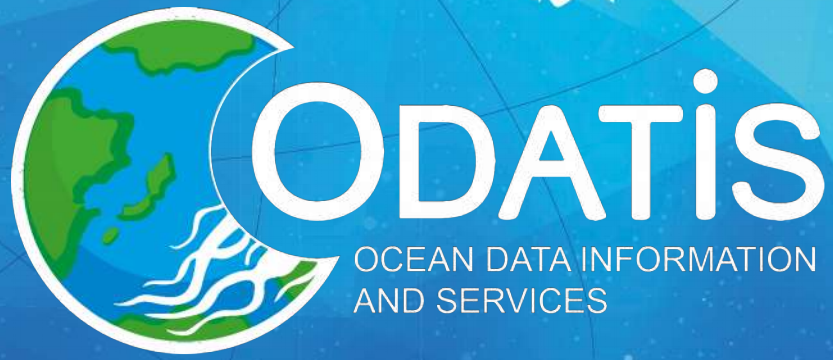
contact@odatis-ocean.fr
www.odatis-ocean.fr



Architecture Odatis
21 novembre 2018

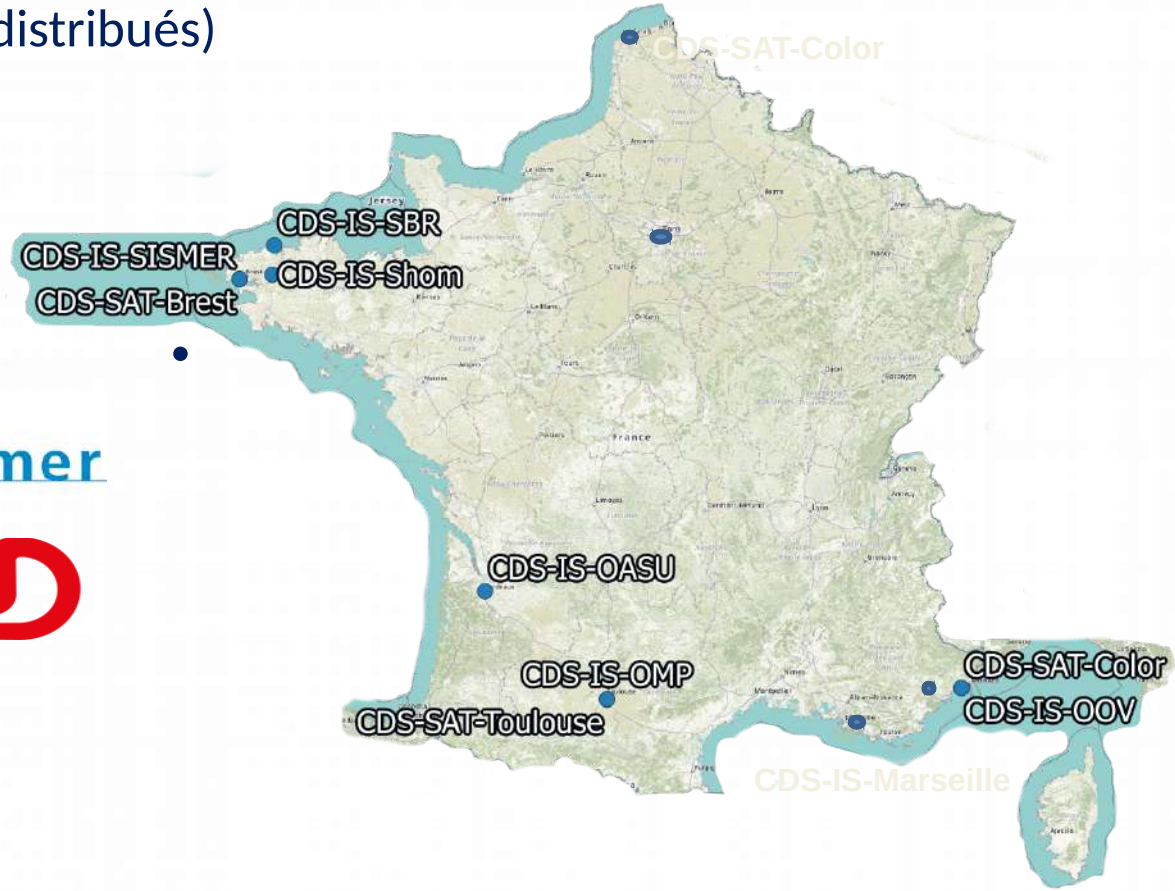
Services de traitement

- **Objectif : fournir une capacité de traitement à la demande « là où les données sont dsiponibles »**
  - Suites logicielles
  - Permettant d'accéder aux seules données utiles (subsetting)
  - Permettant par exemple de visualiser conjointement des données gérés par plusieurs CDS
  - Proposant des « accès directs» aux données (non limités aux téléchargements) suivant des protocoles variés (Cloud, OGC, OpenDap, ...)

- **Exemples**
  - Programmation/ Notebooks : Jupyter (Python), Rstudio (R)
  - Suites logicielles :
  - Pangeo : écosystème python pour les sciences ocean / atmosphere / surfaces continentale/ climat (https://pangeo.io)
  - OceanWorks : Découverte et analyse de jeux de données (visualisation, browsing via Cassandra, …) (*https://github.com/aist-oceanworks*)
  - SeaDataNet : DIVA (Analyse Géostatistique), Ocean Data View Online
  - Galaxy (bioinfo, génomique)

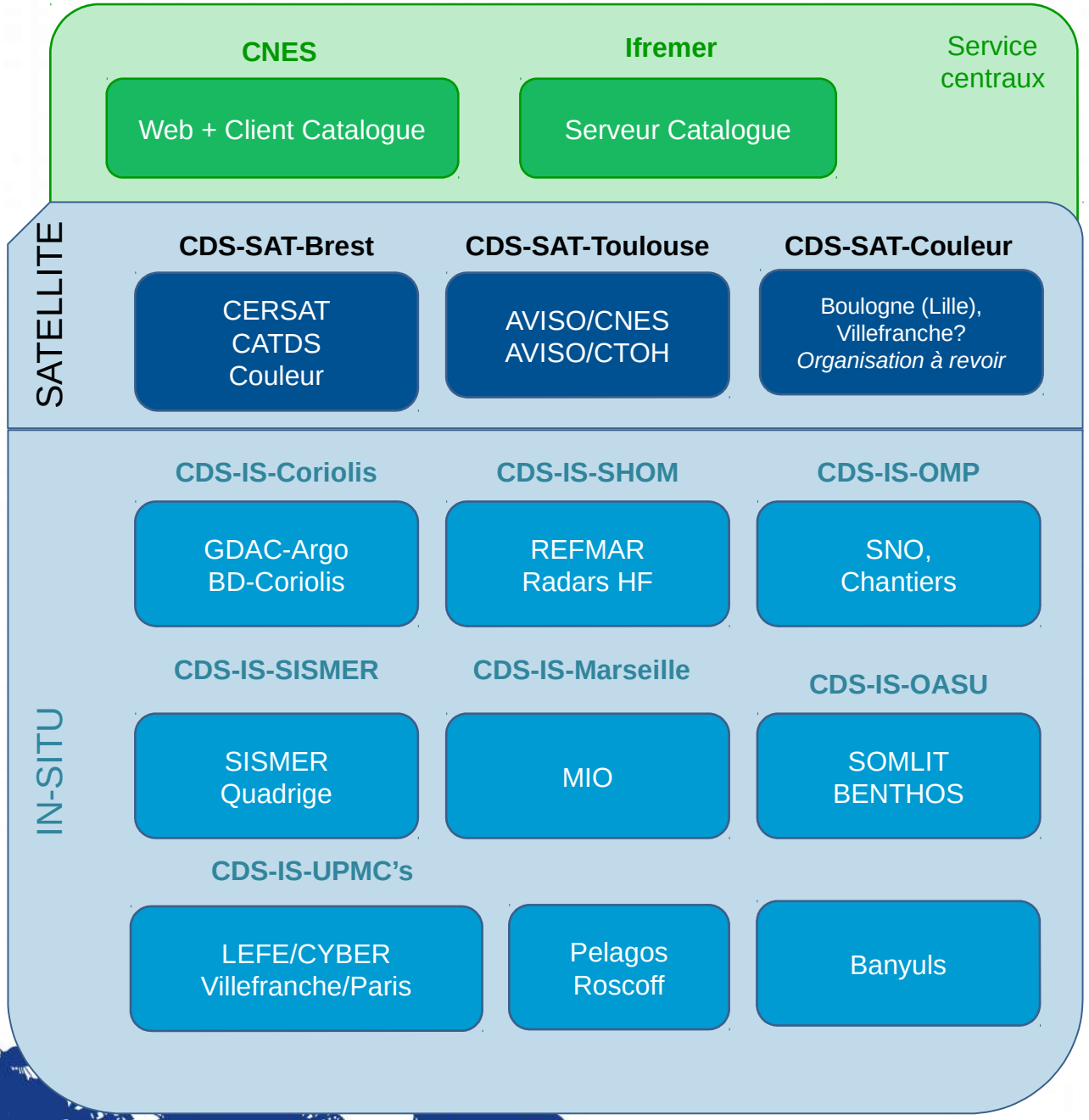
Une infrastructure technique très distribuée

- Centres de données et de services (eux même distribués)



Universités marines

Les centres de données et de service



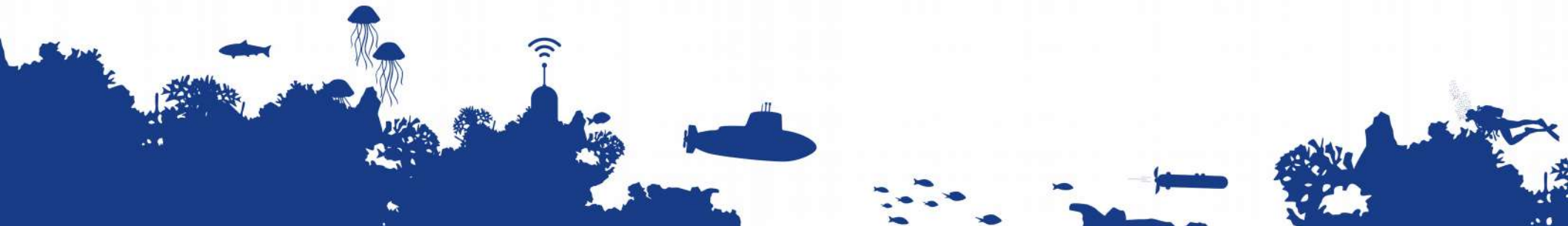
Des données différentes

- **Données satellites (et modélisation)**
  - Forte volumétrie (PetaOctets): capacité de stockage, d'archivage pérenne
  - Capacité de traitement (en particulier pour l'échelle globale)
  - Métadonnées limitées (description de la mission), sauf imagerie à la demande

- **Données in-situ**
  - Volumétrie plus faible (sauf imagerie optique et acoustique) (GigaOctets)
  - Hétérogénéité et complexité importante ⇒ données très structurées :
    - importance accrue des métadonnées et des référentiels (paramètres, unités, instruments, méthodes…)
    - structures « métier » : séries temporelles, profils verticaux, ...
  - Renforcer l'archivage pérenne des données et leur valorisation au-delà de l'usage premier
    - 60% des données d'observation en mer ne sont pas « archivées »
    - Au bout de 5 à 10 ans, les données non archivées dans un centre « ad-hoc » sont perdues (défaillance et obsolescence des supports, mobilités professionnelles….)

Existant : Données satellite

- **Volume global**
  - **Environ 2500 Toctets : Toulouse, Brest + couleur de l'eau (comprend des L1)**
  - **Augmentation annuelle actuelle , environ 500 Teratoctets**
  - **Ne prends pas en compte les nouveaux satellites : Sentinel 3, SWOT, …**
- **Capacité de calcul importante nécessaire**
- **Structures de données**
  - **NetCDF (3 et 4) essentiellement**
  - **Peu de métadonnées : satellite/mission/producteur**
  - **Métadonnées : ISO 19115/19139-version 2/3 ? (en cours au CDS-Brest), fichiers d'index (Thredds par exemple)**
- **Protocoles d'accès**
  - **FTP**
  - **OpenDAP**
  - **Traitements « proche des données », conteneurs (Docker), Notebooks (Jupyter)**



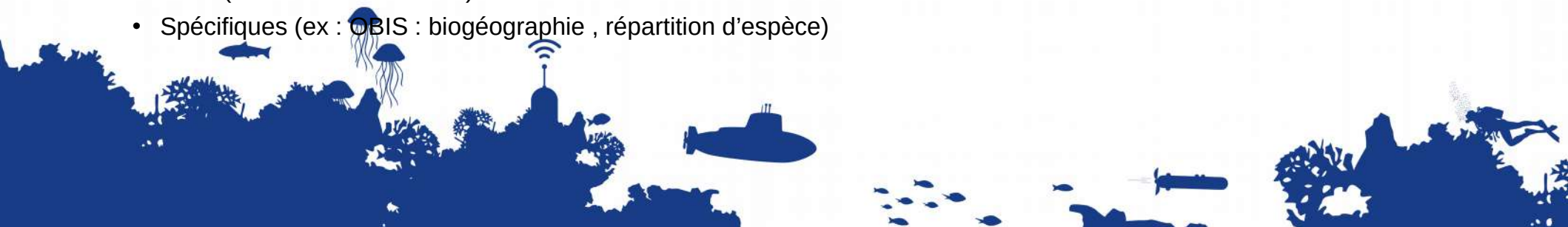
Existant : Données in-situ

- **Volume global**
  - Environ 300 Toctets :
    y compris imagerie optique (vidéos, cytométrie, ...) et acoustique (sondeur, sonar, sismique)
  - Augmentation annuelle actuelle , environ 70 Teratoctets

- **Peu de besoin de puissance de calcul**
  - Serveurs ou machines virtuelles (Linux) pour services en ligne, serveurs de fichiers, bases de données

- **Structures de données**
  - Bases de données relationnelles (chimie, biologie, métadonnées) : Postgre/Postgis, Oracle, MySQL
    et formats textuels « à colonnes » (type CSV)
  - NetCDF (3 et 4) (physique, bathymétrie)
  - Formats SEG (Sismique)
  - Formats images et videos
  - Métadonnées (si structurées) Dublin Core, ISO 19115/19139, fichiers d'index (Thredds par exemple)

- **Protocoles d'accès**
  - FTP
  - OpenDAP (Physique)
  - OGC (domaine littoral/côtier)
  - Spécifiques (ex : OBIS : biogéographie , répartition d'espèce)



Evolutions stockage/calcul

- **Augmentation « naturelle » (i.e. à capteurs constants)**
  - Données in-situ : 10 à 20 % par an

- **Liées aux nouvelles données**
  - Nouveaux satellites : Sentinel-3B, CFOSAT, SWOT
  - Nouvelles mesures in-situ: radar-HF (10To / an / radar), cytométrie, ... par exemple.

- Augmentation notable de la puissance nécessaire (stockage, calcul) / besoins fonctionnels nouveaux
  - Pérenisation : Double copie (données brutes) ou retraitement dans certains cas)
  - Capacité de traitements conjoints satellite / in-situ
  - Apprentissage : classification/reconnaissance groupe morphologique plancton (ex : Villefranche)
  - Analyses géostatistiques (différences finies, filtres de Kalman, classification « big data » : e.g. masses d'eau)
  - Services d'accès et de visualisation

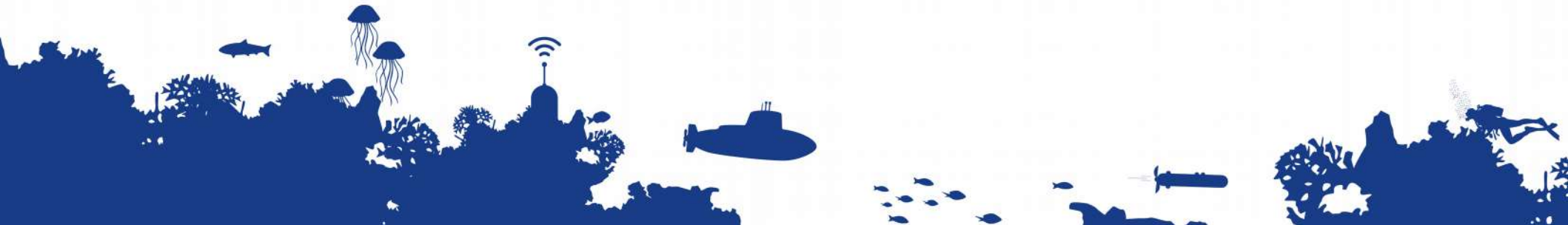
  ⇒ Différentes organisations (imposant duplication) des données pour répondre à différentes utilisation, en particulier pour l'in-situ

Des services variés (1/2)

- **Producteurs**
  - Service d’archivage pérenne de jeux de données « stabilisés »
  - Obligation croissante de déposer les jeux de données associés aux papiers
  - Obtention de DOI
  - Pas trop compliqué à mettre en œuvre : service en ligne, métadonnées simplifiées
  - Succès de Pangeae à Brême ⇒ service à mettre en place

  - Nécessité d’une modération (ne pas archiver des données non adéquates : relecture, données « marines », …)

  Solutions : SeaNoe, Sedoo, logiciel DataVerse, Zenodo (projets européens)

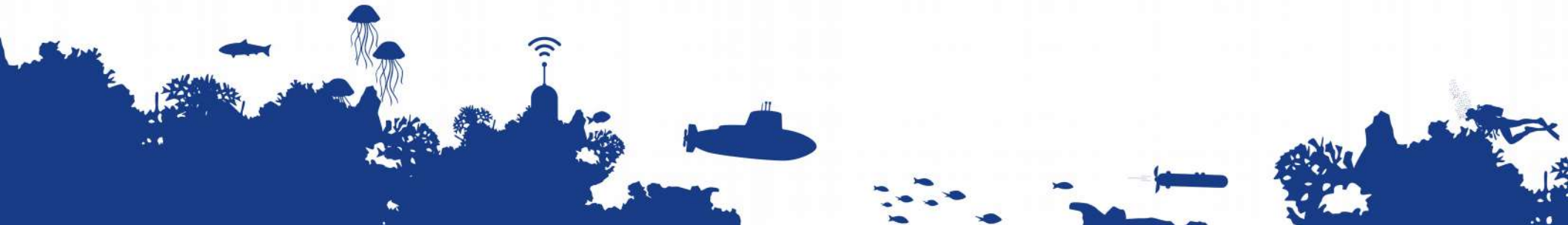


Des services variés (1/2)

- **Utilisateurs**
  - Plusieurs niveaux d'utilisation / plusieurs communautés (recherche, hors recherche)
  - A qui s'adresse le(s) Pôles?
    - Producteur initial, Utilisateur secondaire (reproductibilité des résultats), Intégrateur (Climatologie…), Modélisateur…
- **Services, par ordre d'utilisation**
  - Téléchargements, via le DOI (http), par download (http, ftp)
  - Sélection : Portail (Métadonnées, Facettes ou Pas, Accélérateur : ElasticSearch, …) puis download du sous-ensemble
  - Prévisualisation des données (et des métadonnées)
    - Localisation des poisitions d'observation
    - Prévisualisation de la donnée elle-même
  - Service d'accès directs
    - Fichiers locaux(Calcul Haute Performance), éventuellement Système de Fichier Vituel (Irods…)
    - OpenDAP (physiciens, observation de la terre satellite, support NASA)
    - OGC (WFS, WCS), éventuellement O&M SOS, utilisé dans le domaine côtier (utilisateurs de SIG)
  - Depuis peu, mais une demande forte : physique (Python), biologie (R), …
    - Traitements à la demande (notebooks ou équivalent)
    - Passage à l'échelle d'application développé localement
    - Virtual Research Environment (suites logicielles permettant analyse et traitement)

Un contexte à prendre en compte

- **Une politique nationale de rationalisation des « Data Centres » : InfraNum – MESRI (Patrick Garda)**
  - Trois Centres de niveau « Européen » : Hydris, IN2P3, CINES
  - Des centres régionaux : 1 par région environ : «MésoCentres »
- **Au niveau européen**
  - European Open Science Cloud : Données de la recherche, plutôt in-situ dans le cas des pôles
  - DIAS : Data Information and Access Service : pour l'Océan : Eumetsat – ECMWF - Mercator
- **Obligation pour l'IR « Pôles de Données » : s'adapter / prendre en compte cette politique**
  - Dans le cadre du PIA 3 : un des critères de recevabilité des propositions



Des services hiérarchisés

Observations « individuelles »

Initiatives locales, régionales, ...
Résultats associés à une publication

Dépôt / Publication
- Obtention d'un DOI (Landing page)
- Conservation long terme

Suivi des dépôts
Importation

Observation structurée

Programmes nationaux, européens et internationaux (SNO, IR et TGIR d'observation, ...)
Missions satellites (segments sols)

Centres de Données et de Services
- Curation des données d'observation
- Traitement des observations (en routine)
- Mise à disposition des données

Traitements multi-sources
- Satellite / In-situ
- Constitution/ harmonisation de séries temporelles
- ...

Services avancés de diffusion
- « Cache technique »
- Assemblage multi-source
- Subsetting
- Co-visualisations

Utilisateurs de l'ensemble des services :
- **Accès par DOI,**
- **Requêtage,**
- **Visualisation**
- **Téléchargement**
- **Accès direct**
- **Soumission de traitement**

Services communs (de base)
- Authentification / Autorisation des utilisateurs, sécurisation
- Catalogue(s)
- Serveurs de référentiels : vocabulaires communs (ontologies), fonds géographiques, ...
- Suivi des usages (analyses des logs des services, bibliométrie)

Centres de Données et de Services

- **Le cœur de l'architecture**
- **Objectif : fournir un service thématique de préservation et de production**
  - Garant de la qualité des données et de leur « harmonisation » dans le temps
  - ⁼ métadonnées précises, définition de référentiels communs, contrôle qualité, ...
  - Production régulière de données dérivées
    selon des méthodes établies par les pôles (Conseil Scientifique, Consortium d'Expertise Scientifiques, ...)
  - Garant de l'archivage long terme (avec leurs moyens ou par délégation, en s'appuyant par exemple, sur les services de dépôt)
  - Mise à disposition des données pour leur communauté
  - Interopérabilité technique et sémantique
- **Conditions :**
  - Architecture répartie pour tirer parti des compétences thématiques / géographiques des Centres existants
  - Proximité avec les programmes d'observation (en particulier pour les mesures in-situ)
  - Matériels informatiques, connexion réseau, ... disponibles dans le CDS ou hébergement externe au CDS
  - Organisation interne libre pour répondre aux besoins de la thématique (base de données, …)

➔**nécessaire exigence d'un niveau de service minimal**

➔**adoption de formats, référentiels communs pour les échanges (harmonisation sémantique)**

ex : métadonnées ISO19115 ou équivalent , CSW (ou OAI/PMH)

ex : NetCDF convention CF, CSV / ODV Spreadsheet

➔**interopérabilité, synchronisation avec services de diffusion**

ex : WMS pour prévisualisation,

OpenDAP/Errdap pour synchronisation avec Services de diffusion

Dépôt / publication

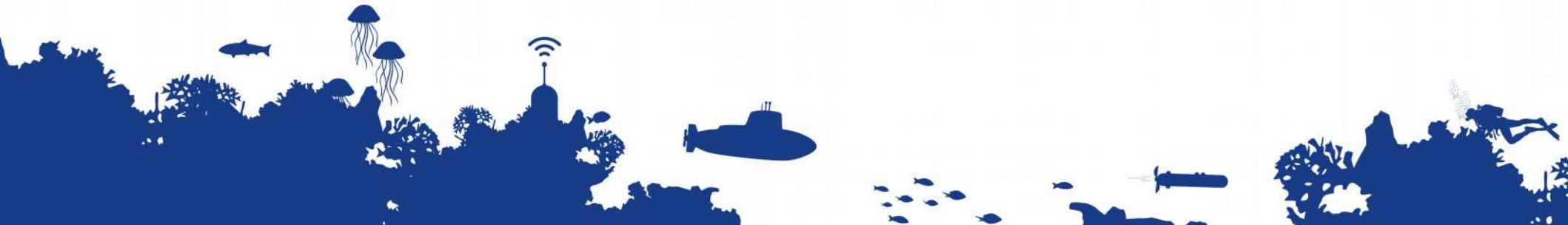
- **Objectif : fournir un service « d'accueil », de base, pour l'ensemble des données produites**
  - Dépôt en ligne d'un jeu de données (format non spécifique)
  - Obtention d'un DOI (association à une publication scientifique)
  - Préservation « en l'état »
  - Certification de type RDA - CoreTrustSeal Data Repository ou ISO 16 363
  - Référencé par les éditeurs scientifiques (1)

- **Conditions :**
  - Authentification des déposants
  - Filtrage des dépôts par CDS (qui sont compétents pour écarter les dépôts indésirables)

- **Valorisation des dépôts par les pôles**
  - Ingestion par les CDS et/ou dans les traitements
  - Utile pour les CDS (dépôts de données d'observation ou dérivées de référence, jeux de données « stabilisés »)

- **Exemples**
  - SeaNoe
  - Sedoo
  - EUDAT (CINES)
  - Zenodo (projets européens)
  - Logiciel DataVerse

[1] http://journals.plos.org/plosone/s/data-availability



Services de diffusion avancés

- **Objectif : fournir des services d'accès performants à l'utilisateur pour l'ensemble des données gérées**
  - Permettant de combiner, lors d'une même requête, des données de différentes origines (assemblage)
  - Permettant d'accéder aux seules données utiles (subsetting)
  - Permettant par exemple de visualiser conjointement des données gérés par plusieurs CDS
  - Proposant des « accès directs» aux données (non limités aux téléchargements) suivant des protocoles variés (Cloud, OGC, OpenDap, ...)

- **Conditions :**
  - « Pot commun » des données, éventuellement « non visible » de l'utilisateur (idée de « cache technique ») pour ne pas mettre les CDS au second plan (en lien avec politique des données)
  - Structuration des données orientée « utilisation » (et non plus observation)
  - Synchronisation des données avec CDS (interopérabilité) :
    avec rsync, OpenDap (synchronisation « nuancée »), ….?
  - ⇐ implique duplication des données (possible pour les données in-situ et les produits élaborés uniquement)
  - ⇐ data centre ayant capacité à héberger des services en ligne (disponibilité/redondance, sécurité, absorption des pics d'audience, connectivité réseau adéquate)

- **Valorisation**
  - Amélioration des temps de réponse et de la disponibilité des services pour les utilisateurs
  - Pour les CDS, possibilité de délégation de tâches informatiques ardues à maintenir 24h/24 et 7jours sur 7 (« hébergement délégué » des services en ligne)

- **Exemples**
  - DIAS (Exemple de Copernicus Marine Services : données centralisées et logiciel d'accès installés 1 seule fois)
  - SeaDataCloud (en partenariat avec EUDAT)