Problématiques de certification des réseaux de centres de données et de services de l'IR Data Terra

G. Maudire, A. Chambodut, J. Sudre

2ème réunion annuelle du nœud national RDA France

Vendredi 13 septembre 2019

Contexte

Infrastructure de Recherche «Pôles de Données et Services pour le Système Terre»

Au niveau national :

- en accord avec la **Stratégie nationale de recherche** France Europe 2020 (2015)
- sur la **feuille de route nationale** (2016-2018) et (2018-2020) des IR du MESRI

Au niveau européen et international :

- **Roadmap ESFRI** (*European Strategy Forum* on RI)
- Développements de services d'accès aux données spatiales et in-situ (EOSC, COPERNICUS/DIAS, ...)
- Internationalisation des dispositifs de partage des données et services (GEO/GEOSS,...)

Enjeux

Observer, comprendre et prévoir de manière intégrée l'histoire, le fonctionnement et l'évolution du système Terre soumis aux changements globaux

- **Globalité de l'observation** du Système Terre (approches intégrées, multi-sources, multi-capteurs, …) à **long terme**
- Augmentation exponentielle du **volume de données** (in-situ & spatiales)
- Besoins en analyses et traitements **intelligents et systémiques** (*deeplearning*, *big data*, intelligence artificielle,...)

➢ Favoriser l'interopérabilité, l'émergence **d'approches multi- et inter-disciplinaires** et l'innovation pour des avancées scientifiques

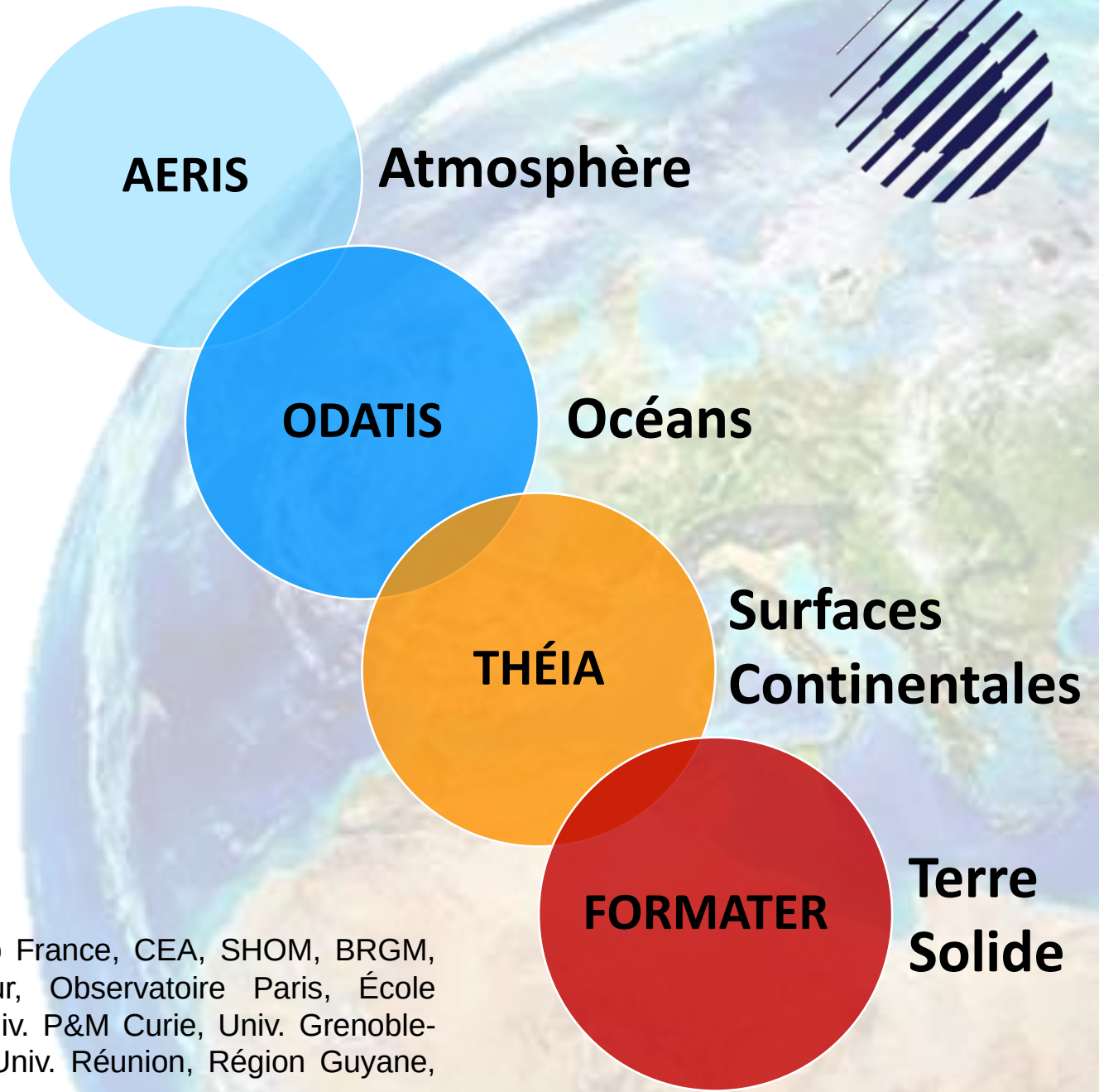
➢ Concilier **recherche d'excellence** et développement de **partenariats durables avec acteurs publics et économiques pour la société**

➢ Structurer et organiser l'offre nationale (inter-organisme) tout en se **projetant au niveau européen et international**

I.R. Data Terra

4 pôles

- **Associer** l'expertise scientifique aux techniques de gestion des données
- **Partager** les informations et les bonnes pratiques
- **Coordonner et fédérer** au sein d'une même IR, l'ensemble des institutions, dispositifs et moyens existants
- **Servir** la recherche nationale et internationale sur la planète, l'environnement, le climat, les risques, …



Plus de 30 partenaires

(CNES, CNRS, IFREMER, IRD, IGN, IRSTEA, INRA, IPGP, Météo France, CEA, SHOM, BRGM, CEREMA, CIRAD, INERIS, ONERA, Observatoire Côte d'Azur, Observatoire Paris, École Polytechnique, Univ. marines, Univ. Lille-1, Univ. Féd. Toulouse, Univ. P&M Curie, Univ. Grenoble-Alpes, Univ. Clermont-Auvergne, Univ. Strasbourg, Univ. Guyane, Univ. Réunion, Région Guyane, Région Hauts de France)

Projets vers la certification

cadre de l'Appel Flash "Science Ouverte : Pratiques de Recherche et Données Ouvertes" de l'

ANR AGENCE NATIONALE DE LA RECHERCHE

ODATIS ➔ **COPiLOtE** – "***C****ertificati****O****n* ***P****o****L****e* ***O****c****E****an*"

ForM@Ter ➔ **CEDRE** – "*towards* ***C****ertification of solid* ***E****arth* ***D****ata* ***RE****positories in France*"

- Projets de **même philosophie** visant à **promouvoir les bonnes pratiques** de *Data Management* d'une manière durable dans une démarche «*FAIR*» auprès des communautés scientifiques et techniques

Certification et architectures distribuées

Réseaux

organisations fédératrices
représentant des groupes de CDS

agents de coordination
pour des nœuds aux caractéristiques communes et disciplines connexes

➔ ***WDS Network member***

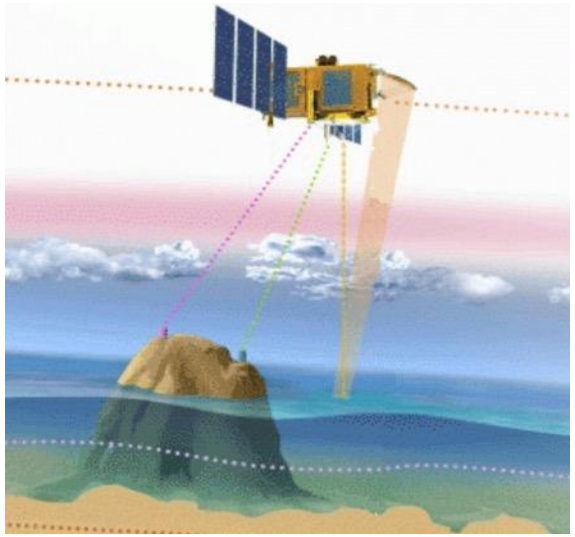
Centre de Données et de Services (CDS)

répertoires de données,
gestionnaires de données et/ou services d'analyse de données

➔ ***CoreTrustSeal Certified Repository***

(➔ avant 2017:
WDS Regular Member)

Projet COPiLOtE - Contexte

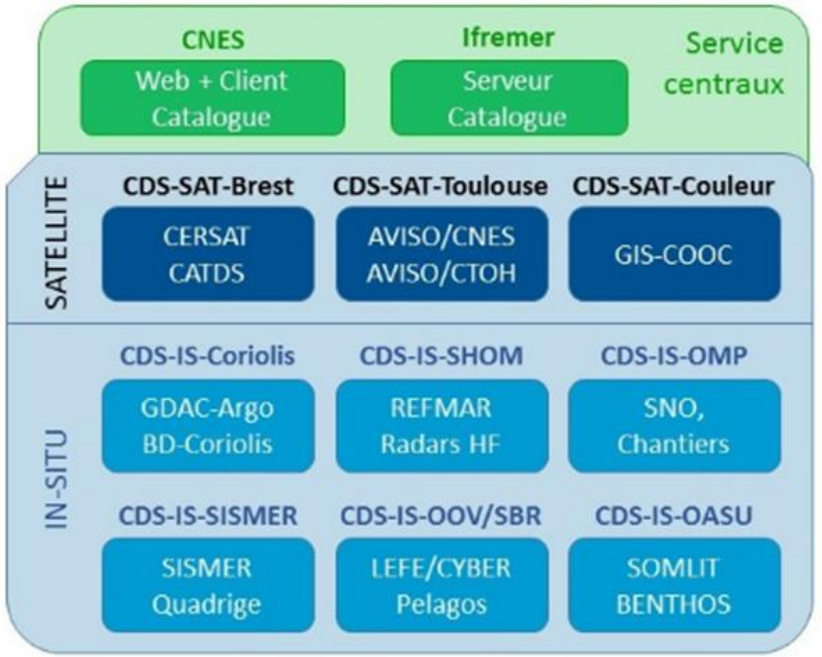


- **Pôle de données Odatis : la gestion des données marines**
  - du littoral au hauturier
  - de la surface au fond, avec les interfaces : terre/mer, océan/atmosphère, sous-sol sous-marin
  - Physique, Chimie, Biologie
  - dans différents compartiments : Eau, Sédiments, Biota

- **Un milieu restant difficile et coûteux à observer**
  - **Collaboration étroite entre les missions satellites «marines» et les observations in-situ**
  - IR d'observations :

    Flotte Océanographique Française, Argo et futur O-Hisse (hauturier), Ilico (côtier), EMSO (fond de mer), ...
  - Systèmes d'observations labellisés (SNO et SO)
  - Nombreuses observations moins structurées : objectif scientifique ponctuel, appui aux politiques publiques mais contribuant néanmoins à la connaissance du milieu

Projet COPiLOtE - Acteurs et objectifs



- **9 Centres de Données et de Services (CDS)**
  - 3 satellites et 6 in-situ,
  - opérés par 6 partenaires (+ 1 en rapprochement), souvent en liaison avec de grands programmes européens et internationaux

- **Grandes hétérogénéités structurelles et culturelles**
  - Sources de données variées (du microscope au satellite),
  - Champs disciplinaires multiples,
  - Infrastructures organisationnelles et techniques très différentes (du méga-octets au péta-octets),
  - Plusieurs degrés de maturité dans les pratiques de gestion des données, …
  - Méthodes hétérogènes (exemples *DOI*, …),

- **Nécessaire harmonisation**

  pour faciliter la vie d'utilisateurs accédant à des jeux de données multiples, **selon deux axes** convergents :
  - « *FAIR*-isation » des données
  - Certifications des CDS

Projet COPiLOtE - Méthode

« Auto-évaluation » initiale puis soumission d'un dossier à *CoreTrustSeal* :

- Critères de *CoreTrustSeal*
- Principes « FAIR »
- Décliné dans le cahier des charges des CDS

Production de guides de « bonnes pratiques »
(Ateliers préexistants : InterPôles et Odatis + experts)

- Métadonnées et vocabulaires
- DOI
- Formats supportés
- Services
- Outils
- Pérennité et sécurité

Implémentation progressive

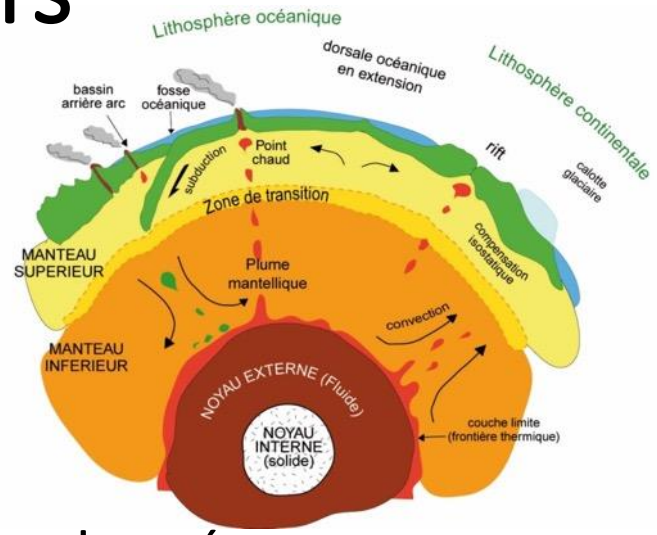
- 3 Centres Pilotes
- 3 cours par an / ensemble des CDS
  - Coordinateurs
  - *Data managers*
  - Informaticiens
- Support en continu

➢ Les CDS pilotes ont finalisé la procédure de certification *CoreTrustSeal* (critères auto-évalués au niveau 3 minimum)

➢ L'ensemble des CDS finalisent un dossier de certification *CoreTrustSeal*, sans obligation de soumission, tous les CDS améliorant de manière significative leurs pratiques en gestion de données

DATATERRA FORMATER

Projet CEDRE – Contexte & Acteurs



- **Grandes hétérogénéités structurelles et culturelles**
  - multiples disciplines,
  - nombreux partenaires,
  - différents répertoires de données,
  - divers formats de données et méta-données,
  - méthodes hétérogènes (*DOI minting*, Web service, …),
  - plusieurs degrés de maturité dans les pratiques de gestion des données, …
- Travaux d'amélioration de fond (comme la certification) souvent individuellement repoussé face à la **pression de l'opérationnel**

- **Consortium de 13 centres de données et de services** (CDS)
- tous **internationaux** ou ayant des liens internationaux (EPOS, EMSO, …)
- **correspondants à des Services Nationaux d'Observations (SNO) ou à des répertoires institutionnels**

Projet CEDRE – Acteurs

Géodésie
Géologie
Géomagnétisme
Géophysique Marine
Géothermie
Gravimétrie
Sismologie
Volcanologie
…

BCMT

FRANCE

RGP

Projet CEDRE – Objectifs & Méthode

- minimum de 2/3 des CDS impliqués finalisant leur dossier de certification,
- 100% des CDS impliqués améliorant de manière significative leurs pratiques en gestion de données

---

- approche **coopérative** → transformer le processus individuel de préparation à la certification en un effort collectif,
- répondre aux questions communes comme individuelles
- **travail conjoint et croisé de révision** → transfert de compétences et de savoir-faire

Chaque CDS engage des ressources pour mettre en œuvre les améliorations requises en vue de la certification *CoreTrustSeal*

DATATERRA FORMATER

Projet CEDRE - Programme

Approche très pragmatique sous la forme de 3 ateliers

➔ 1 binôme par CDS : [scientifique-*Data Scientist]* + [technique - *Data Manager]*

1er atelier (fév. 2020) :

- informations sur la certification *CoreTrustSeal* (experts externes)
- mise en place d'une ontologie et d'un vocabulaire commun
- travaux en petits groupes pour identifier les forces et les faiblesses de chacun
- matrice d'amélioration

2ème atelier (~oct. 2020) :

- lecture croisée et évaluation des premières ébauches de dossiers
- identification des points bloquants restants pour lever les ultimes verrous

3ème atelier (~juin 2021) :

- finalisation des dossiers de soumission à la certification

Plusieurs niveaux d'intérêts à l'outil qu'est la certification

CDSs

Reconnaissance locale et Internationale

- Profil renforcé et exposition mondiale
- Visibilité accrue
- Amélioration des perspectives de financement

Engagement manifeste à l'égard de la science ouverte

- Adhésion au principe Open Science & FAIR
- Engagement envers la qualité et la gestion à long terme

Performances et souplesse accrues

- Interactions et échange de données facilités (interopérabilité)
- Découverte et citation améliorées (réputation et reconnaissance)
- Pratiques et processus améliorés

Utilisateurs *Data Provider*

Reconnaissance (paternité des données, propriété, licence, …)
Conservation de qualité assurée à long terme
…

Utilisateurs *Data User*

Visibilité (métadonnées répondant à l'état de l'art)
Performances (interopérabilité)
…

Ce sont des Initiatives en cours!…

Projets ANR (**Kick-offs prévu pour fin 2019 – début 2020**) constituent une 1ère étape :

- Accent mis sur la certification de chaque CDS
- Concertation avec Data Terra dans son ensemble (concertation inter-Pôles pour thésaurus, portail, ...)

➢ **Consolider un système réparti, vers un plus haut degré d'harmonisation des CDSs**

Nécessaires implications de nombreuses parties prenantes :

- au sein des pôles : coordinateurs des CDS, *Data Scientists*, *Data Managers*, IT, …
- au niveau de la Direction de l'IR et de chacun des 4 Pôles
- Organismes de tutelle : Mandats, Politiques de Données, Licences, ...
- Le cas échéant, implication de ressources externes : experts, infrastructures informatiques, ...

➔ Création d'un groupe RDA – France : « Données d'Observation de la Terre » ?