Institut Pythéas
Observatoire des Sciences de l'Univers
Aix-Marseille Université

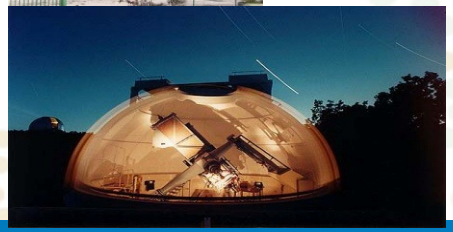
ERDDAP, un outil pour la Science Ouverte

*pour des données **F**aciles à trouver, **A**ccessibles, **I**nteropérables et **R**éutilisables*

M. LIBES & D. MALLARINO : OSU Pytheas CNRS

Findable Accessible Interoperable Reusable

jres MARSEILLE 2021 2022



cnrs

Institut Pythéas
Observatoire des Sciences de l'Univers
Aix-Marseille Université

ERDDAP, un outil pour la Science Ouverte

1) Le contexte institutionnel et l'*Open Science*
2) Les principes FAIR
3) Cycle de vie des données
4) … et ERDDAP là-dedans ?
   - Objectifs, bénéfices
   - Le protocole DAP et les requêtes

- ***Tuto ERDDAP « Howto » : travaux pratiques***

Le cadre institutionnel de l'Open Science

- Plan national pour la Science ouverte (07/2018, F. Vidal)
  - « les résultats de la recherche scientifique **ouverts à tous, sans entrave, sans délai, sans paiement** ***« ouvert autant que possible, fermé autant que nécessaire… »***
  - « … rend obligatoire l'accès ouvert pour les publications et pour les données issues de recherches sur projets »
  - « faire sortir la recherche financée sur **fonds publics** du cadre confiné des bases de données fermées. Réduire les efforts dupliqués dans la collecte, la création, le transfert et la réutilisation du matériel scientifique. **Améliorer ainsi la reproductibilité, l'efficacité, l'éthique et la transparence de la recherche** »

- ***https://www.ouvrirlascience.fr***
  - **« … science sera plus cumulative, plus fortement étayée par des données, plus transparente, plus rapide et d'accès universel »**
  - **« … accès ouvert aux publications et – autant que possible – aux données, aux codes sources et aux méthodes de la recherche »**
  - **« … La communauté scientifique doit œuvrer à la construction d'un écosystème de la publication scientifique ouvert, éthique et transparent »**

Le cadre institutionnel de l'Open Science

- Feuille de route CNRS pour la Science ouverte (nov. 2019)
  - « La Science Ouverte ne favorise pas seulement une approche transversale du partage des résultats de la science … »
  - « En ouvrant les **données**, les processus, les codes, les méthodes ou encore les protocoles, elle offre aussi une **nouvelle façon de faire de la science** »

- Le Plan données de la Recherche du CNRS (nov. 2020)
  - **« pour inciter les scientifiques à rendre leurs données accessibles et réutilisables »**
  - « le cadre légal des données de la recherche produites dans un contexte de financement sur **fonds publics** a évolué vers un principe de mise à disposition en **données ouvertes** …»
  - **« … incite toutes les structures de recherche à se doter de politiques des données »**

Les principes « FAIR »

***F**indable, **A**ccessible, **I**nteroperable, **R**eusable*

- **F :** Les données doivent être ***findable*** (faciles à trouver) et identifiables par les humains et les machines :
  - → **catalogues**, **métadonnées**, mots clés issus de **thésaurus** disciplinaires,
  - → identifiants uniques et pérennes (DOI)

- **A :** Les données doivent être **accessibles** facilement, avec des conditions connues :
  - → des licences claires, des protocoles « ouverts »,
  - → données présentes dans des entrepôt certifiés … et accessibles

- **I :** Les données doivent êtres **interopérables** à plusieurs niveaux :
  - **Sémantique** : **vocabulaires contrôlés, métadonnées disciplinaires** précises
  - **Syntaxique** : **protocoles d'échanges** ouverts et standards *(CSW, WMS, SOS, DAP …)*
  - **Contenus** : **formats** de fichiers aux **standards** internationaux disciplinaires (ex : NetCDF, ODV, etc.)

- **R : Réutilisables** : l'objectif final des principes FAIR : la pérennité et réutilisation des données :
  - Pas de « R » sans « FAI »
    - Identifiants uniques et pérennes (DOI) pour l'identification et la citation des données
    - Licences claires d'utilisation des données
    - Standards communs : protocoles d'échanges et formats standards des données qui répondent à des normes communautaires pertinentes pour le domaine
  - Authentification d'accès, si nécessaire

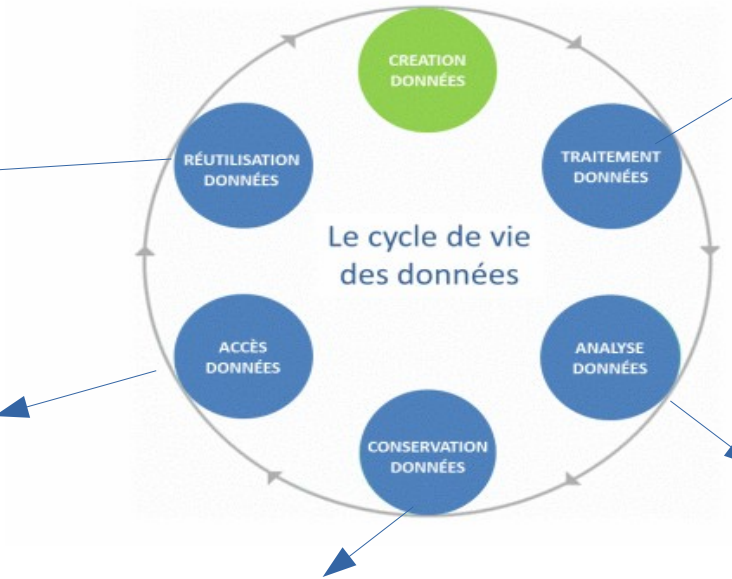
Le « cycle de vie des données »

Le « ***cycle de vie*** » est une représentation structurante qui décrit la vie des données en *plusieurs phases*

Instrumentation, capteurs, mesures, modélisation

Conversion d'un format brut ou spécifique vers un format interopérable ou standard, qualification, nettoyage

Croisement des données, création de nouvelles données



Accès et diffusion et publication des données

Outils logiciels de traitements & calculs, programmation ad-hoc

Format des données & archivage pérenne

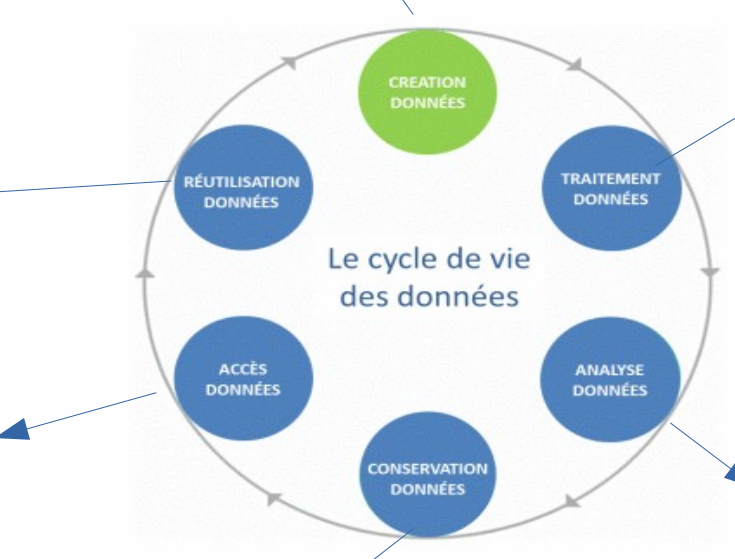
ERDDAP dans le « cycle de vie des données »

Sur chaque phase, on devra être attentif aux principes FAIR et leur application. ERDDAP de son coté apporte de l'aide ou une solution à certaines étapes.

Instrumentation, capteurs, mesures, modélisation

Conversion d'un format brut ou spécifique vers un format interopérable ou standard, qualification, nettoyage

Croisement des données, création de nouvelles données : ERDDAP apporte des solutions



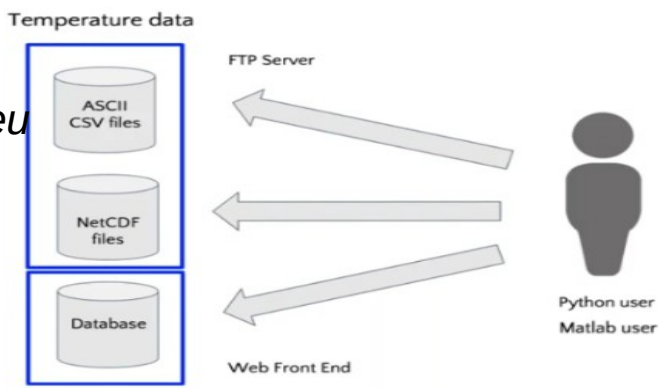
ERDDAP est dans son rôle : ***making it easier for you to get scientific data***

Outils logiciels de traitements & calculs, programmation ad-hoc

Format des données & archivage pérenne : ERDDAP facilite la centralisation

Quelques problématiques de l'*open data*

- ***Pour le fournisseur :***
  - ***Fournir*** *les données dans plusieurs formats standards interopérables différents* ***sans*** *dupliquer les données*
  - *Agréger les fichiers en un seul jeu de données homogène*
  - *Assurer le catalogage et la découverte des données*
  - *Associer des métadonnées aux données*
  - *Homogénéiser les formats de temps différents* sur les jeux de données



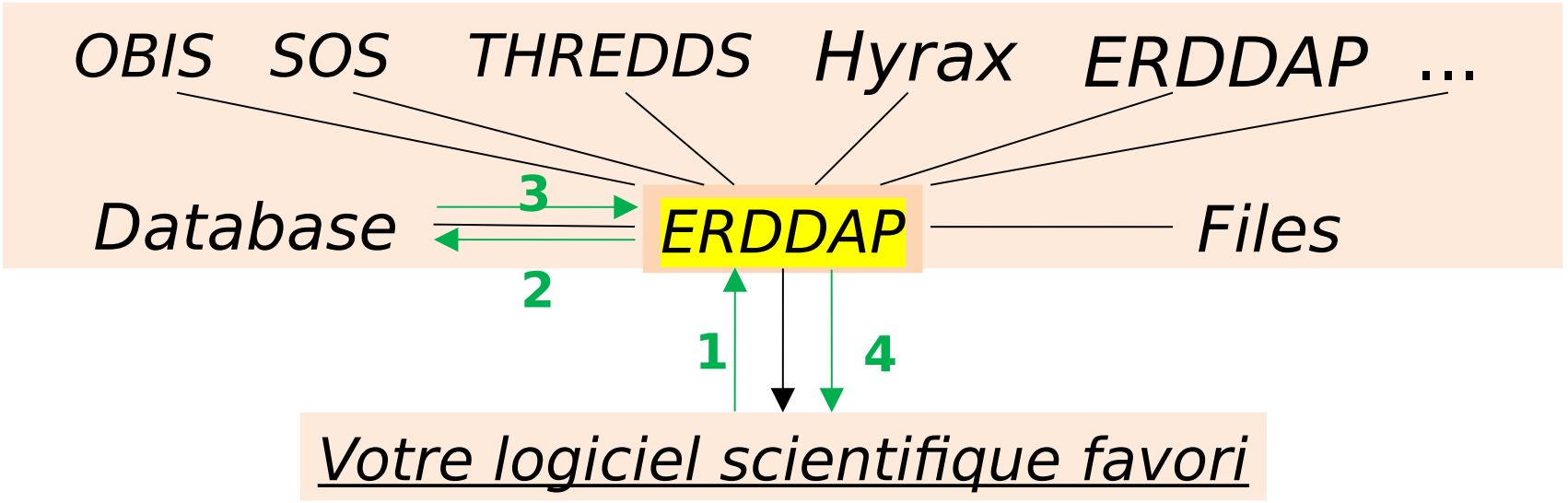
- ***Pour l'utilisateur***
  - *Les données sont dans des formats très variés*
  - *Les jeux de données (dataset) peuvent se composer de centaines de fichiers*
  - *Les utilisateurs aimeraient avoir les données dans leur application (format) favorite*

Auxquelles ERDDAP fournit des solutions...

- **ERDDAP** est une plateforme logicielle interopérable pour l'accès et la distribution de données scientifiques (**auteur** : Bob Simons de la NOAA)
  - Le leitmotiv est ***Easier Access to Scientific Data***
    → but : faciliter l'affichage, l'accès et la diffusion de données scientifiques
  - Agit comme un *middleware* entre les données et des clients, humains ou logiciels
    → s'appuie sur de nombreux formats et protocoles standards **en entrée et sortie**

- Fournit un ensemble complet de fonctionnalités pour la gestion des jeux de données :
  - Lire et convertir des jeux de données dans de nombreux formats
  - Fournir un catalogue et une recherche avancée des jeux de données
  - Afficher et fournir les métadonnées
  - Interroger et filtrer les données au travers de formulaires
  - Créer des graphiques, des cartes simples pour visualiser les jeux de données
  - *Agréger automatiquement les nouvelles données*
  - *Créer un réseaux de données distribuées*
  - *Homogénéiser les dates au format ISO, etc...*

ERDDAP vu depuis un fournisseur de données



Agit comme un *middleware (data broker, data provider)* entre les données et des clients (humains ou programmes)… s'appuie sur de nombreux formats et protocoles standards **en entrée et sortie**

ERDDAP cause « *DAP* » *(Data Access Protocol)*

ERDDAP utilise **le protocole *openDAP*** pour accéder aux données *au travers de requêtes standardisées HTTP REST*

- *https://earthdata.nasa.gov/esdis/eso/standards-and-references/data-access-protocol-2*
- *https://earthdata.nasa.gov/collaborate/open-data-services-and-software/api/opendap*

- openDAP : protocole ouvert, client-serveur, d'accès aux données, soutenu par la NASA, largement *utilisé dans la diffusion des données scientifiques*
  - permet l'extraction sélective de données à distance sous la forme d'un service Web facile à invoquer qui repose sur HTTP
  - Requête via ligne de commande, un navigateur Internet ou une interface utilisateur personnalisée
  - Fonctionne avec divers outils : Matlab, R, jupyter-notebook, IDL, Panoply… etc.
  - OpenDAP extrait les données de divers formats standards :
    - *NetCDF, GeoTIFF, JPEG2000, JSON, ASCII, HDF*

The Data Access Protocol - DAP

Un utilisateur (ou programme) peut construire des expressions de contraintes élaborées qui renverront précisément les données

- **https://erddap.osupytheas.fr/erddap/tabledap/Emso_Ligure_Ouest_Albatross_Aquadopp_NetCDF_2021.graph?time,Speed&depth=500.0&time>=2021-09-26T00:00:00Z&time<2021-09-28T00:00:00Z&.draw=linesAndMarkers&.marker=63&.color=0x000000&.colorBar=%7C%7C%7C%7C%7C&.bgColor=0xffccccff)**

- Le serveur : **https://erddap.osupytheas.fr/erddap/tabledap**
- Le dataset : **Emso_Ligure_Ouest_Albatross_Aquadopp_NetCDF_2021**
- Les données à extraire : **time,Speed**
- Les filtres : **depth=500.0&time>=2021-09-26T00:00:00Z&time<2021-09-28T00:00:00Z**

En savoir plus

- Réseau ***SIST*** de l'INSU :

  ➔ *http://sist.cnrs.fr*

- *« Guide de bonnes pratiques sur la gestion des données de la recherche*

  ➔ *https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html*

- ***EcoInfo*** : AGIR sur les données de la Recherche :

  ➔ *https://ecoinfo.cnrs.fr/2021/03/01/agir-sur-les-donnees/*

Time to play : le Tuto… Erddap

- La doc en ligne du tuto ERDDAP : ~~*https://maurice.libes.pages.in2p3.fr/tuto-erddap-jres/*~~
  - ***Nouveau lien corrigé =>*** ***https://mauricelibes.pages.in2p3.fr/tuto-erddap-jres/***
- La machine virtuelle VirtualBox du TP préinstallée (Debian11, Tomcat, Java et ERDDAP - 7Go)
  - ici : **https://nuage.osupytheas.fr/s/Brm3Tw6Bja5ZkJK**
  - ou là : **https://amubox.univ-amu.fr/s/9XMP8iTeJGWZ86a**

Plan du tutorial

- *Installation, configuration*
- *Utilisation*
  - *Chargement de divers jeux de données en CSV, NetCDF*
  - *Agrégation automatique des fichiers*
  - *Visualisation graphique*
  - *Interrogation par formulaires requêtes*
  - *Exportation, reformatage à la volée*
  - *Accès aux données via un jupyter notebook*