20 ans de science ouverte au laboratoire LEGI :

Quels succès ? Quelles difficultés ? Quels enseignements ?

Un tremplin vers le futur

Cyrille Bonamy (IR-CNRS), Gabriel Moreau (IR-CNRS),
Julien Chauchat (MCF-GINP), Joël Sommeria (DR-CNRS)

Laboratory LEGI - CNRS / UGA / Grenoble-INP - France

18 mai 2022 / Marseille

La Science Ouverte

Données

- Principes **FAIR**
- Plan de Gestion des Données : une obligation aujourd'hui
- Problématique de la durée de vie

Logiciels

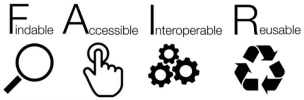
- **Libre**, OpenSource
- Nécessité d'associer logiciels et données
- FAIR4RS (RS = Research Software)
- Reproductibilité (computationnelle ?)

Publications

- Green **OpenAccess** : pre-print et articles soumis
- Gold OpenAccess : facile mais onéreux ; Attention aux éditeurs qui en profitent !

- Objectifs : transparence, reproductibilité, efficacité, confiance, sobriété, pérennité
- Pour qui ? **vous**, votre laboratoire, votre communauté, tout le monde
- Obstacles : le coût humain à court terme

Les principes FAIR



Findable

Les **métadonnées**, données et logiciels doivent pouvoir être trouvées tant par les humains que par les ordinateurs ; identifiant persistant et unique : Digital Object Identifier (doi)

Accessible

Une fois trouvées, les utilisateurs doivent savoir comment y accéder (dépôts, protocoles. . .)

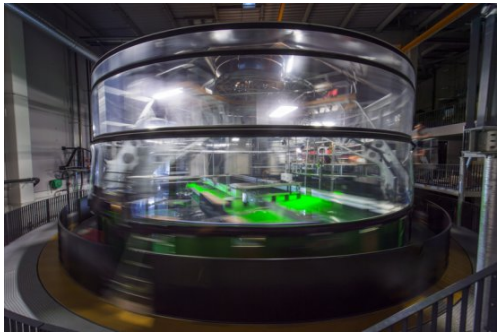
Interoperable

Les données doivent être publiées sous un **format standard** facilitant les échanges.

Reusable

L'accès est défini au travers d'une **licence** qui précise les conditions d'usage.

Ce qui s'est fait au LEGI ?



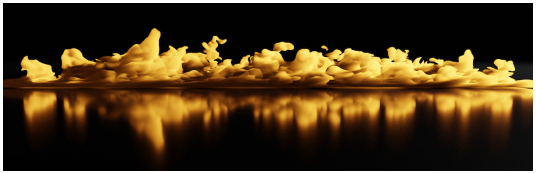
Il y a 20 ans...

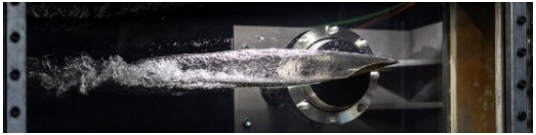
- Objectif : **échange** de données expérimentales (champs de vitesse, concentration...)
- Coriolis : grand équipement accueillant des équipes du monde entier
- Exemple de données : des fichiers XML, PNG et NetCDF(quelques Go par expérience en 2000, des dizaines de To aujourd'hui)
- **Associé** à quelques logiciels libres (UVmat, puis fluidimage)
- Loin d'être parfait : organisation, reproductibilité, formats

Ce qui se fait actuellement au LEGI ?

Aujourd'hui

- Application de ces principes bien plus large que CORIOLIS
- Idéalement : association données et logiciels avec publications
- Quelques réticences, mais des débats
- Au mieux en fonction de nos moyens





Une vraie dynamique

- Un groupe de travail regroupant chercheurs et ingénieurs
- Une méthodologie rigoureuse
- Des financements sur projet (WP du projet européen Hydralab)
- Une diffusion de notre savoir-faire et expérience : ANRs, Univ du Delaware

Focus concernant les données

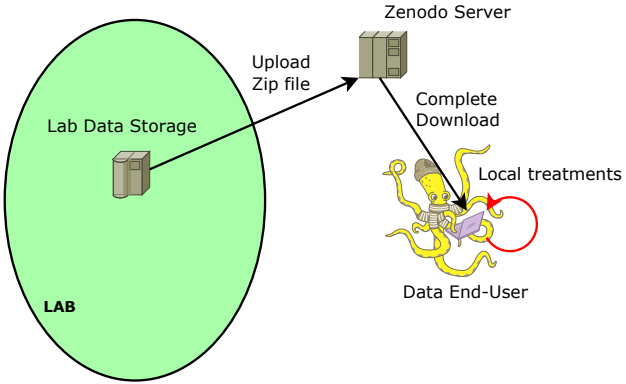
Plusieurs solutions pour le partage et la diffusion des données ouvertes

- Simplement sur Zenodo
- Via un accès distant OPeNDAP
- Via le partage de scripts de post-traitement à distance : (Jupyter Notebook + OPeNDAP)

Quelques points d'intérêt

- Importance d'associer des métadonnées
- Nécessité d'apposer une licence et quelques fichiers indispensables (README, AUTHOR...)
- Choix de la solution en fonction des besoins et contraintes (Volume, Durée de vie)

Niveau 1 : Zenodo (CERN)



Pour

- Simple à utiliser
- ID unique par jeux de données (`doi`)

Contre

- Limité à 50 Go
- Téléchargement intégral de l'archive

Outils libres développés

- project-meta

Niveau 1 : Zenodo - Création d'un jeux de données libres

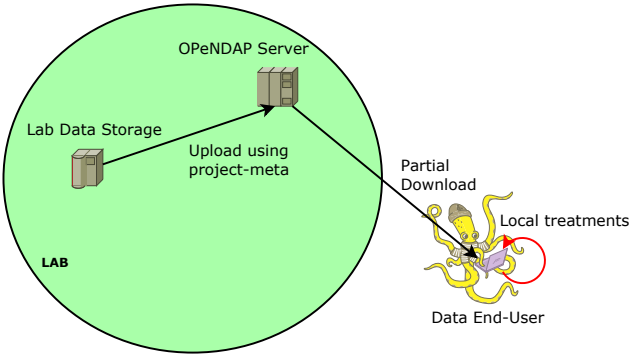
project-meta : boîte à outils libre en ligne de commande pour l'open-data

- Déclaration des métadonnées dans un fichier au format YAML
- Basée sur les spécifications **Dublin Core Metadata Initiative** (DCMI)
- Format descriptif simple composé de quinze propriétés relatives :
  - au contenu (titre, sujet, description, source, langue, relation, couverture)
  - à la propriété intellectuelle (créateur, contributeur, éditeur, gestion des droits)
  - à l'instanciation (date, type, format, identifiant de la ressource)
- Extension au Dublin Core pour déclarer les fichiers à mettre en données ouvertes
- Génération automatique à la racine de l'archive des fichiers : `LICENSE.txt`, `README.txt`, `AUTHORS.txt`, `COPYRIGHT.txt`...

Ligne de commande :

```
project-meta make-zip
```

Niveau 2 : OPeNDAP - Data Access Protocol



Pour

- Pas de limite en taille
- Téléchargement partiel possible (ex. sous-ensemble d'un fichier NetCDF)
- Encourage l'utilisation de **formats ouverts et auto-documentés** (NetCDF, HDF5. . . OPeNDAP backend)

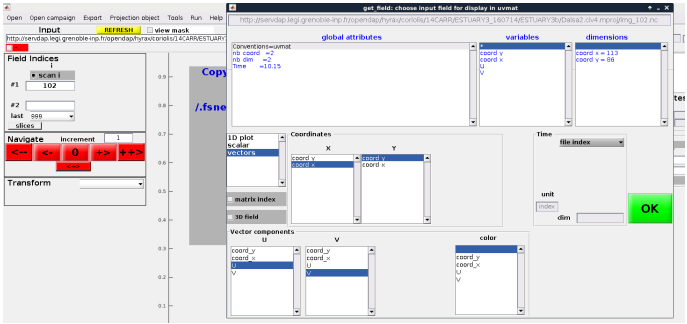
Contre

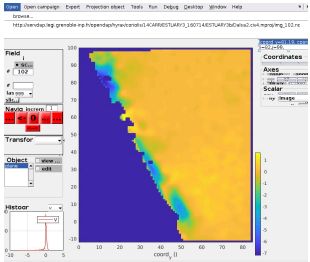
- Pas d'ID unique (`doi`)
- Obligation d'avoir son propre serveur OPeNDAP (pas de serveur publique)

Outils libres développés

- project-meta
- UVmat : compatibilité OPeNDAP

Niveau 2 : OPeNDAP - Visualisation distante (PIV) avec UVmat

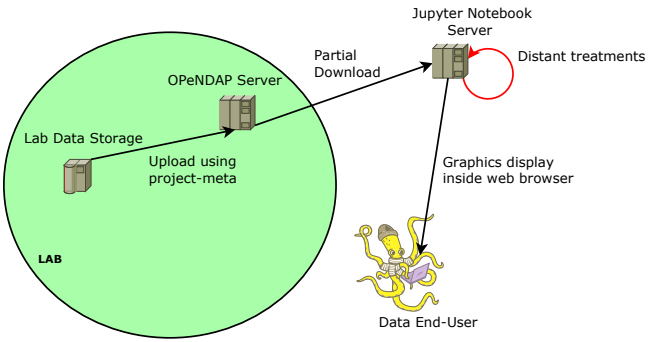




UVmat : boîte à outils graphique libre Matlab pour la PIV

- Développeur principal **Joël Sommeria** (CNRS LEGI)
- **100% compatible OPeNDAP**
- Traitement local sur un petit sous ensemble de données distantes PIV sans téléchargement complet de l'archive !

Niveau 3 : Jupyter Notebook + OPeNDAP



Pour

- Pas de limite en taille, téléchargement partiel...
- Pas besoin d'installer de logiciel sur le terminal utilisateur
- Utilise un langage ouvert pour le post-traitement (Python)
- Partage des scripts de post-traitement

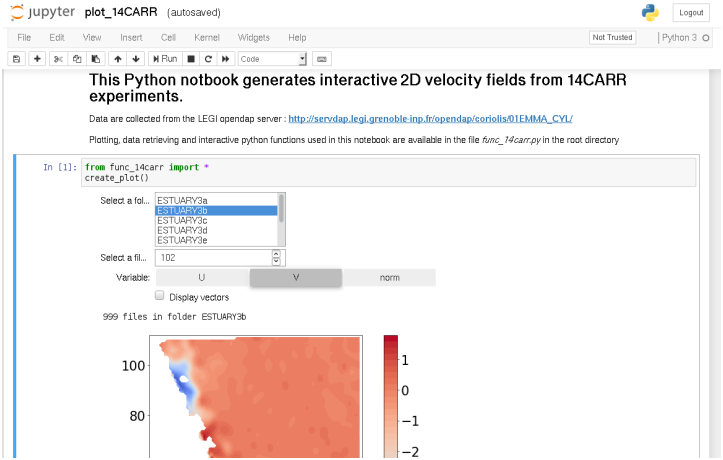
Contre

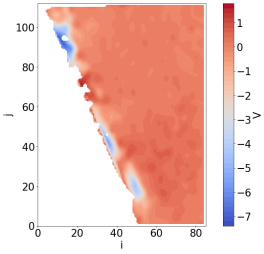
- Pas d'ID unique (`doi`)
- Obligation d'avoir son propre serveur OPeNDAP (pas de serveur publique)

Outils libres développés

- De nombreux notebooks Python

Niveau 3 : Jupyter Notebook + OPeNDAP - Traitements distants

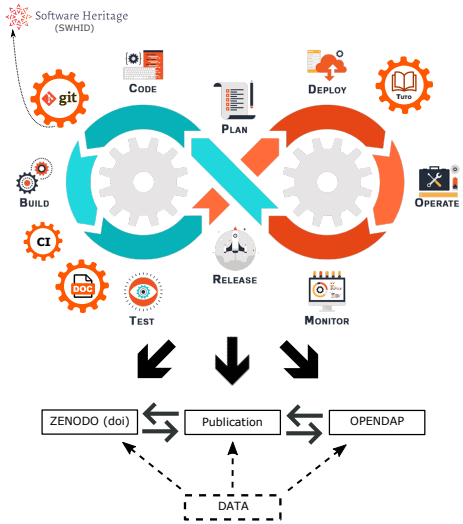




- Tous les post-traitements sont effectués depuis un navigateur Web
- Pas de dépendance au terminal utilisateur

https://mybinder.org/v2/gh/CyrilleBonamy/notebook_opendata/HEAD?labpath=plot_14CARR.ipynb

En matière de logiciels



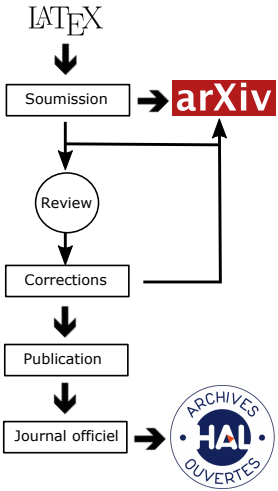
FAIR4RS autant que possible

- Gestion de version, intégration continue, documentation, licence
- Diffuser des version pérennes et identifiables via Zenodo (identifiant unique : doi) et Software Heritage (SWHID)
- Formats (standards) des fichiers d'entrée et sortie
- Reproductibilité (tests)
- Impacts environnementaux (CI, volume)
- Ne pas attendre ! Dès le début !

En matière de publications

Publications : *Ouverture*

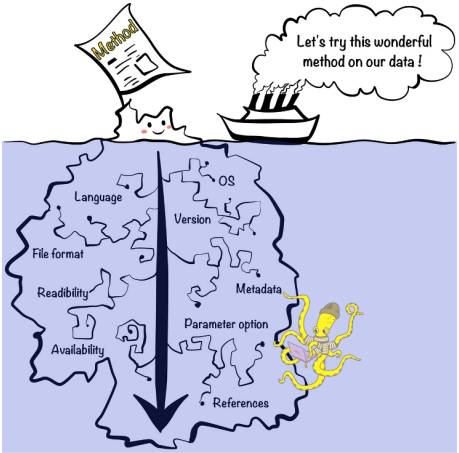
- Partager les « données » (*a minima* les configurations numériques) et logiciels associés à la publication
- Fournir les scripts libres de post-traitement pour recréer les images
- Documenter !
- Ne pas hésiter à ouvrir la publication dès que possible : Green OpenAccess
- Utiliser les outils de nos tutelles (Hal, forge...)



Difficultés

- Certaines craintes : critiques, mauvaise utilisation ou à des fins négatives, nécessité de fournir du support...
- La phrase clamée par le CNRS :

  « *accessible autant que possible, et fermé autant que nécessaire* »
- Coût : humain (service aux utilisateurs, documentation), publications
- Licences d'origine parfois bloquantes
- Démarche trop peu valorisée



Épilogue

Points saillants

- 1er bénéficiaire → initiateur
- Dès le montage du projet (DMP, SMP)
- Licences au plus tôt
- Apposer une durée de vie aux données
- Citer des données via un doi n'est pas naturel
- Coût humain à court terme, mais bénéfique à moyen terme
- Science ouverte et éco-responsabilité non opposées

Quelques liens intéressants

- Initiative européenne RDA
- Science Ouverte au CNRS
- Plan de données ouvertes du CNRS
- Cellule data-stewardship Grenobloise
- *Open Acess Funding* à Manchester
- Project-meta

Licence

Merci de votre attention !

Cette présentation est sous : Licence Art Libre

http://artlibre.org/

Code source