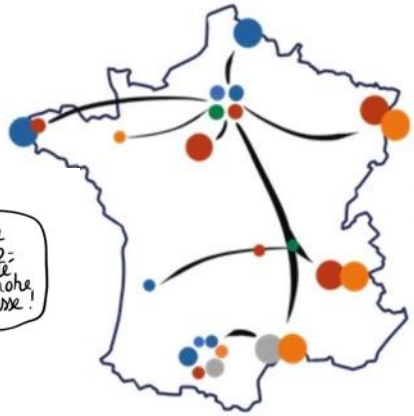
Infrastructure distribuée de services



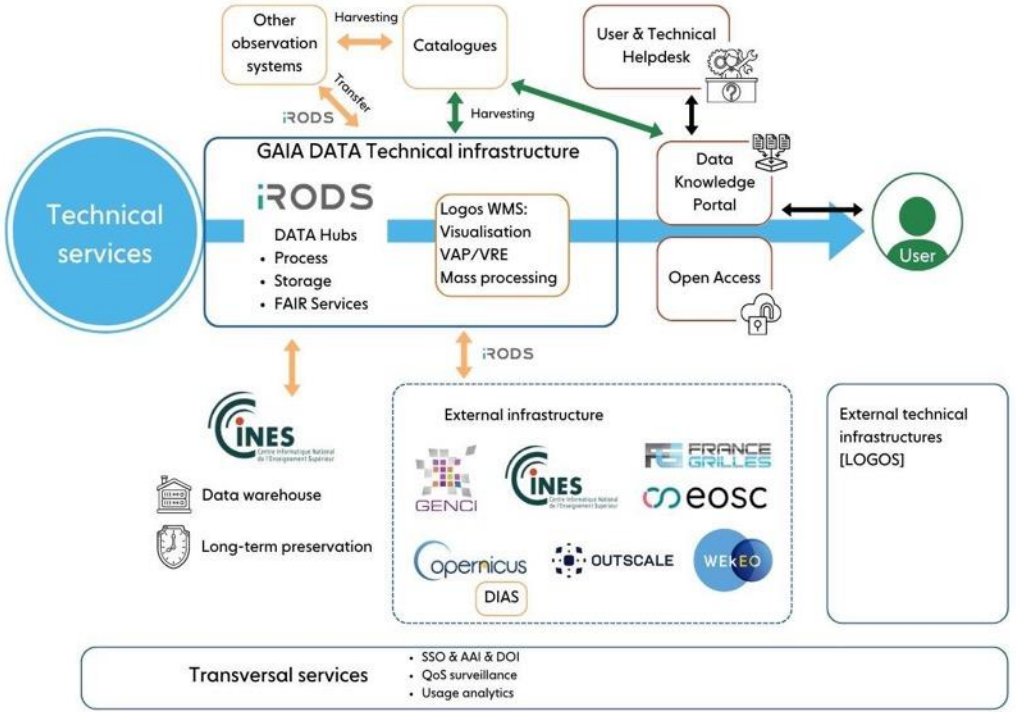
Mettre en œuvre au plan national, européen et international une infrastructure distribuée de services innovante du domaine système Terre et environnement

UN PROJET PORTÉ PAR CLIMERI, DATA TERRA ET PNDB POUR FACILITER LES USAGES DES DONNÉES SYSTÈME TERRE ET ENVIRONNEMENT





8 sites principaux
30 sites existants



INFRASTRUCTURE GAIA DATA

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

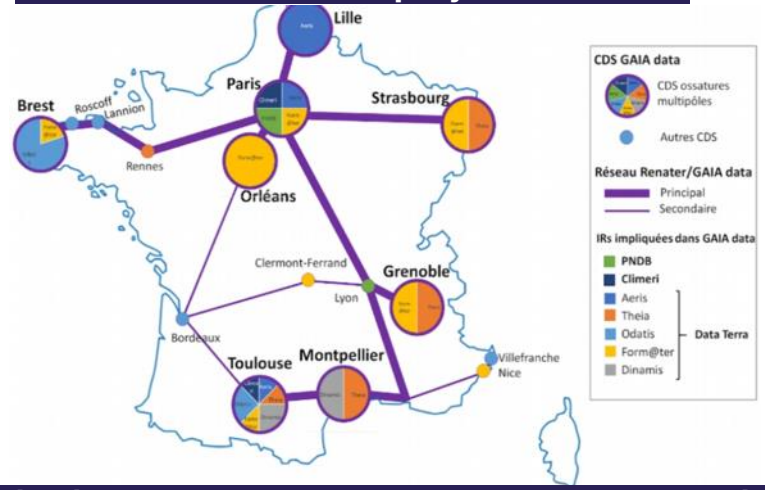
Projets Equipex+ ou PIA4 infra
- FITS
- MesoNet
- Clusster

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

En relation avec des projets connexes



Intégré dans le paysage international / Européen

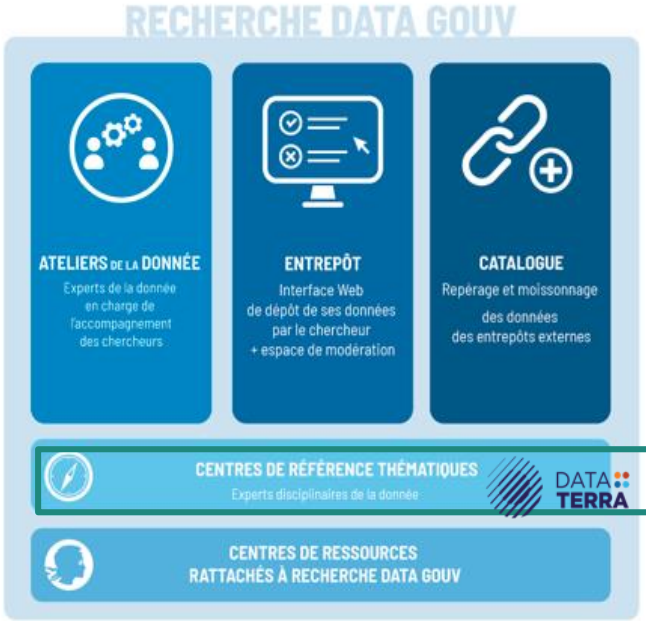
Copernicus · GEO · EOSC · CEOS · EUDAT · EGI · FRANCE GRILLES · EuroHPC Joint Undertaking · HORIZON EUROPE · PRACE · DESTINATION EARTH · GAIA-X

- **Mise en place d'un réseau dédié haut-débit et sécurisé**
- **Déploiement d'une grille de données (système iRODS AC) / S3** 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)

Recherche Data Gouv

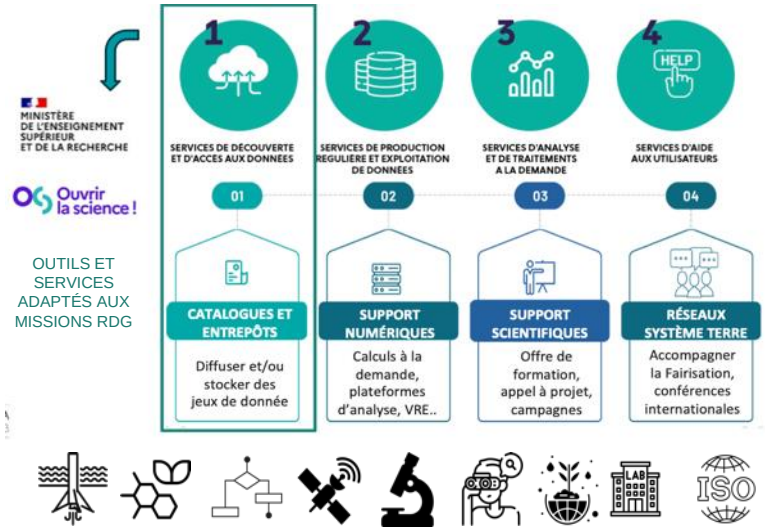
Data Terra => Centre national de Référence Thématique système Terre et Environnement

Un écosystème au service du partage et de l'ouverture des données de recherche FÉDÉRER, ACCOMPAGNER, PARTAGER, OUVRIR, RÉUTILISER



Les centres de référence thématiques de *Recherche Data Gouv* :

- Définissent les normes internationales de description des données
- Définissent les normes internationales de diffusion des données (ouverture, période d'embargo, accès restreint)
- Définissent et diffusent les bonnes pratiques de collecte, documentation, traitement, et diffusion des données
- Définissent la liste des entrepôts de données de référence de leur domaine thématique (nationaux et/ou internationaux) vers lesquels orienter les chercheurs pour le dépôt des données et que *Recherche Data Gouv* moissonnera
- Contribuent à la définition de l'arborescence thématiques des données de l'entrepôt *Recherche Data Gouv*
- Soutiennent l'articulation entre les dispositifs thématiques spécialisés et *Recherche Data Gouv*



L'ENTREPÔT DES DONNÉES DE LONGUE TRAÎNE DU SYSTÈME TERRE ET DE L'ENVIRONNEMENT

POUR QUI ?

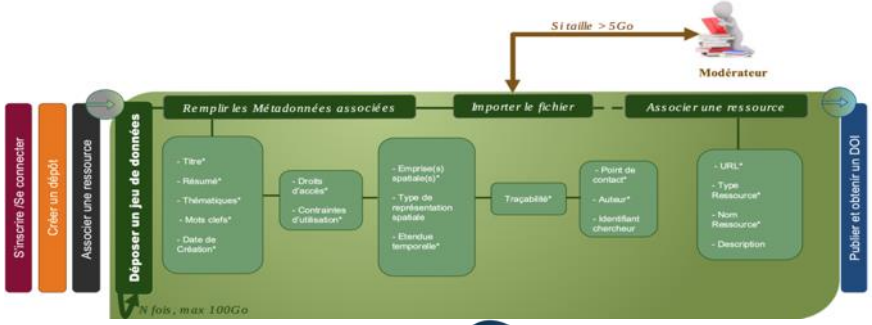
- ✓ LES SCIENTIFIQUES DU SYSTEME TERRE ET ENVIRONNEMENT
- ✓ LES PARTENAIRES NATIONAUX, EUROPÉENS ET INTERNATIONAUX

QUEL TYPE DE DONNÉES ?

- ✓ DONNÉES LONGUE TRAINE
- ✓ ISSUES DE RECHERCHE EN VUE DE PUBLICATION
- ✓ ISSUES DE PROJETS

Un procédé simple en 5 étapes clés

Un modérateur pour accompagner le déposant



Pérennisation et visibilité des données :

- Stockage adapté et diffusion de données de qualité
- Attribution de DOI aux données déposées / citation
- Référencement des données déposées dans le catalogue de la plateforme nationale fédérée

Simplicité du dépôt :

- Authentification via EduGAIN/Renater, ORCID
- Prise en compte d'identifiants chercheurs,
- Dépôt en quelques étapes
- Modération et accompagnement par des spécialistes

Qualité des métadonnées et interopérabilité :

- Norme ISO 19115-3,
- Interface adaptée aux spécificités des données géo-référencées
- Thesaurus communautaires
- Services avec possibilité d'associer des ressources aux données (publication, code source, site web projet, ...)

Les particularités des données dans le domaine système Terre et environnement

- Une donnée d'observation est une donnée **unique** dans l'espace (lat, lon, altitude ou profondeur) et le temps -> Impossible à reproduire

Constat immédiat : Si elle est perdue, elle est perdue à tout jamais!

Conséquence immédiate : Doit-on la conserver? **OUI**

Combien de temps garde-t-on les données d'observation? ⟹ **ad vitam eternam!**

Les questions que l'on doit se poser (recommandation):

- Quelles données doit-on conserver?
- Qui doit la conserver?
- Comment devons nous la conserver?

- Acquisition
- Traitement (Calibration)
- Traitement (Validation)
- Distribution
- Pérennisation

Application des principes FAIR

Les particularités des données dans le domaine système Terre et environnement

Qui doit la conserver? **Et comment la conserver?**

Au plus près du producteur de la donnée (principe FAIR - le producteur est celui qui connait le mieux « sa » donnée et qui est capable expliquer son contenu, son traitement,…).

1. Donnée brute (épurée de la donnée scientifiquement non exploitable?) -> le producteur ou un entrepôt d'un OSU, la structure qui traite la donnée
2. On conserve les données intermédiaires le temps du projet (Est-ce mieux de la conserver ou de la recréer!! Mise en balance des coûts énergétiques -> Ecoinfo!)
3. On fait du nettoyage après le projet! (des centaines de TO voire plus dorment dans les laboratoires et ne seront JAMAIS réutiliser)
4. Donnée distribuée -> un entrepôt d'un centre de donnée et de service, DOI, archive pérenne (Pôle de données, CDOS, entrepôt certifiés (Core Trust Seal)

Pour les entrepôts de donnée: Entreprendre une certification (Core Trust Seal)!

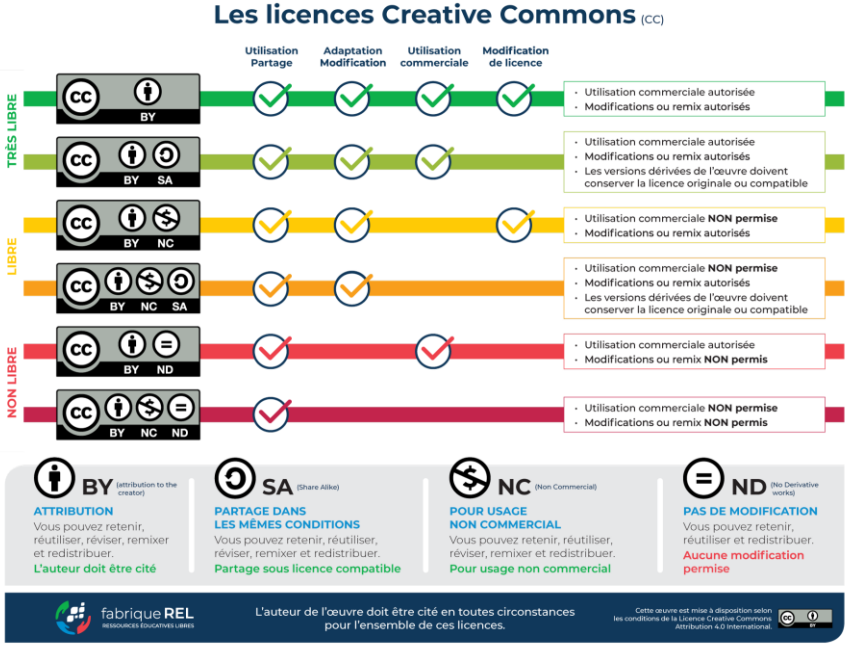
ABANDONNER l'usage des disques amovibles, des clés usb, du disque dur du PC ! SVP

Les particularités des données dans le domaine système Terre et environnement

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté Égalité Fraternité

Comment devons nous la conserver? (donnée distribuée)

1. Un format d'échange (format ouvert et interopérable: NetCDF, HDF, TSV,mp3,mp4,…) **Compression Oui…ou Non? (coût énergétique)**
2. Un vocabulaire contrôlé (donnée ET métadonnée) -> thésaurus, web sémantique?, FAIR pour IA,…), chaque paramètre doit répondre à une définition claire et un protocole d'obtention bien définie
3. Un DOI (traçabilité, bien identifier chaque personne ayant contribuer à son cycle)
4. Une licence (Etalab 2, Creative Commons)



Le chemin qu'il reste à faire!

Sur les données du système Terre et de l'environnement les jeux de données (+ de 15000!) sont encore très hétérogènes:

- Nécessité de « FAIRISER » les données ET les métadonnées
- Utilisation du web sémantique et des serveurs de vocabulaire
- Avoir des ontologies qui se correspondent entre les différents domaines
- Mettre en place des formations sur les bonnes pratiques et sur le FAIR:

  ➡Rendre l'utilisation des entrepôts nationaux courants
  ➡Changer les pratiques sur les données
  ➡Prendre l'habitude d'écrire un PGD

Quelques conseils et informations complémentaires!

Ecrire un plan de gestion de donnée permet de se poser les bonnes questions!

- Permet de planifier l'arrivée de nouvelle donnée
- De structurer une offre de service
- D'avoir une visibilité à long terme
- Rassurer le producteur de donnée
- D'assurer la protection de la donnée (propriété intellectuelle)
- D'avoir un retour sur l'usage de la donnée

✓ Data Terra : Centre de de référence thématique de Recherche Data Gouv pour les données du Système Terre: **https://recherche.data.gouv.fr/fr/page/centres-de-reference-thematiques-expertises-par-domaine-scientifique**

✓ Recensement des services accompagnant la rédaction des PGD au sein des établissements d'enseignement supérieur et de la recherche: **https://scienceouverte.couperin.org/sos-pgd/**

✓ Rédaction de votre PGD (fortement conseillé): **https://dmp.opidor.fr/**

Utilisation du MADMP (PDG machine actionnable - **https://dmp.opidor.fr/help#content**)

✓ Aide à l'amélioration du contenu de votre PGD : **DoRANum – Grille de relecture de PGD**

✓ Ressources pédagogiques DoRANum: **https://doranum.fr**

Merci de votre attention!

contact@data-terra.org
+33 (0)4 67 54 87 08

www.data-terra.org

Les particularités des données dans le domaine système Terre et environnement

Quelles données doit-on conserver?

Est-ce que à un moment on supprime des choses et si oui quoi ?

Les données d'observation et leur cycle interne de l'acquisition à sa réutilisation:

1. La donnée est acquise par un capteur (un instrument ou un observateur) -> donnée brute (raw data)
2. La donnée est calibrée, validée, traitée, FAIRisée pour être exploitable scientifiquement
3. La donnée est ensuite mise à disposition et pérennisée

1. Donnée brute: Contient des données inutilisables mais permet d'être retraitée si les algorithmes de traitement change ⟹ **Toutes les données inutilisables à supprimer! (Quality control)**
   (ex :https://www.seadatanet.org/Standards/Data-Quality-Control)

2. Donnée distribuée: Scientifiquement exploitable, citer dans les articles, jeu de donnée avec DOI (et ses différentes versions!) ⟹ **Eviter de multiplier les copies du jeu de donnée (Des bonnes pratiques à mettre en place avec les éditeurs et les entrepôts)**

Un exemple de table de QA

Key	Entry Term	Abbreviated term	Term definition
0	no quality control	none	No quality control procedures have been applied to the data value. This is the initial status for all data values entering the working archive.
1	good value	good	Good quality data value that is not part of any identified malfunction and has been verified as consistent with real phenomena during the quality control process.
2	probably good value	probably_good	Data value that is probably consistent with real phenomena but this is unconfirmed or data value forming part of a malfunction that is considered too small to affect the overall quality of the data object of which it is a part.
3	probably bad value	probably_bad	Data value recognised as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena.
4	bad value	bad	An obviously erroneous data value.
5	changed value	changed	Data value adjusted during quality control. Best practice strongly recommends that the value before the change be preserved in the data or its accompanying metadata.
6	value below detection	BD	The level of the measured phenomenon was too small to be quantified by the technique employed to measure it. The accompanying value is the detection limit for the technique or zero if that value is unknown.
7	value in excess	excess	The level of the measured phenomenon was too large to be quantified by the technique employed to measure it. The accompanying value is the measurement limit for the technique.
8	interpolated value	interpolated	This value has been derived by interpolation from other values in the data object.
9	missing value	missing	The data value is missing. Any accompanying value will be a magic number representing absent data.
A	value phenomenon uncertain	ID_uncertain	There is uncertainty in the description of the measured phenomenon associated with the value such as chemical species or biological entity.