ÉCOLE CENTRALE DE LYON

Formation doctorale : Bibliothèque Michel Serres

ÉCOLE CENTRALELYON

Gestion des données de recherche : optimisation de l'impact de vos travaux(2/2).

ÉCOLE
CENTRALELYON

Code & données de la recherche (2/2)

Formateurs : Code & données de la recherche

- Linda Angulo, Chargée de Mission Données de la Recherche
- Matteo Camier, Responsable HPC @pmcs2i.ec-lyon

Formateurs : Publications

- Nicolas Jardin, Directeur adjoint Biblio. Michel Serres
- Stéphanie Lamaison, Bibliothécaire Biblio. Michel Serres

36, avenue Guy de Collongue 69130 Écully - France
+33 (0)4 72 18 60 00

Commentaires et rappels

```
netcdf filename {
dimensions:
        lat = 3 ;
        lon = 4 ;
        time = UNLIMITED ; // (2 currently)

variables:
        float lat(lat) ;
                lat:long_name = "Latitude" ;
                lat:units = "degrees_north" ;
        float lon(lon) ;
                lon:long_name = "Longitude" ;
                lon:units = "degrees_east" ;
        int time(time) ;
                time:long_name = "Time" ;
                time:units = "days since 1895-01-01" ;
                time:calendar = "gregorian" ;
        float rainfall(time, lat, lon) ;
                rainfall:long_name = "Precipitation" ;
                rainfall:units = "mm yr-1" ;
                rainfall:missing_value = -9999.f ;

// global attributes:
                :title = "Historical Climate Scenarios" ;
                :Conventions = "CF-1.0" ;

data:
 lat = 48.75, 48.25, 47.75;
 lon = -124.25, -123.75, -123.25, -122.75;
 time = 364, 730;
 rainfall =
   761, 1265, 2184, 1812, 1405, 688, 366, 269, 328, 455, 524, 877,
   1019, 714, 865, 697, 927, 926, 1452, 626, 275, 221, 196, 223;
}
```

Coordinate variable

Variable attribute

Global attribute

Guide de Bonnes Pratiques sur la gestion des données de la Recherche

Groupe de travail "Atelier Données"

Mission pour les Initiatives Transverses Interdisciplinaires

Voir: http://corist-shs.cnrs.fr/sites/default/files/ressources/guide_bonnes_pratiques_gestion_donnees_recherche_v2.pdf

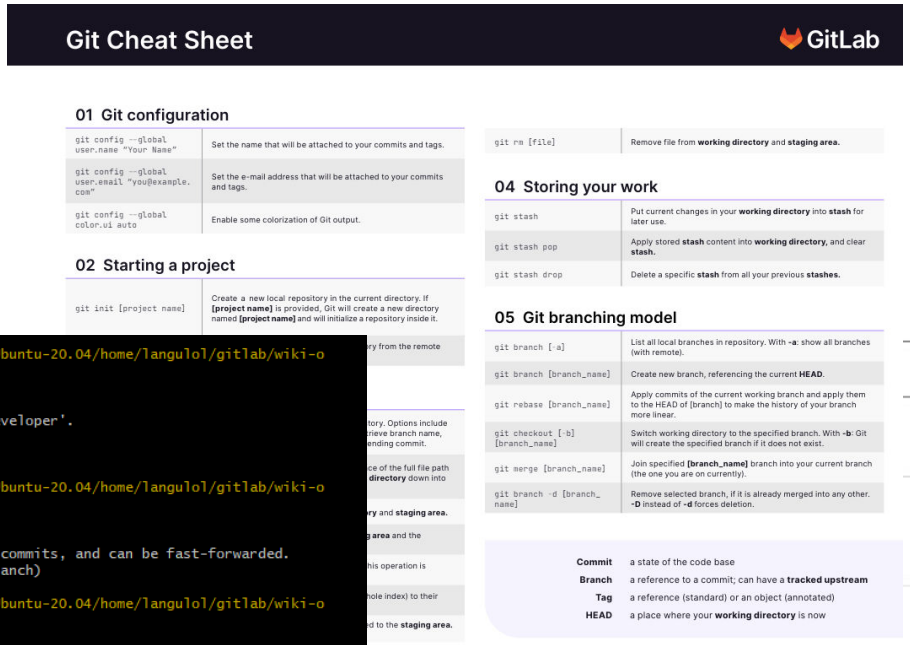
Génération de fichiers NetCDF avec Python3

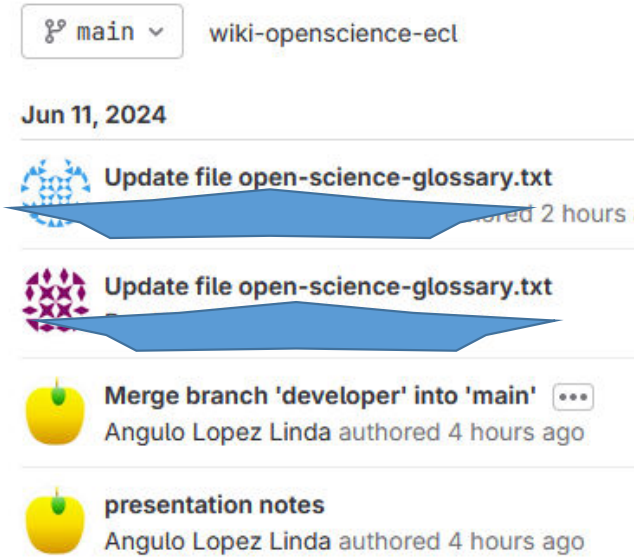
Voir: https://desktop.arcgis.com/fr/arcmap/latest/manage-data/netcdf/fundamentals-of-netcdf-data-storage.htm

https://www.cs.ubbcluj.ro/weadl/wp-content/uploads/2022/06/Arild%20Burud%20-%20Introduction%20to%20netCDF%20and%20THREDDS.pdf

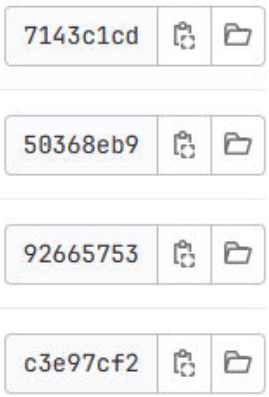
EXO-1 : clone > pull> checkout –b branch > switch > edit > commit > push > merge (owner, maintainer)

- fork, si pas invité !!
- Git Cheat Sheet

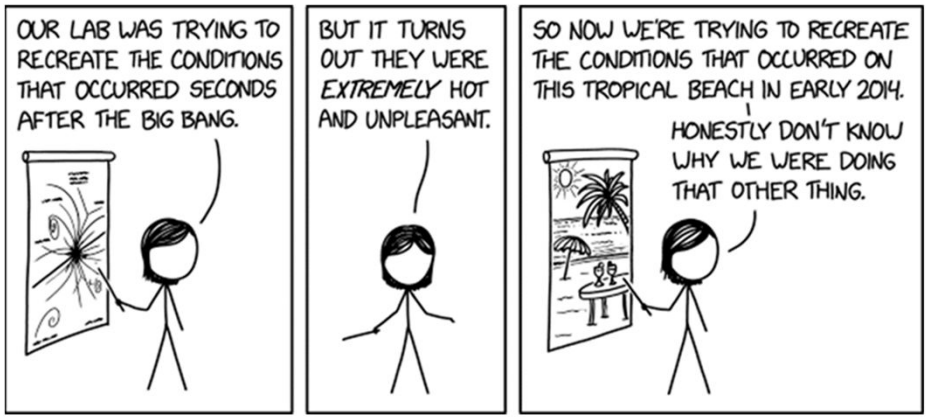


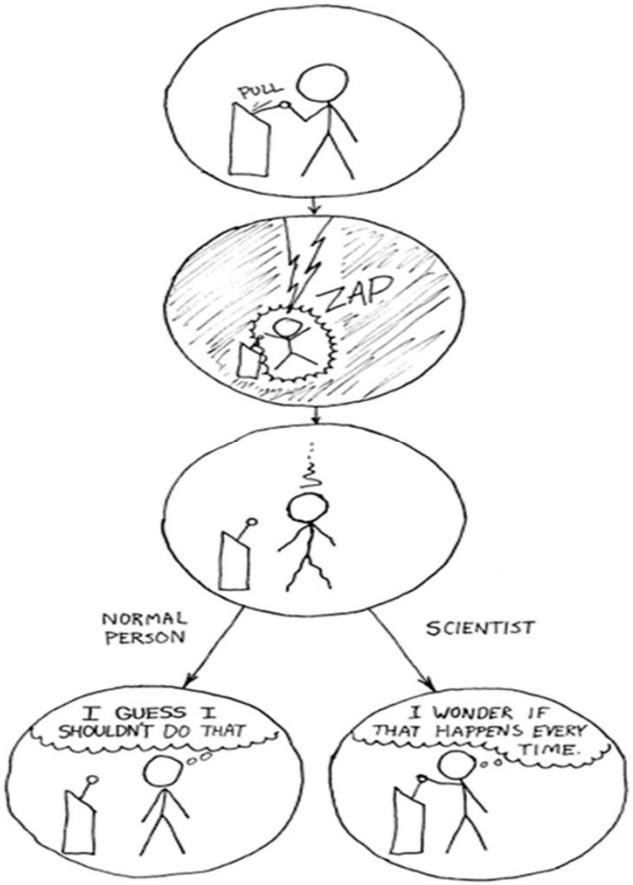



```
langulol@IN1958 MINGW64 //wsl.localhost/Ubuntu-20.04/home/langulol/gitlab/wiki-openscience-ecl (developer)
$ git status
On branch developer
Your branch is up to date with 'origin/developer'.

nothing to commit, working tree clean

langulol@IN1958 MINGW64 //wsl.localhost/Ubuntu-20.04/home/langulol/gitlab/wiki-openscience-ecl (developer)
$ git switch main
Switched to branch 'main'
Your branch is behind 'origin/main' by 7 commits, and can be fast-forwarded.
  (use "git pull" to update your local branch)

langulol@IN1958 MINGW64 //wsl.localhost/Ubuntu-20.04/home/langulol/gitlab/wiki-openscience-ecl (main)
$ git pull
Updating 67b4c44..7143c1c
Fast-forward
 ...51es-et-codes-de-recherche-centrale-lyon_1.pdf" | Bin 0 -> 5419352 bytes
 ...1es-et-codes-de-recherche-centrale-lyon_1.pptx" | Bin 0 -> 11282177 bytes
 open-data-code-2024/exo-1/exo-1.txt                |  88 +++
 .../exo-1/open-science-glossary.txt                |   2 +
 open-data-code-2024/notes_ppt-1.txt                | 795 ++++++++++++++++++++++
 5 files changed, 885 insertions(+)
 create mode 100644 "open-data-code-2024/donn\303\251es-et-codes-de-recherche-centrale-lyon_1.pdf"
 create mode 100644 "open-data-code-2024/donn\303\251es-et-codes-de-recherche-centrale-lyon_1.pptx"
 create mode 100644 open-data-code-2024/exo-1/exo-1.txt
 create mode 100644 open-data-code-2024/exo-1/open-science-glossary.txt
 create mode 100644 open-data-code-2024/notes_ppt-1.txt

langulol@IN1958 MINGW64 //wsl.localhost/Ubuntu-20.04/home/langulol/gitlab/wiki-openscience-ecl (main)
$
```



Code & données de la recherche (2/2)





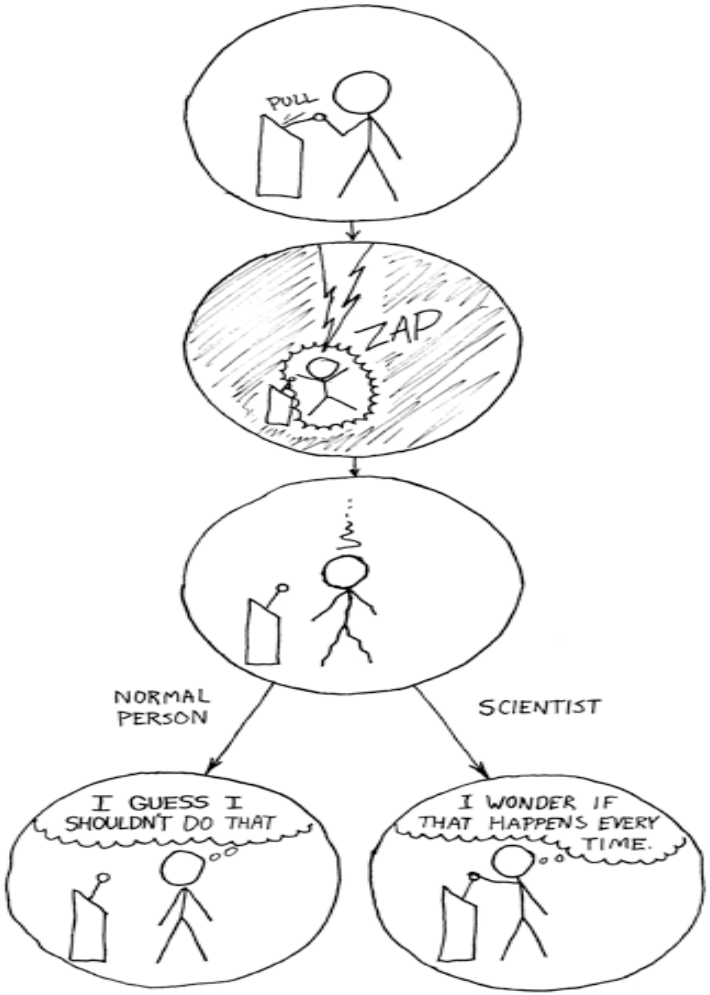
UNIVERSITÉ DE LYON
ÉCOLE CENTRALE LYON

36, avenue Guy de Collongue 69130 Écully - France
+33 (0)4 72 18 60 00

www.ec-lyon.fr


ÉCOLE
CENTRALE
DE LYON
Produire des faits vérifiables
Reproducibility
ÉCOLE
CENTRALELYON

L'environnement logiciel, source de variabilité



Reproductibilité

- recréer exactement le même environnement

Réplicabilité

- suffisamment bien documenté et partagé

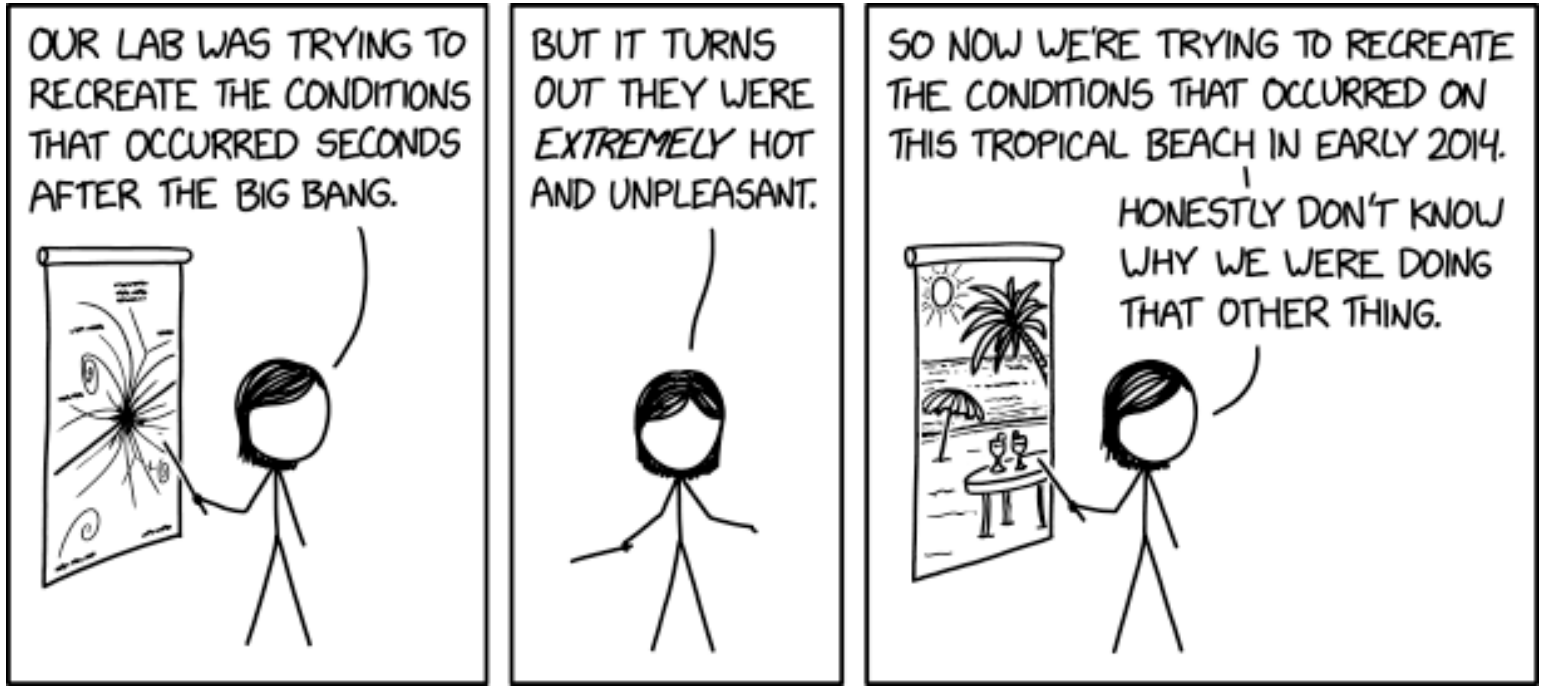
Validation des résultats

- cohérent et bien défini

Produire des faits vérifiables

Trois axes de reproductibilité :

- **méthode**
- **résultats**
- **inférence**

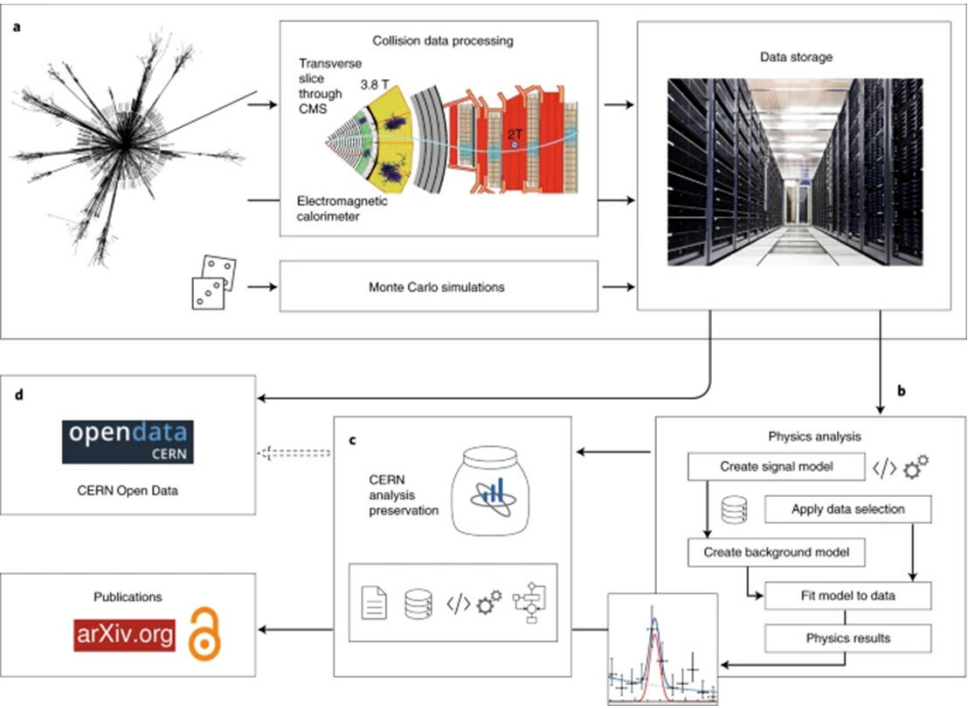


Défis et enjeux de la reproductibilité

- L'amélioration continue des connaissances scientifiques.

- Evénement bien décrit, observé à plusieurs reprises et provoqué par différentes personnes.

- La répétabilité et la réplicabilité



- Evaluée par des méthodes statistiques et probabilistes.
- Contraintes temporelles, financières et techniques.
- CERN, des stratégies pour la recherche reproductible

Rétractations : Scandales majeurs

- **Équipe de Ranga Dias**

- Article de 2023 dans Nature sur le supraconducteur à température ambiante rétracté

- Données fabriquées et falsifiées découvertes

- **Équipe de Microsoft Quantum**

- Article de 2018 dans Nature sur les particules de Majorana rétracté

- Données sélectionnées, résultats invalidés

Les physiciens s'inquiètent pour leur réputation et leur avenir

"If you think of research as a product that is paid for by the taxpayer, then Reproducibility is the Quality Assurance Department,..."

Frolov, ICRCMP, 2024

Processus et attitudes actuel

- Publications soumise aux éditeurs, revue par des référents anonymes
- Mauvais résultats publiés, problèmes de responsabilité
- Besoin d'une meilleure assurance qualité et reproductibilité
- Aborder la pression systémique à publier

Quoi faire

- Encourager des vues positives sur le partage des données
- Créer une culture de résultats partageables et reproductibles

Renforcer la confiance dans la science

- Les journaux et les agences de financement
- Créer une culture de résultats partageables et reproductibles

Défis pratiques

- Soutien nécessaire pour les étudiants en doctorat pour la préparation des données

Travailler en Collaboration à l'échelle Européen

- Les projets européens, depuis 1984
- 31 projets Horizon 2020
- Environnements et Dépendances

HORIZON 2020
LE PROGRAMME DE RECHERCHE ET D'INNOVATION DE L'UNION EUROPÉENNE

Travailler en Collaboration avec vos collègues

Proposer des outils Informatiques

- CSV, netCDF, … + Code
- Jupyter, RStudio, Emacs ou Git
- DMP-Opindor
- Research Data Gouv
- Newton et CC-IN2P3

Façons de procéder

- l'enthousiasme
- la gentillesse
- l'indifférence

L'environnement logiciel, source de variabilité

Programme (code; script)

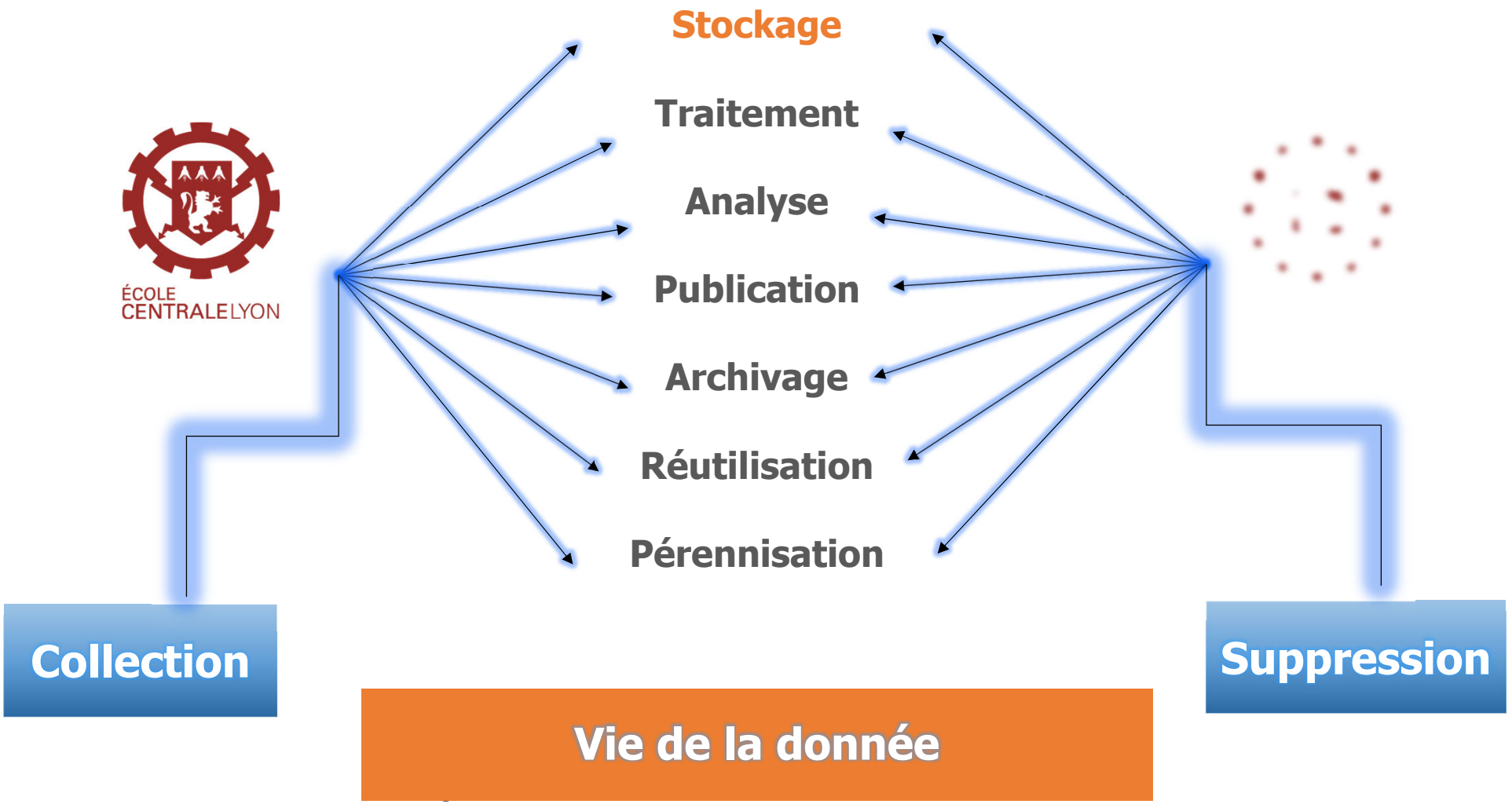
Données (Bruts; traitées)

Environnement (Matériel; logiciel)

Workflow

Recherche Reproductible

Cycle de vie des données



Newton HPC

- Espace utilisateur, $HOME
- Volume scratch
- Espaces de stockage par laboratoire
- Volume dédié aux IO intensifs
- Espace des modules et logiciels communs

bash Copy code

```
Disk quotas for user USER (uid XXXX):
     Filesystem   space   quota   limit   grace   files   quota   limit   grace
service2:/home/LABO/USER
                 1010G   1048G   1048G            837k       0       0
```

Les gestionnaires d'environnement logiciels

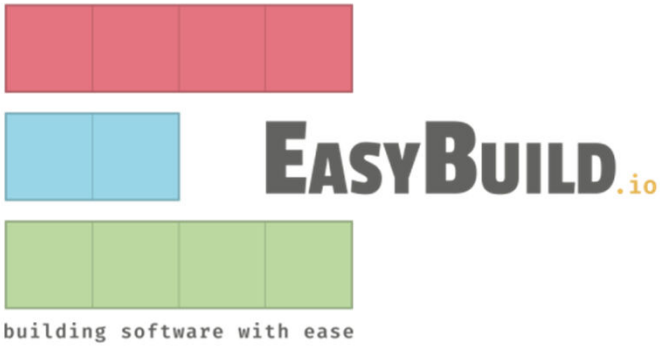
- Systèmes d'exploitation (apt-get; yum)
- Packages de langage (pip_Python)
- Généralistes HPC (spack ; easybuild)
- Pour la reproductibilité HPC (nix; guix)

Le gestionnaire de paquets, Python

- Environnements virtuels
- Gestion des dépendances
- Documentation des environnements

- Contrôle de version
- Partage et distribution

Cohérence des compilations de logiciels



pmcs2i.ec-lyon

- Mise en place initiale
- Assurer la cohérence
- Faciliter le développement collaboratif
- Déploiements & reversions
- Corrections de Bugs
- Revenir à une version précédente

Cycle de vie des données

Le standard d'échange de données pour l'archivage (SEDA)

Issu d'une collaboration lancée dès 2006 entre les Archives de France et l'ancienne Direction générale de la modernisation de l'État (DGME), le standard d'échange de données pour l'archivage (SEDA) vise à faciliter l'interopérabilité entre le système d'information d'un service d'archives et les systèmes d'information de ses partenaires dans le cadre de leurs échanges de données.

Transaction et acteurs + Formalisme

Préservation du patrimoine culturel et scientifique

Type de document	DUA	Sort final	Observations
Comptes rendus des réunions du conseil de service et du comité des directeurs de laboratoires	8 ans	C	Les autres collections de comptes rendus qui peuvent exister à la délégation sont à détruire.

- Lieux de versement des archives historiques, CC-IN2P3
- Soutien technique et administratif aux unités de recherche et de service, PMCS21
- la typologie des documents, durée d'utilité administrative
- le sort final qui doit être appliqué aux documents (C, T, D)
- Observations nécessaires à la compréhension et à la mise en œuvre
    - Schémas de métadonnées (descriptives, structurelles et administratives)
    - Définir clairement les droits d'utilisation
    - Mettre régulièrement à jour les supports de stockage
    - Assurer un financement adéquat et un soutien politique
    - Prévoir l'obsolescence technologique
- https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006074236/LEGISCTA000006129161/

Pérennisation des Données et Code au CC-IN2P3

- Bandes magnétiques
- Redondance et sauvegarde
- Accès à long terme

- Gestion des versions
- Documentation
- Partage et collaboration

Cycle de vie des données

Code, Types de Licences open source

Produits gratuits

- Licence publique générale GNU (GPL)

Produits commerciaux

- Licence Apache
- Licence MIT

Licence permissive

- Licence BSD
- Creative Commons

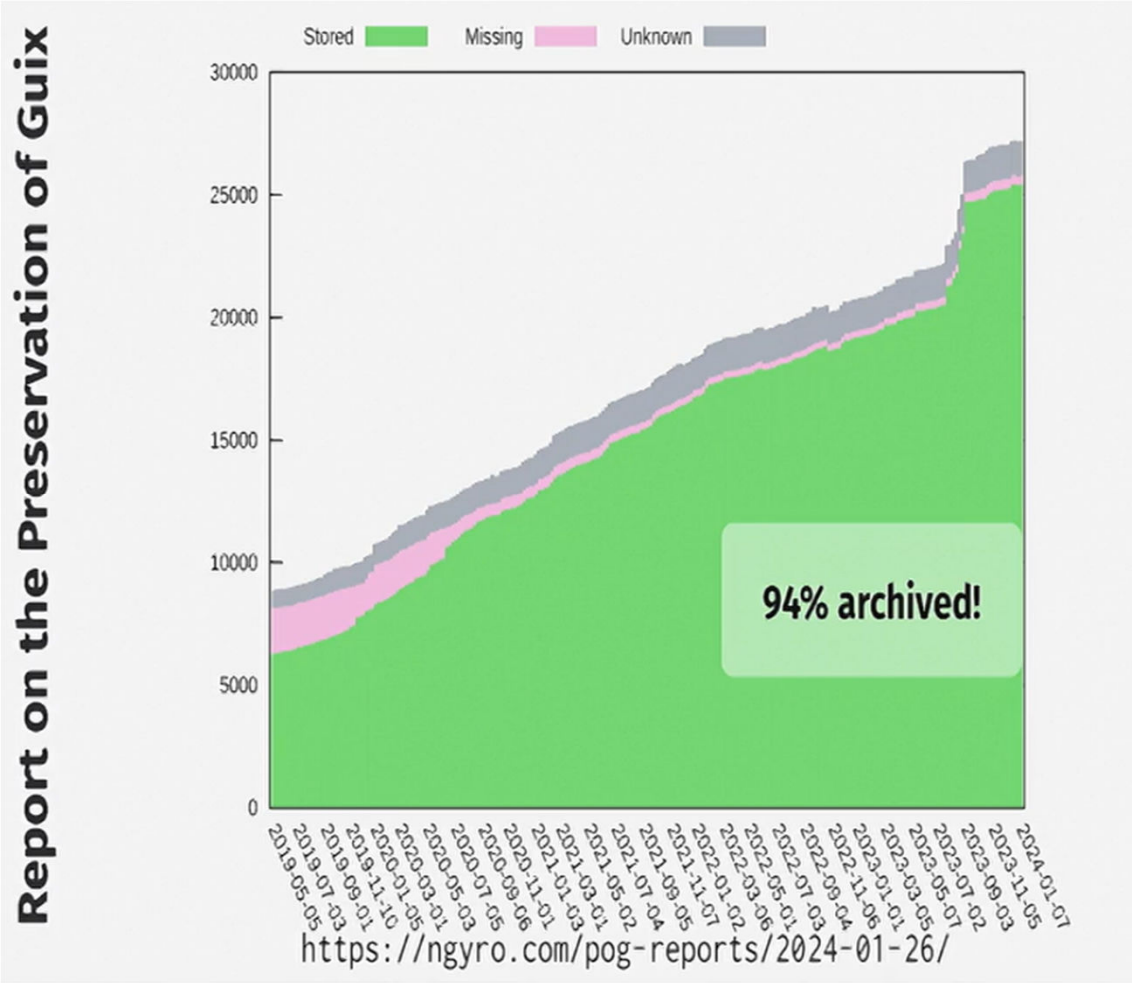
Cycle de vie des données

Assurer la durabilité et la disponibilité continue

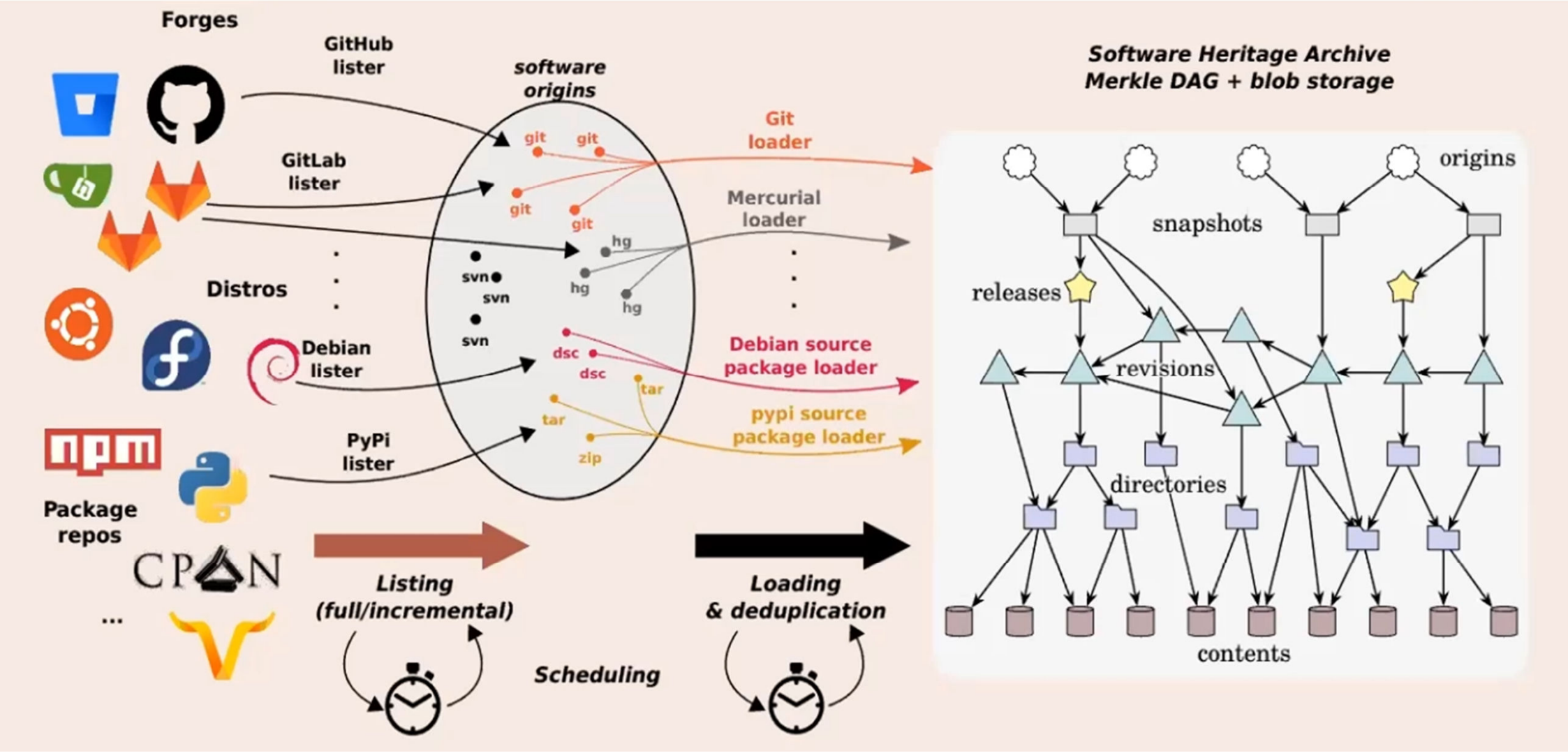
- Normes et Formats Durables
- Gestion des Métadonnées
- Politiques et Cadres Juridiques
- Maintenance et Actualisation
- Gestion de la Confidentialité et de la Sécurité
- Engagement Communautaire

Guix, conteneur avec guix time-machine

- Software Heritage
- env. virtuel universel
- voyage dans le temps
- & l'espace



Code, Forges logicielles et Software Heritage



L'Impératif de la Recherche Ouverte et Reproductible

nature physics

PERSPECTIVE

https://doi.org/10.1038/s41567-018-0342-2

Corrected: Publisher Correction

OPEN

Open is not enough

Xiaoli Chen[1,2], Sünje Dallmeier-Tiessen[1]*, Robin Dasler[1,11], Sebastian Feger[1,3], Pamfilos Fokianos[1], Jose Benito Gonzalez[1], Harri Hirvonsalo[1,4,12], Dinos Kousidis[1], Artemis Lavasa[1], Salvatore Mele[1], Diego Rodriguez Rodriguez[1], Tibor Šimko[1]*, Tim Smith[1], Ana Trisovic[1,5]*, Anna Trzcinska[1], Ioannis Tsanaktsidis[1], Markus Zimmermann[1], Kyle Cranmer[6], Lukas Heinrich[6], Gordon Watts[7], Michael Hildreth[8], Lara Lloret Iglesias[9], Kati Lassila-Perini[4] and Sebastian Neubert[10]

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness.

L'environnement logiciel, source de variabilité

Programme (code; script)

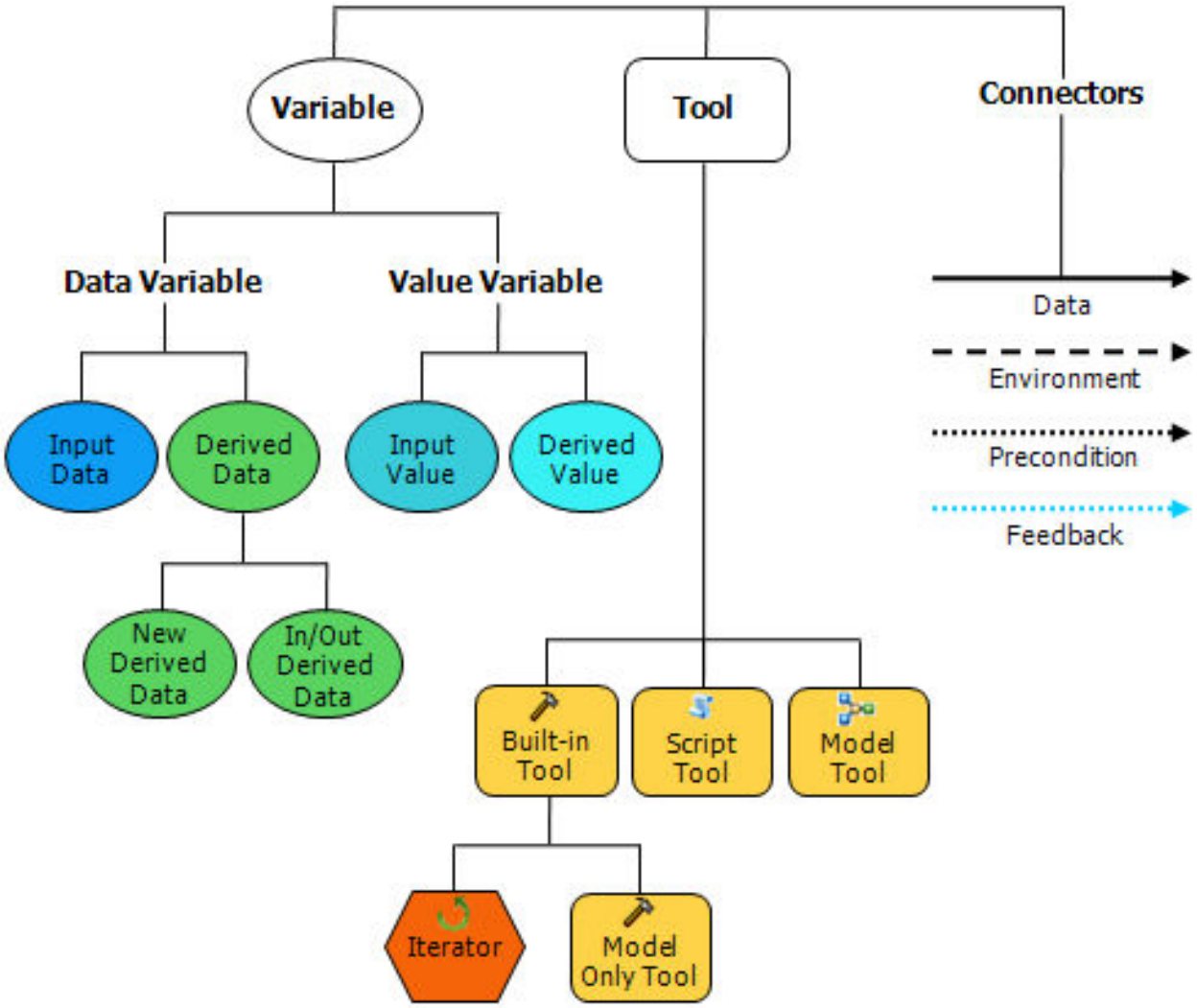
Données (Bruts; traitées)

Environnement (Matériel; logiciel)

Workflow

Recherche Reproductible

Work flow, Tâches Ordonnées et Automatisées



- Gestion des versions de code
- Documentation
- Automatisation
- Environnements
- FAIR

Exemple de workflow reproductible

Développement du code : Utilisation de Git pour le contrôle de version
Automatisation : Création de scripts pour le traitement des données
Documentation : Rédaction d'un guide utilisateur
Environnements : Déploiement avec Docker

Cycle de vie des données

Data papers et data journals

- Communiquer sur l'existence des données et permettre de les trouver
- Créditer les auteurs (reconnaissance, référence citable) et valoriser les données
- Faciliter la réutilisation des données, en les rendant intelligibles

- Dans un data journal, revue dédiée à ce type de publication
- Dans une revue classique qui publie des data papers en plus des articles traditionnels.
- Éléments communs aux articles classiques et spécifiques liés aux données.

Exemple de structure de data paper

UN ACCÈS AUX DONNÉES

Le data paper fournit l'accès aux données qu'il décrit. Les données peuvent être :

INTÉGRÉES AU DATA PAPER

ELSEVIER

Volume 2, March 2015, Pages 42–47

Open Access

Titre : le titre doit se concentrer sur les données spécifiques partagées, ce n'est pas un titre d'article de recherche
Auteur(s) : nom, affiliations, email,..
DOI : https://doi.org/10.1016/j.dib.2014.12.001
Type de licence : CC-By...
Date de l'article : date de soumission de l'article, date de publication, date de validation, date de révision.
Résumé : présentation du contexte d'obtention des données (front de recherche, question de recherche)
Métadonnées :

Domaine de recherche	Physique, chimie, psychologie, etc.
Domaine plus spécifique	Ex : physique nucléaire
Type de données	Tableau, image (radiographie, microscopie...), texte, graphique, figure, etc.
Moyens d'acquisition des données	Microscopie, enquête (vue générale), MEB, RMN, spectroscopie de masse, etc.
Format des données	Brut, filtré, analysé, etc.
Facteurs expérimentaux	Description brève de la préparation des échantillons
Caractéristiques expérimentales	Description expérimentale très brève
Emplacement de source de données	Ville, Pays, coordonnées GPS pour échantillons ou données
Accès aux données	Nom de l'entrepôt, son DOI, ou l'Url directe aux données
Article de recherche lié	Si les données accompagnent un article de recherche, citez-le

Intérêt du jeu de données : décrire la valeur scientifique de ces données
Description des données : décrire brièvement les données partagées
Matériels et méthodes : description complète de la méthode d'obtention des données. Inclure n'importe quelles figures/tables pertinentes à la compréhension des données.
Remerciements
Références : références des données citées.

JOINTES AU DATA PAPER

sous forme de matériel supplémentaire

Transparency document. Supplementary material

Download Word document (10KB) Help with docx files

Supplementary material

Appendix A. Supplementary material

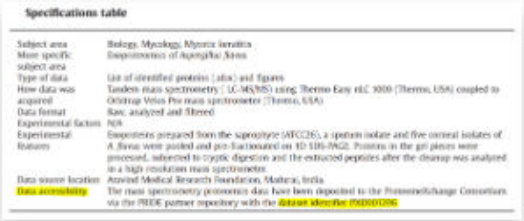
Download text file (24KB) Help with txt files

Supplementary material

Dassou, A. G., Carval, D., Dépigny, S., Fansi, G., & Tixier, P. (2016). Dataset on the abundance of ants and Cosmopolites sordidus damage in plantain fields with intercropped plants. Data in Brief, 9, 17-23. http://dx.doi.org/10.1016/j.dib.2016.08.027

DÉPOSÉES DANS UN ENTREPÔT

auquel cas le data paper fournit l'identifiant pérenne (type DOI) permettant de faire le lien vers l'entrepôt en question.



Specifications table

Selvam, R. M. et al. (2015). Data set for the mass spectrometry based exoproteome analysis of Aspergillus flavus isolates. Data in brief, 2, 42-47. http://dx.doi.org/10.1016/j.dib.2014.12.001

Rédiger les data papers

Outils pour rédiger les data papers

- Boîte à outils de publication intégrée du GBIF (IPT)

- NephilaPapier

- Outil d'écriture arpha

Critères d'acceptation

- Prend en charge les principes de données FAIR

- les données doivent être collectées à l'aide d'une méthode scientifique et avoir une valeur pour la communauté des chercheurs

- N'acceptera généralement pas les soumissions de descriptions de données qui ont déjà été publiées comme matériel supplémentaire dans un article de recherche original.

Crédits image et sources principales

Formations

- Ecole Centrale de Lyon, A.S.P.I.C.S.
- FUN MOOCS
  - Sorbonne Université, La science ouverte
  - INRIA, Recherche reproductible : principes méthodologiques pour une science transparente
  - INRIA, Recherche reproductible II: Practices and tools for managing computations and data

Ressources en ligne

- https://mi-gt-donnees.pages.math.unistra.fr/site/guide.html
- https://www.recherche-reproductible.fr/
- https://www.ouvrirlascience.fr/science-ouverte-donnees-de-la-recherche
- https://recherche.data.gouv.fr/fr/page/classes-virtuelles
- https://dmp.opidor.fr
- https://cat.opidor.fr
- https://printempsdeladonnee.fr/calendrier/
- https://doranum.fr/
- https://www.canal-u.tv/chaines/ad/journee-gitlab/software-heritage-l-archive-universelle-a-la-croisee-des-forges
- https://xkcd.com
- http://corist-shs.cnrs.fr/sites/default/files/ressources/guide_bonnes_pratiques_gestion_donnees_recherche_v2.pdf
- https://desktop.arcgis.com/fr/arcmap/latest/manage-data/netcdf/fundamentals-of-netcdf-data-storage.htm
- https://desktop.arcgis.com/fr/arcmap/latest/manage-data/netcdf/fundamentals-of-netcdf-data-storage.htm
- https://www.dgdr.cnrs.fr/bo/2007/12-07/433-bo1207-insdAf-dpAci-res-2007-002.htm
- https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006074236/LEGISCTA000006129161/
- https://www.gbif.org/fr/event/7j0sFhaug80CIMqs6aWegK/webinar-introduction-to-data-papers

GLOSSAIRE DONNEES DE LA RECHERCHE (2/2)

CHEATSHEET, DONNEES DE LA RECHERCHE (2/2)
Formation Doctorale Ecole Centrale de Lyon

Données de la recherche	Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont reconnus par la communauté scientifique comme nécessaires pour la validation des résultats.
Données à caractère personnel	Données concernant une personne physique qui est identifiée ou identifiable, par exemple par corrélation avec d'autres jeux de données.
Éditeurs prédateurs ou revues prédatrices	Éditeurs ou revues ayant des pratiques d'évaluation par les pairs ou des pratiques commerciales et/ou contestables (prix excessifs, contrôle des réutilisations, etc.).
Embargo	Période pendant laquelle une production scientifique ne peut pas être diffusée en accès ouvert.
Publication en accès ouvert	Publication accessible à son lectorat immédiatement et sans restriction d'accès. Son financement peut provenir de subventions publiques ou de sociétés savantes, de paiements par les institutions des auteurs (voir APC), de financement par les bibliothèques universitaires, etc.
Trustworthy Repositories	Criteria for the Selection of Trustworthy Repositories, ils reçoivent le label de certification qui vise à promouvoir des entrepôts de données fiables et durables.
Logiciel	Texte, écrit dans un ou plusieurs langages informatiques, décrivant des calculs destinés à être exécutés par un ordinateur. Il peut prendre diverses formes : fichiers de code, assemblage graphique, formules de tableur, cahier computationnel, etc.
Forge	Environnement de développement logiciel facilitant le travail collaboratif autour d'un projet logiciel. Une forge contient des outils tels que le dépôt versionné de code source, des forums de discussion, un environnement de tests automatisés, etc.


ÉCOLE
CENTRALE
DE LYON
Je vous remercie pour votre attention.
ÉCOLE
CENTRALELYON