Cycle de vie des données

Sabine Schmidt (Directrice scientifique ODATIS, pôle Océan)
Sébastien Payan (Directeur scientifique AERIS, pôle Atmosphère).

Objectifs de l’exposé

Sensibiliser sur le cycle de vie et la gestion des données,

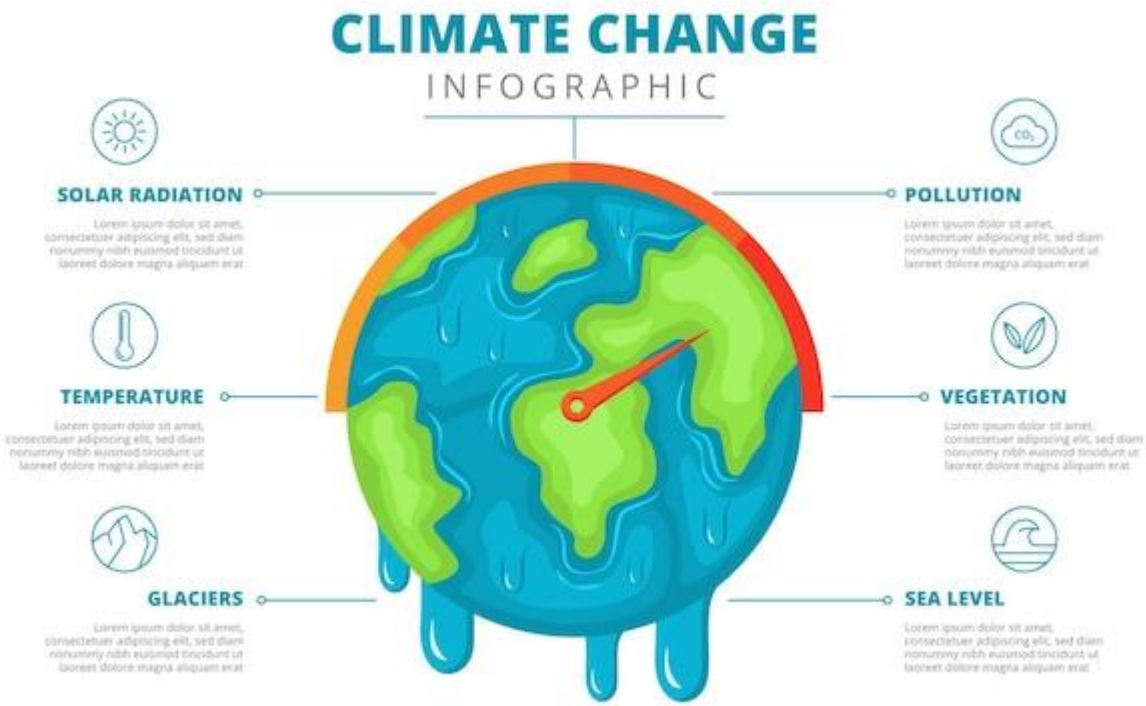
notamment sur le besoin de bien renseigner les données (méta-données), la place et liens avec les producteurs de données, le contrôle de l’usage des données sous forme de DOI ou de licences, et le vocabulaire.

L’anthropocène

Depuis la révolution industrielle, l'empreinte des activités humaines sur l'environnement mondial s'est accrue.

Les conséquences actuelles et attendues du changement global sur l’environnement sont multiples:

- réchauffement
- pollution
- niveau marin,
- fonctionnement des écosystèmes …



Un besoin crucial de mieux comprendre pour prévoir les impacts du changement global

Des observations sont nécessaires à tous les stades du processus scientifique :

description, compréhension, modélisation et prévision

Considérant que:

- l'acquisition des données est difficile et onéreuse :

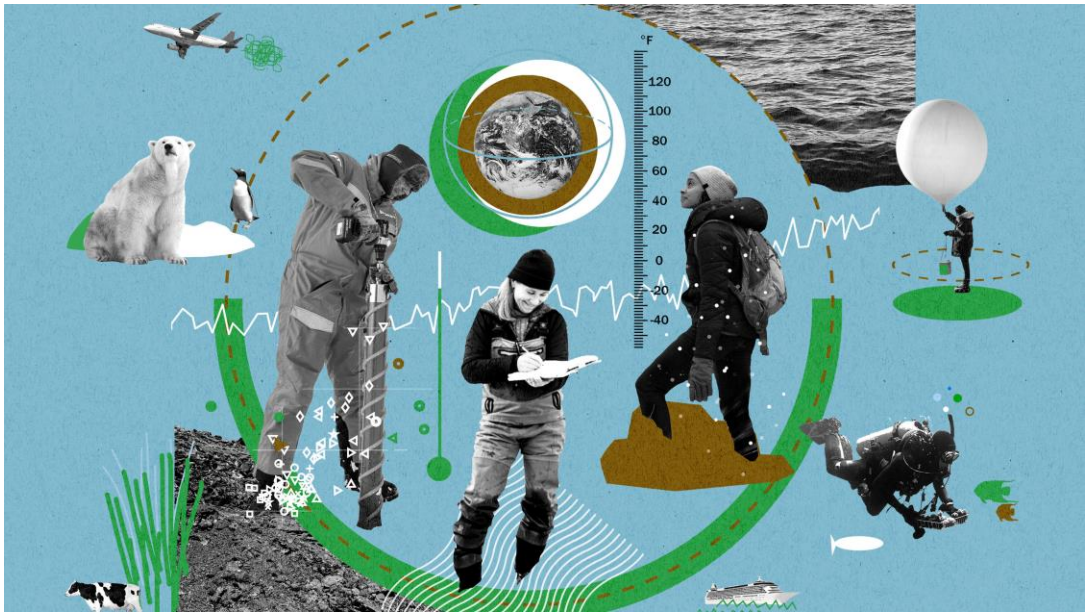
  Nécessité d'accéder à des sites distants et d'utiliser de nombreux moyens techniques (flotte aéroportée et océanographique, sites instrumentés, ballons, gliders, bouées Argo, …)

- sans archivage, > 30% des données sont perdues ou inutilisables 10 ans après leur acquisition (source: Ifremer).

➡ La préservation des données environnementales est un enjeu majeur,

Les défis de l’accès aux données du système Terre

Augmentation du nombre des observations (*in situ ;* télédétection; analyses au laboratoire) au cours des dernières décennies.



Pour tirer le meilleur parti de ce flux de données au profit de la connaissance et de la société, les centres de données doivent respecter des principes communs.

Le cadre réglementaire

La directive 2007/2/CE INSPIRE du 14 mars 2007:

Directive du Parlement européen visant

- à rendre disponible une information géographique, appropriée, harmonisée et de qualité, pour aider à l'élaboration, l'exécution, la surveillance et l'évaluation des décisions politiques environnementales européennes.
- Ce qui impose d'établir une infrastructure d'information géographique = ensemble de services d'information disponibles sur Internet, répartis sur les sites des différents acteurs concernés, et permettant la diffusion et le partage de données géographiques.
- concerne les autorités publiques (État, collectivités territoriales et leurs groupements, établissements publics, ainsi que toute personne physique ou morale fournissant des services publics en rapport avec l'environnement).
- impose aux autorités publiques, de rendre ces données accessibles au public en les publiant sur Internet, d'autre part de les partager entre elles.

Le cadre réglementaire

La Science Ouverte

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Le Plan national pour la science ouverte 2021-2024 : vers une généralisation de la science ouverte en France

vise à généraliser les pratiques de science ouverte, à partager et ouvrir les données de la recherche, et à promouvoir les codes sources produits par la recherche.

→ création de la plateforme nationale des données:

www.data.gouv.fr

données environnementales : IR Data Terra

Le cycle de vie de la donnée

Le schéma d'ODATIS

Le cyc

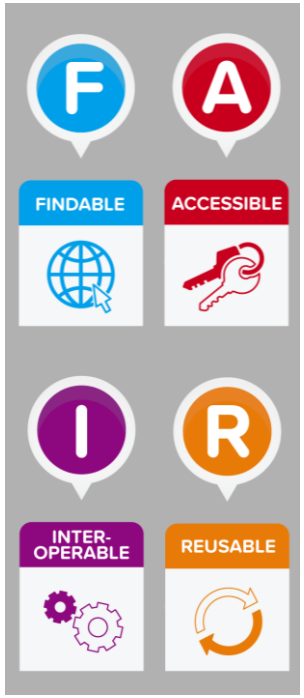

LE CYCLE DE VIE DES DONNEES
Un outil pour améliorer la gestion / la mise en qualité / l'ouverture des données
Produire des données Faciles à trouver, Accessibles, Interopérables et Réutilisables (FAIR)
Respecter le droit des personnes pour les Données à Caractère Personnel (DCP)
Ouvrir ses données autant que possible, les fermer autant que nécessaire
PAR QUOI COMMENCE-T-ON ?
PGD
PEUT-ON RÉUTILISER DES DONNÉES ?
CONCEVOIR LE PROJET
Réfléchir aux données produites ou réutilisées
Rédiger un Plan de Gestion de Données (PGD)
Anticiper la pertinence du recueil des DCP
CRÉER LES DONNÉES
Collecter/acquérir des données
Créer les 1ères métadonnées descriptives des données
Si DCP, déclarer le traitement et préparer les formulaires (information / consentement)
TRAITER LES DONNÉES
Appliquer des processus adaptés pour préparer les données brutes
Vérifier, valider, nettoyer les données (curation)
Compléter les métadonnées
Stocker et échanger les données de manière sécurisée sur la durée du projet
ANALYSER LES DONNÉES
Utiliser des méthodes et des outils pour produire des résultats liés à la problématique initiale
Tracer les différentes étapes des analyses
Si DCP et modification de l'analyse initialement envisagée, mettre à jour le dossier de conformité
CONSERVER/ ARCHIVER LES DONNÉES
Réfléchir aux données pertinentes à préserver : conservation sur le long terme et périmètre du partage
Structurer et organiser les jeux de données à préserver, accompagnés de leurs métadonnées
Si DCP, adapter la sécurité des données à leur confidentialité
SUPPRIMER LES DONNÉES
Dans certains cas, supprimer les données brutes (exemple : données personnelles non anonymisées) et ne conserver que les résultats des analyses et les métadonnées
DIFFUSER DONNÉES / MÉTADONNÉES
Selon le PGD, donner accès aux données communicables et aux métadonnées par le moyen le plus adapté (entrepôts, catalogue de données, article, ...)
Restreindre l'accès si nécessaire (exemple DCP, propriété intellectuelle, ...)
Attribuer des identifiants pérennes (DOI)
Se questionner sur la dimension juridique (licences)
RÉUTILISER LES DONNÉES
Utilisateur
Découvrir, exploiter les données, s'informer sur les licences et contrats associés, citer les données utilisées
Producteur de données
évaluer la réutilisation (citations, nombre téléchargements, ...)
PENSER FAIR
creative commons
ETALAB
Batifol V., Burnel L., Cardona A., Johany F. 2021. Affiche « Cycle de vie des données : un outil pour améliorer la gestion, la mise en qualité et l'ouverture des données ». Réseau Qualinous & Mission RGPD, département ACT – INRAE. DOI : 10.15454/hsc3-b796
Création graphique : www.ctookom.com
Qualinous
INRAE

Les défis de l’accès aux données

Défi 1 : la qualité des données (principes FAIR)

Les métadonnées et les données doivent être faciles à trouver et à (ré)utiliser (décrire vos données, appliquer des identificateurs persistants)

Les données doivent être intégrables à d'autres données (format ouvert, vocabulaire cohérent, normes de métadonnées)



Envisager ce qui sera partagé et comment on peut y accéder

Les données doivent être réutilisables, avec des métadonnées bien décrites et une licence appropriée

Les défis de l'accès aux données marines

Défi 2: certification des dépôts de données

La Research Data Alliance (RDA) fournit un cadre commun pour la mise en œuvre et la maintenance des dépôts numériques.

the CoreTrustSeal requirements

La certification est importante pour garantir: :

- ✓ la fiabilité et la durabilité des dépôt de données,
- ✓ l'archivage et le partage à long terme des données

pour les utilisateurs et les financeurs.

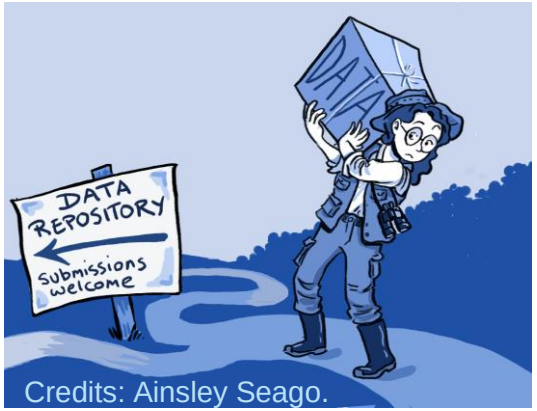
www.coretrustseal.org

Les défis de l’accès aux données marines

Défi 3: Amener les producteurs à partager leur données

L’ICSU (International Council for Science) promouvait déjà en 2011

« un accès complet et ouvert aux données scientifiques, en particulier lorsque la recherche est financée par des fonds publics ».



Credits: Ainsley Seago.

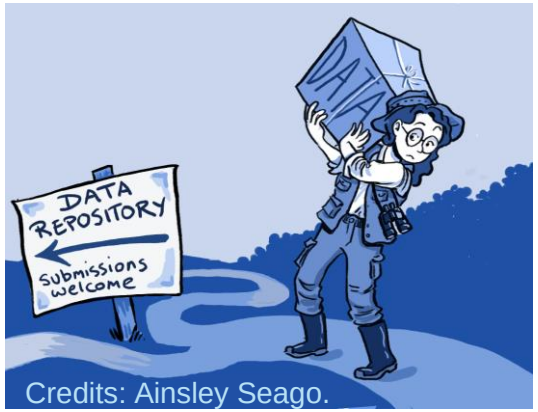
Les défis de l'accès aux données marines

Défi 3: Amener les producteurs à partager leur données

L'ICSU (International Council for Science) promouvait déjà en 2011
« un accès complet et ouvert aux données scientifiques, en particulier lorsque la recherche est financée par des fonds publics ».

Les collègues peuvent être réticents à partager leurs données **en raison des coûts réels et/ou perçus**

Sentiment de - perte de contrôle sur les données,
- contraintes sans bénéfice,
Manque de - ressources informatiques/humaines
- formation.



Credits: Ainsley Seago.

Les dépôts de données doivent rendre le partage de données

plus faciles pour le producteur
pas seulement pour l'utilisateur.

Le besoin d’infrastructures interopérables

Afin d'accélérer la collecte et l'utilisation des données, il est nécessaire de disposer d'infrastructures interopérables pour aider :

- les producteurs à archiver et à partager leurs données,
- les utilisateurs à obtenir un accès relativement facile aux données

coordonnées au moins au niveau national, voire international.

Initiative nationale (FR): L'IR Data Terra

Infrastructure de recherche « *pôle de données et services pour le Système Terre* » : *de la donnée à la connaissance du Système Terre*

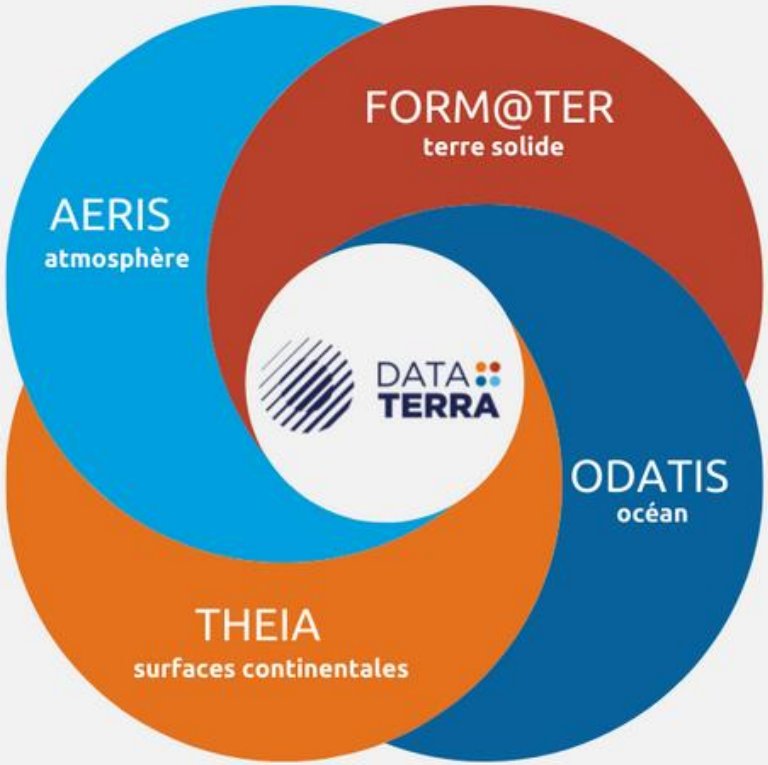
L'IR DATA TERRA
ses pôles de données & dispositifs transversaux

Cette IR est fondée sur quatre pôles de données correspondant à chacun des grands compartiments du système Terre :

- **THEIA pour les données surfaces continentales** *(agriculture, forêts, biodiversité …)*
- **AERIS pour les données atmosphère** *(gaz, aérosols, nuages …)*
- **ODATIS pour les données océan** *(niveau moyen des mers, hydrologie, risques littoraux…)*
- **ForM@Ter pour les données terre solide** *(volcanologie, érosion des sols, sismologie …)*

Elle regroupe aussi deux dispositifs transversaux :

- **DINAMIS** (Pour accéder aux données spatiales haute résolution)
- **INTER-PÔLES** (Fédérer et animer des communautés d'experts en données)



Dispositifs transversaux DATATERRA DINAMIS DATATERRA INTER-PÔLES

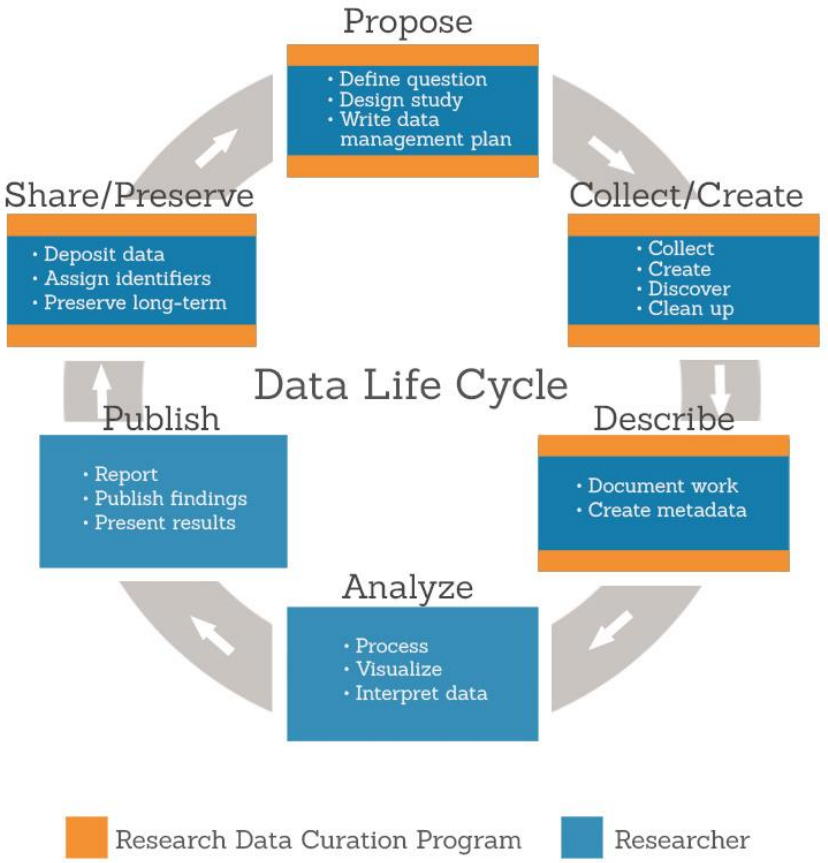
Le DMP:

Avant toute manipe,

voire dépose de projet de recherche:

→ besoin d'établir un DPM pour établir
comment les données seront gérées
pendant le projet
mais surtout à la fin !

(metadata, bancarisation, préservation pour le futur).

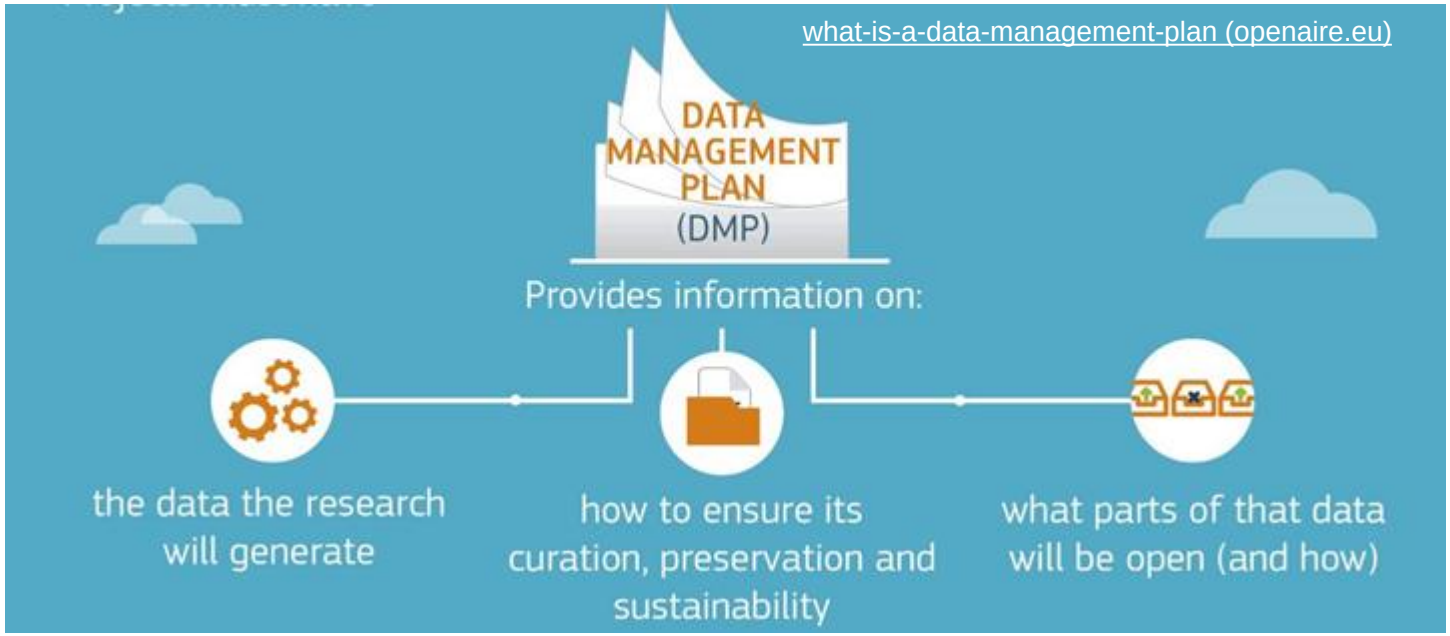


Le DMP:

Multiplication de l'instrumentation in situ (capteurs innovants, drones, gliders, …) :

➔production de données très diverses (stations fixes ou lagrangiennes, fréquence d'acquisition, paramètres mesurés).

pour garantir l'acquisition de bases de données bien documentées en vue d'exploitation dans le futur, par des politiques d'archivage, d'interopérabilité et de licences adaptées.



La donnée

Pourquoi définir une donnée ?

Comment définir une donnée ?

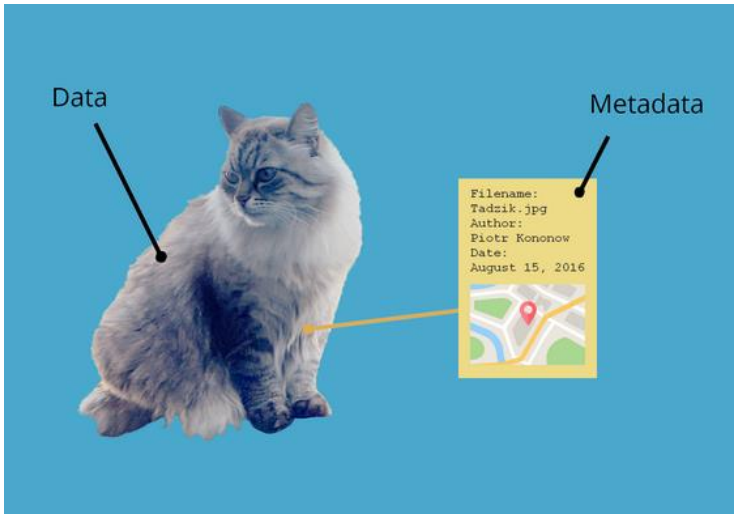
Vocabulaire:

- différentes disciplines différents sens,
- les unités
- les capteurs
- les méta-données

La donnée

Les méta-données

Données sur les données

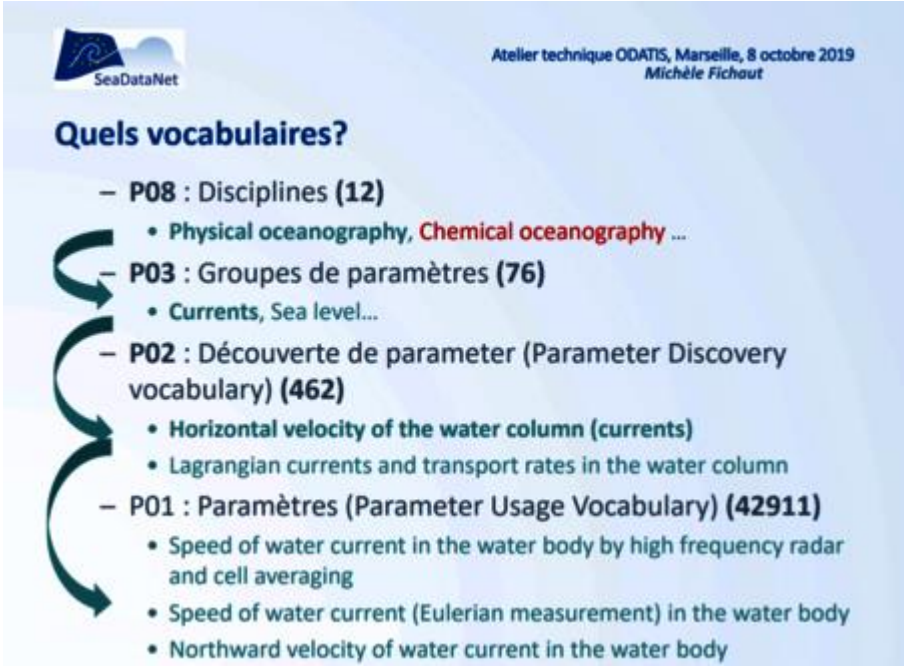


Le vocabulaire

Un paramètre est défini par:
 un vocabulaire contrôlé et un modèle sémantique pour faciliter l'échange de donnée et l'interopérabilité

Exemple: serveur de vocabulaire British Oceanographic Data Centre-BODC). adoptés par plusieurs projets (nationaux et européens)

Pour plus de détails:
https://www.odatis-ocean.fr/donnees-et-services/principes-de-gestion-des-donnees/referentiels-et-vocabulaires



Favoriser le partage des données

L'archivage

Quels sont les freins à la mise en base de données ?

complexité
méconnaissance des bases, format
le temps
enjeux de propriétés intellectuelle
pb de la transformation de la donnée (licences)

Favoriser le partage des données

Ce qui pourrait inciter les producteurs à mettre en base ?

- lien avec le producteur de données
- assurer la protection (propriété intellectuelle)
- retour sur l'usage des données

Favoriser le partage des données

Quels formats de fichier ?

Extrait de Formats, attributs, conventions (odatis-ocean.fr)

Convention paramètres et données du NetCDF-CF [Fortement recommandée]

Bien que le NetCDF version 4 apporte 5 nouveaux types de donnée utilisateur (« UserDefinedType » : « Enum », « Opaque », « Compound », « VariableLength »), il est fortement déconseillé de les utiliser.
Le NetCDF-4 apporte un niveau de **hiérarchisation** supplémentaire avec la notion de groupe dans son modèle (afin être conforme avec le HDF-5), il est fortement déconseillé de l'utiliser. En effet, l'utilisation de plusieurs groupes dans un fichier NetCDF complexifie grandement l'utilisation de ce fichier.
En ce qui concerne les paramètres **date** et **heure** dans les fichiers NetCDF, il est fortement recommandé de les insérer sous forme d'entier (type « long ») avec un offset (optionnel) et un scale-factor (obligatoire). L'échelle de temps à adopter est obligatoirement l'**UTC** (Universel Temps Coordonné).
Lorsque des **axes** sont nécessaires dans un fichier NetCDF, il est fortement conseillé de bien définir les axes et l'orientation. La profondeur est souvent insérée en positif dans les fichiers NetCDF (ex : 2000m), ce qui pose des problèmes lorsque l'on va utiliser ces fichiers avec des fichiers ayant des paramètres atmosphériques qui eux vont être aussi en altitude positive (2000m).

Présentation de pôles thématiques

ODATIS et AERIS

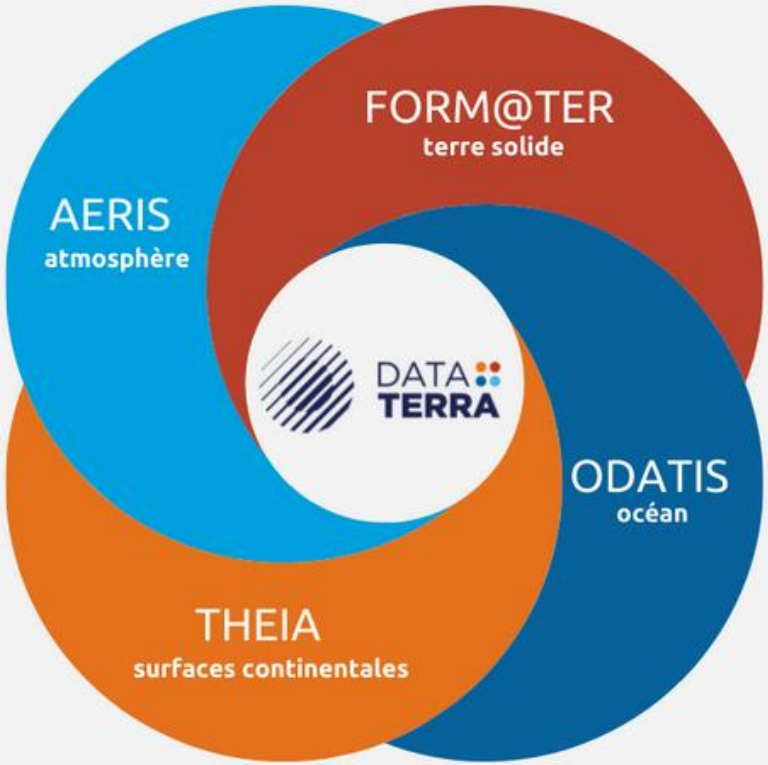
L'IR DATA TERRA
ses pôles de données & dispositifs transversaux

Cette IR est fondée sur quatre pôles de données correspondant à chacun des grands compartiments du système Terre :

- **THEIA pour les données surfaces continentales** *(agriculture, forêts, biodiversité …)*
- **AERIS pour les données atmosphère** *(gaz, aérosols, nuages …)*
- **ODATIS pour les données océan** *(niveau moyen des mers, hydrologie, risques littoraux…)*
- **ForM@Ter pour les données terre solide** *(volcanologie, érosion des sols, sismologie …)*

Elle regroupe aussi deux dispositifs transversaux :

- **DINAMIS** (Pour accéder aux données spatiales haute résolution)
- **INTER-PÔLES** (Fédérer et animer des communautés d'experts en données)



Dispositifs transversaux DATATERRA DINAMIS DATATERRA INTER-PÔLES