Code & données de la recherche (1/2)

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

ÉCOLE CENTRALE DE LYON

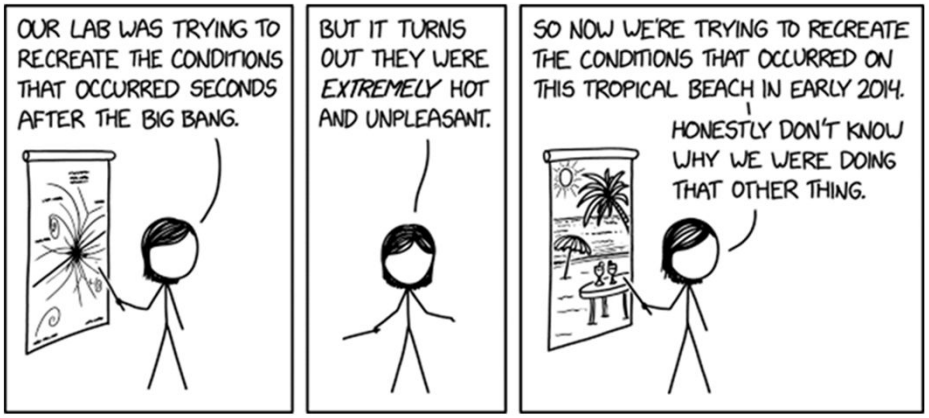
Formation doctorale : Bibliothèque Michel Serres

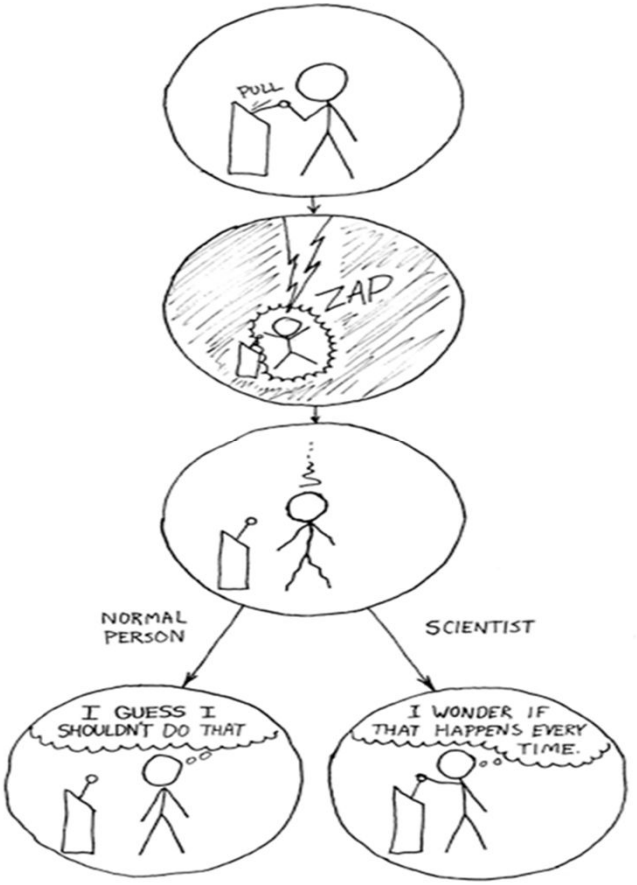
ÉCOLE CENTRALELYON

Gestion des données de recherche : optimisation de l'impact de vos travaux(1/2).

ÉCOLE
CENTRALELYON

Code & données de la recherche (1/2)





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

Budapest Open Access Initiative, 22eme anniversaire

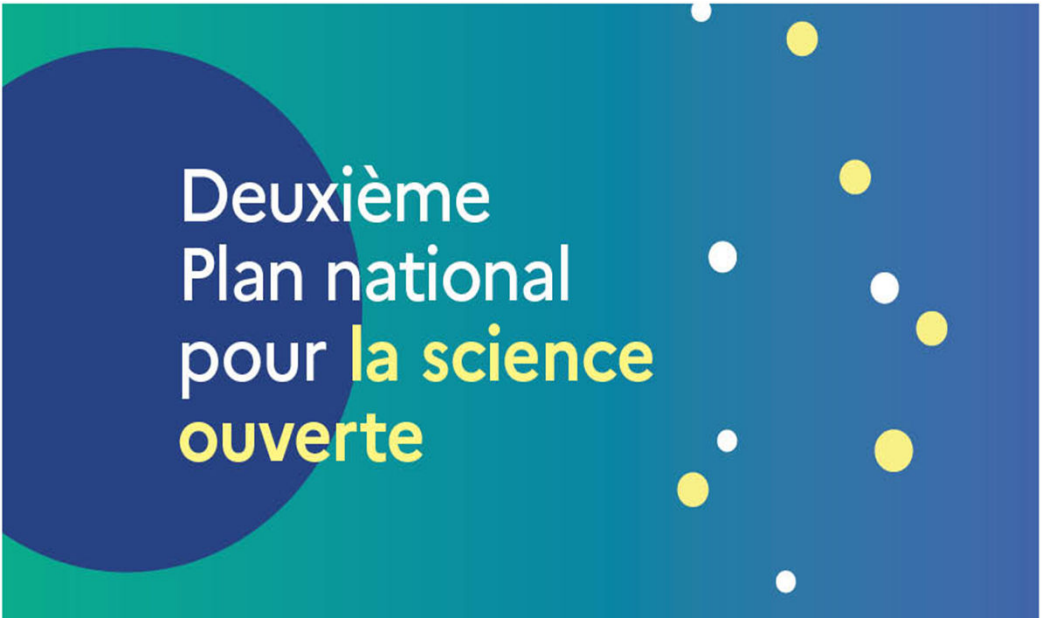
- **Héberger la recherche OA sur des infrastructures ouvertes**
- **Réformer l'évaluation de la recherche et les récompenses pour améliorer les incitations**
- **Favoriser des canaux de publication et de distribution inclusifs**
- **Dépenser l'argent pour publier la recherche OA en gardant en mémoire les objectifs de l'OA**

#BOAI20

2eme Plan Nationale Pour La Science Ouverte 2021-2024

Les 4 axes du 2e Plan national pour la science ouverte

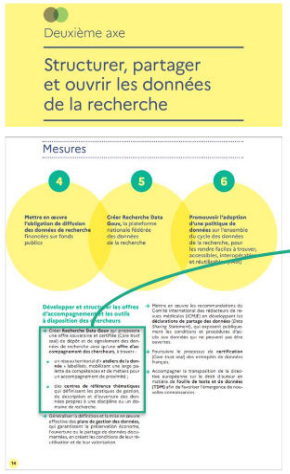
- **Généraliser l'accès ouvert aux publications**
- **Structurer, partager et ouvrir les données de la recherche**
- **Ouvrir et promouvoir les codes sources produits par la recherche**
- **Transformer les pratiques pour faire de la science ouverte le principe par défaut**



Objectif de 100% de publications en accès ouvert en 2030 fixé par la loi de programmation de la recherche.

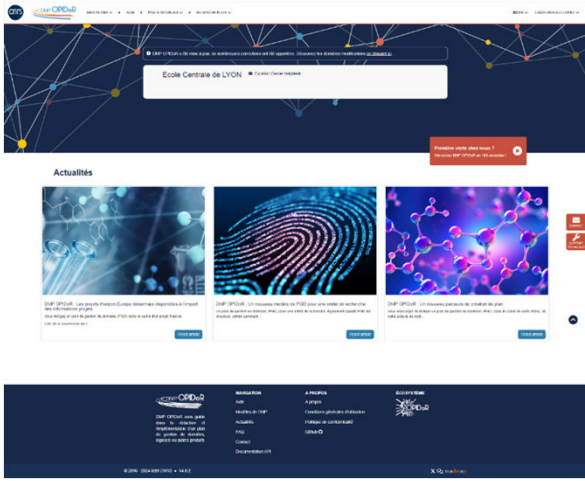
2e Feuille de Route Science Ouverts, à l'ECL

- Ouvrir le processus de recherche et les résultats à tous les acteurs de la société
- Faciliter les collaborations scientifiques et tirer profit de la science des données pour nos enjeux de recherche
- Saisir les opportunités et maîtriser les risques potentiels liés à l'ouverture
- Chargée des données de Recherche



Créer **Recherche Data Gouv** qui proposera une offre souveraine et certifiée (*Core trust seal*) de dépôt et de signalement des données de recherche ainsi qu'une **offre d'accompagnement des chercheurs**, à travers:

- un réseau territorial d'« **ateliers de la donnée** » labellisés, mobilisant une large palette de compétences et de métiers pour un accompagnement de proximité ;
- des **centres de référence thématiques** qui définissent les pratiques de gestion, de description et d'ouverture des données propres à une discipline ou un domaine de recherche.



Acteurs d'Ouverture de Code et Données de la Recherche

Clarisse MARANDIN
Direction Bibliothèque

Christophe CORRE
Direction de la Recherche

Bénédicte MARTIN
Resp Centrale Innov
Resp Aff. EU DPRV

Elisabeth DALVERNY
Responsable DRPV

Camille ZAMI-PIERRE
Juriste ECL

Nicolas JARDIN
Resp. services recherche

Matteo CAMIER
Resp. tech Pôle calcul / gestion des données de la recherche

Laurine MAIRE
Chargée d'Aff.
Centrale Innovation

Véronique MERAT
Ingénierie de projet
DPRV

Guillaume EMPTAZ
FSD

Stéphanie LAMAISON
HAL/open accès

Anne CADIOU
IR CNRS LMFA
Pôle calcul

Marine PICO
Chargée projets EU
Centrale Innovation

Angélique CATEUX
Suivi contrats Recherche

Christophe FESSART
DPO

Linda ANGULO LOPEZ
Chargée mission données

Benoît PIER
Dir. Rech CNRS LMFA
Membre GT COSO

AMI données
GT juridique

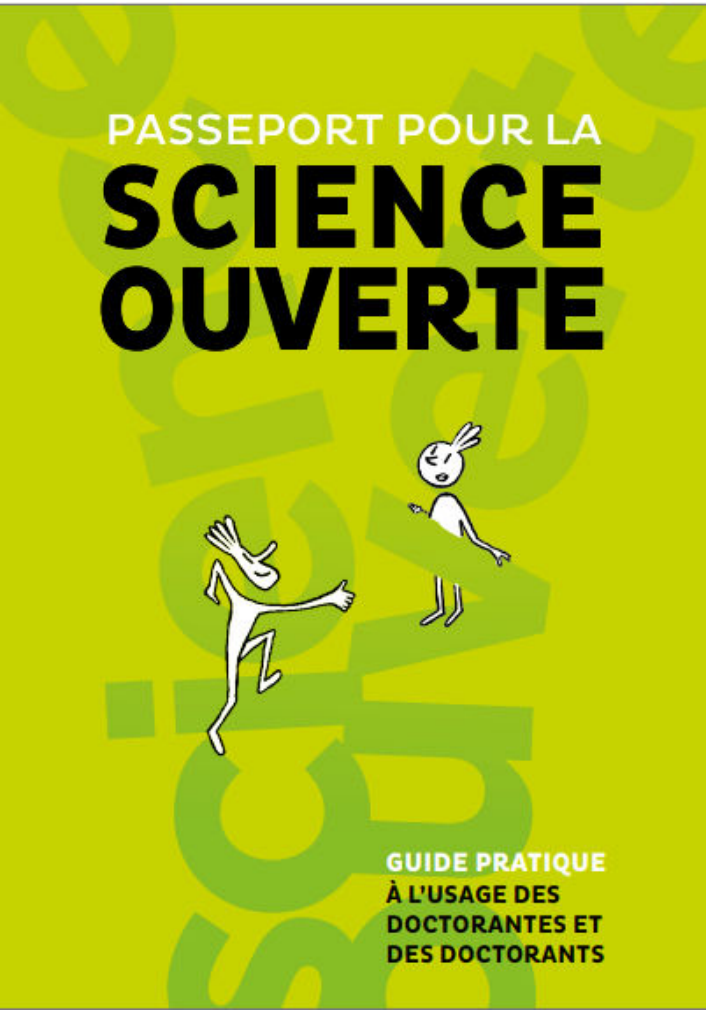
Robin MATEJICEK
Suivi projets Recherche ENISE

Groupe de Travaille Science Ouvert de Centrale Lyon

</> Fork

GT Traitement des données DataLyste

Recommandations aux doctorantes



Gestion des données de recherche : optimisation de l'impact de vos travaux(1/2).

ÉCOLE
CENTRALELYON

Cycle de vie des données

Politiques de sécurité et de confidentialité

La Loi République numérique

Données Publiques

Exceptions

- Droits d'auteur
- Droit sui generis
- Brevets
- Obtentions végétales
- Essais cliniques
- Biodiversité

RGPD

Données à caractère personnel avec dérogations :

- durée de conservation;
- recherches exploratoires;
- droit à l'information des personnes;
- droit à l'opposition

Aspects juridiques des données de la recherche

Droits d'auteur et Propriété intellectuelle

- Chercheurs conservent leurs droits d'auteur sur les œuvres créées dans leurs fonctions et peuvent les céder.

- Protège les investissements dans la création de bases de données et s'oppose à des extractions substantielles de contenu. Valable 15 ans, renouvelable.

Loi République numérique (2016) et Open Data

- **Données de recherche, assimilées à des données publiques, doivent être accessibles en ligne et réutilisables gratuitement.**

- **Exception Text and Data Mining (TDM), permise pour la recherche publique sous conditions (non-lucratif, accès licite, sécurité des systèmes et des fichiers).**

Exception & Politiques de Sécurité et de Confidentialité

La Loi République numérique

Données Publiques

Exceptions

- Obtention Droits d'auteur
- Droit sui generis
- Brevets
- Végétales
- Essais cliniques
- Biodiversité

RGPD

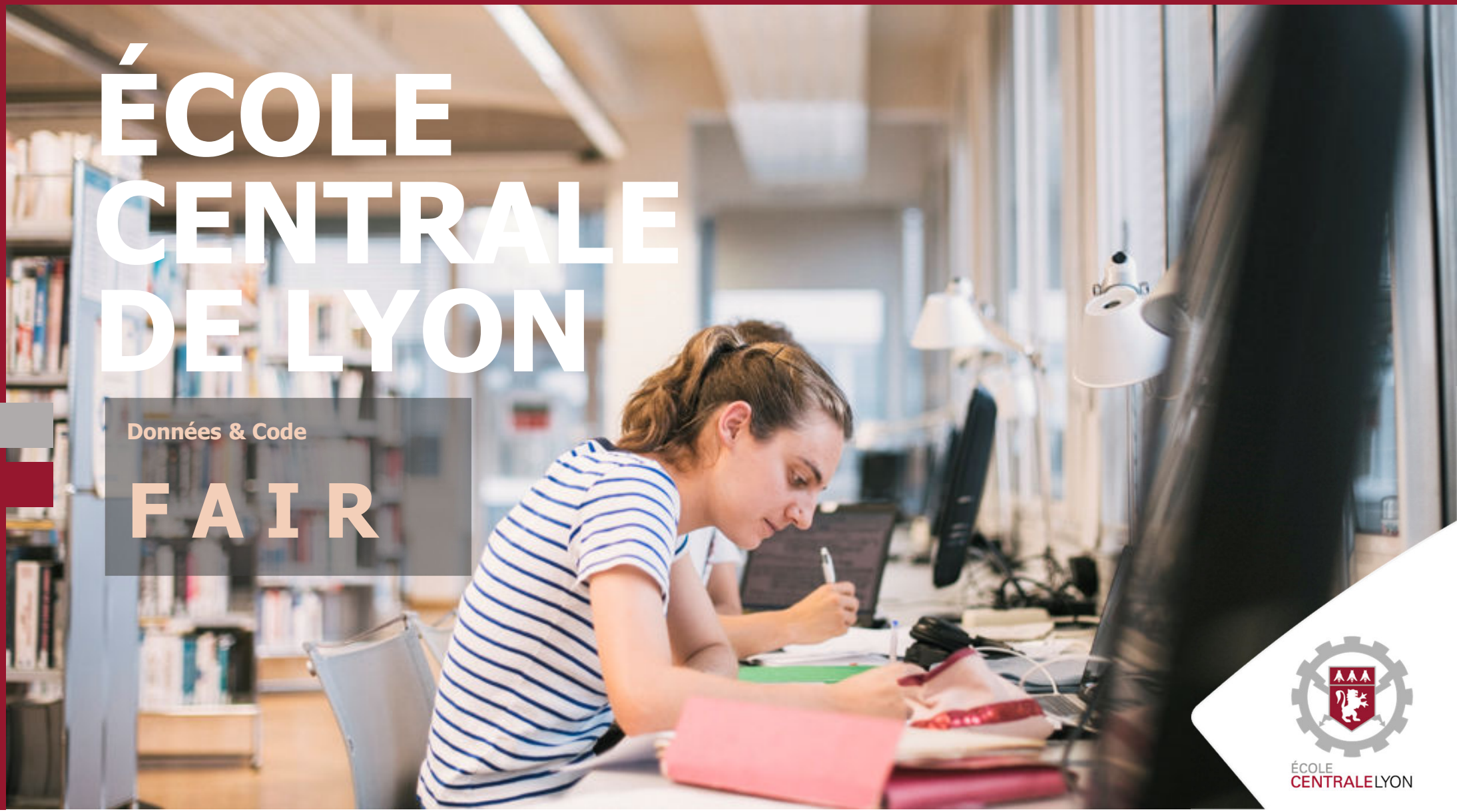
Données à caractère personnel avec dérogations :

- durée de conservation;
- recherches exploratoires;
- droit à l'information des personnes;
- droit à l'opposition

ÉCOLE CENTRALE DE LYON

Données & Code

F A I R



ÉCOLE CENTRALELYON

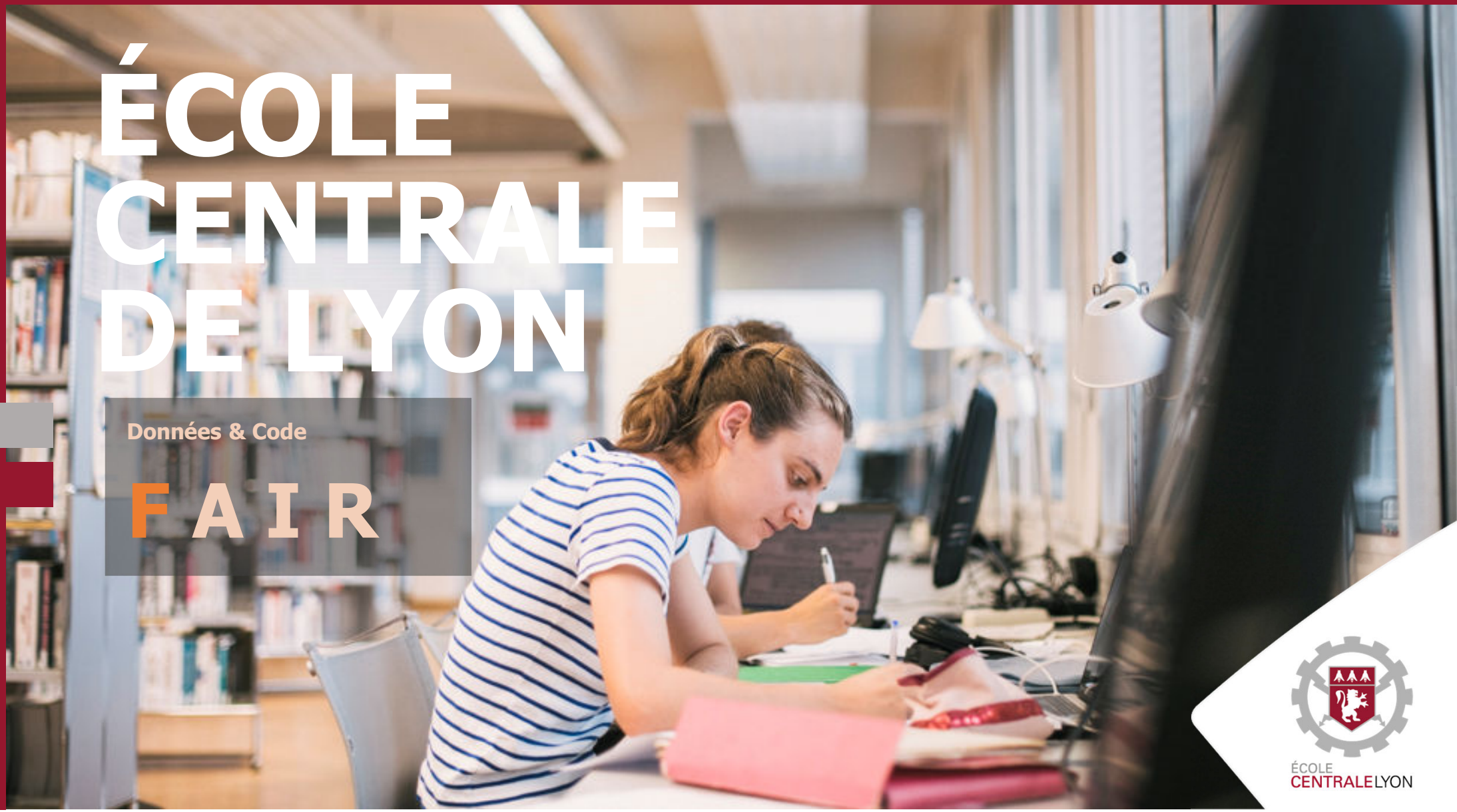
FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable)

Facile à trouver	**DOI ; métadonnées indiquant l'identifiant ; entrepôt permettant la recherche ; Décrie les métadonnées riches et utilisant des vocabulaires**
Accessible	**Données accessibles avec le DOI ; métadonnées toujours disponibles; Déposer des données ouverte, éventuellement sous conditions**
Interopérable	**Métadonnées standardisées et machine-readable ; Format de fichiers ouverts et machine-readable ; vocabulaires avec URI et URL**
Réutilisable	**Licences disponibles ; Décrie provenance des données ; standards correspondant à la communauté**

ÉCOLE CENTRALE DE LYON

Données & Code

FAIR



ÉCOLE CENTRALELYON

ÉCOLE CENTRALE DE LYON

Plan de Gestion des Données

ÉCOLE CENTRALELYON

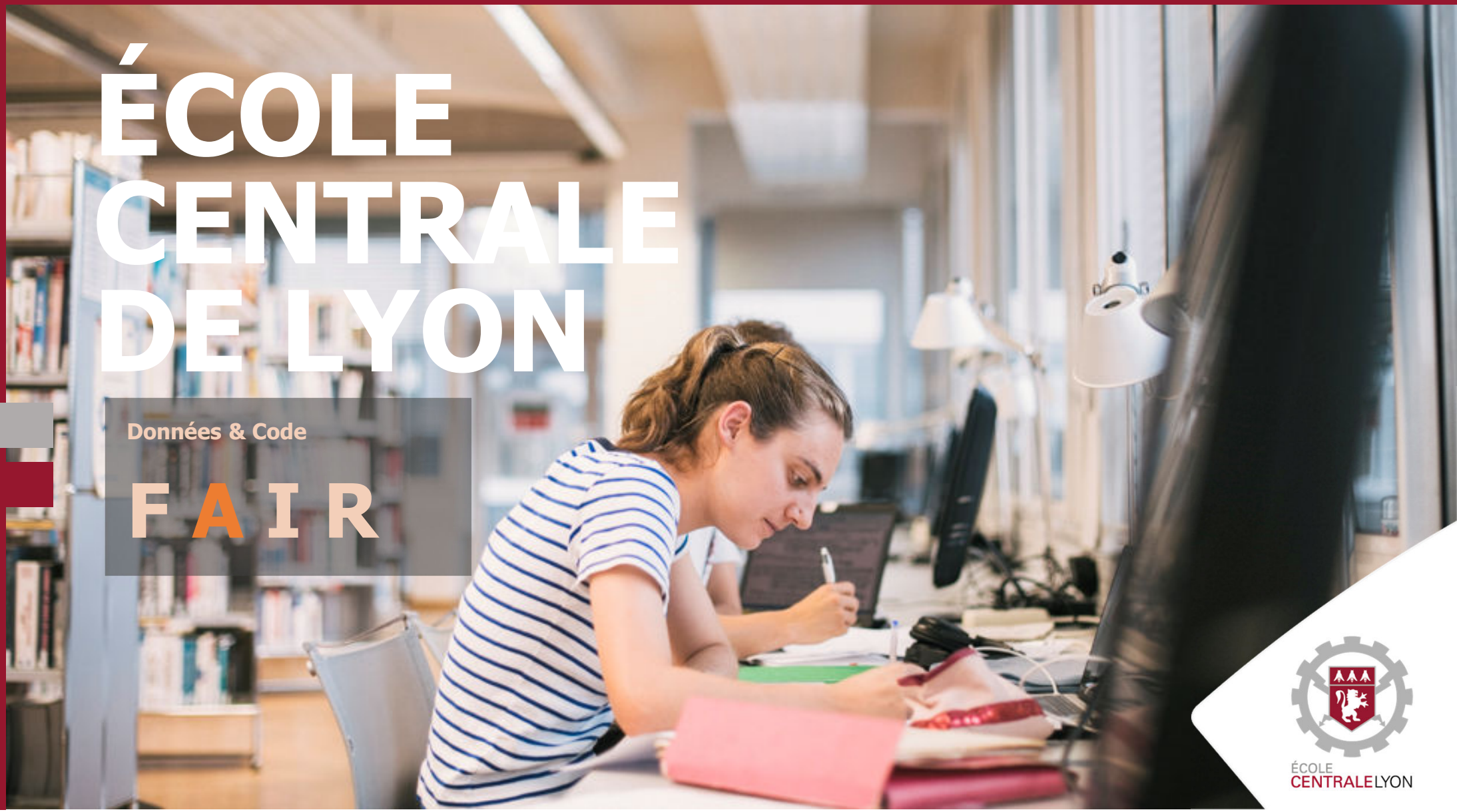
Plan de gestion des données

- L'équilibre entre ouverture et sécurité
- Le PGD évolue tout au long du projet
- Edité collectivement par exemple sur DMP OPIDoR
- Agences de financement de la recherche demandent désormais la fourniture d'un PGD

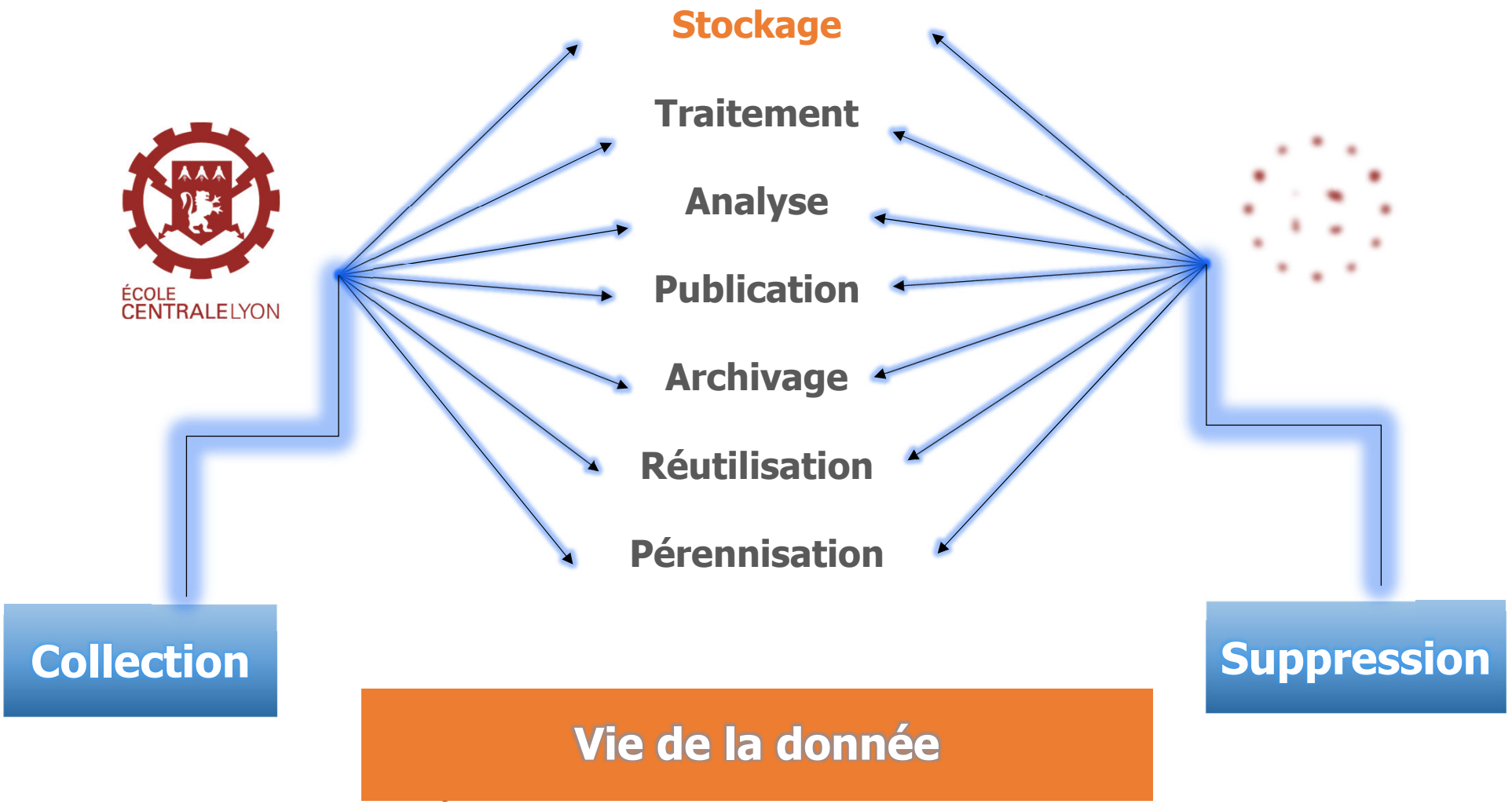
Cycle de vie des données

La collecte de données, RGPD

- Activités principales d'une équipe de recherche
- Garantie de la confidentialité de la donnée doit être intégré dès cette première étape
- Recueil doit se faire de façon justifiée, sécurisée et cloisonnée
- Consentement des personnes, active ou passive
- Limitation facilite le stockage et surtout la protection des données


ÉCOLE
CENTRALE
DE LYON
Données & Code
FAIR
ÉCOLE
CENTRALELYON

Cycle de vie des données

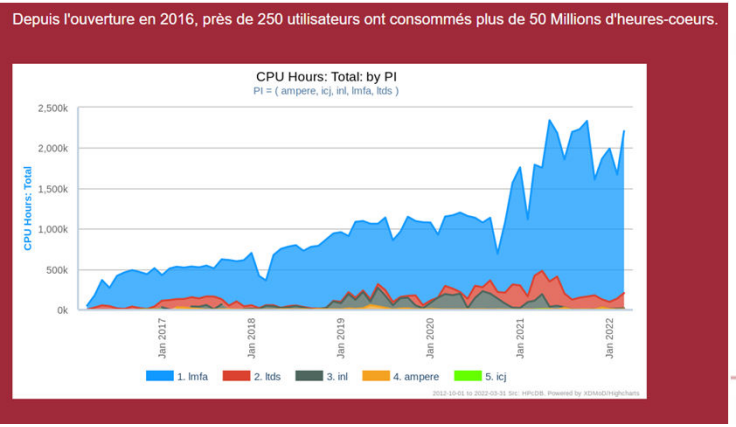


Le stockage des données

- DMP incitent à mieux préparer cette étape
- Fichiers nommés, retrouver facilement & distinguer versions
- La sauvegarde , 3 copies sur 2 supports différents, dont 1 copie à distance.
- Stocker ses données en cours de projet
- Données confidentialité, nécessite un espace adapté et s'interroger sur la durée de conservation
- Stockage Centraliser

Le Pôle de Calcul de l'École Centrale

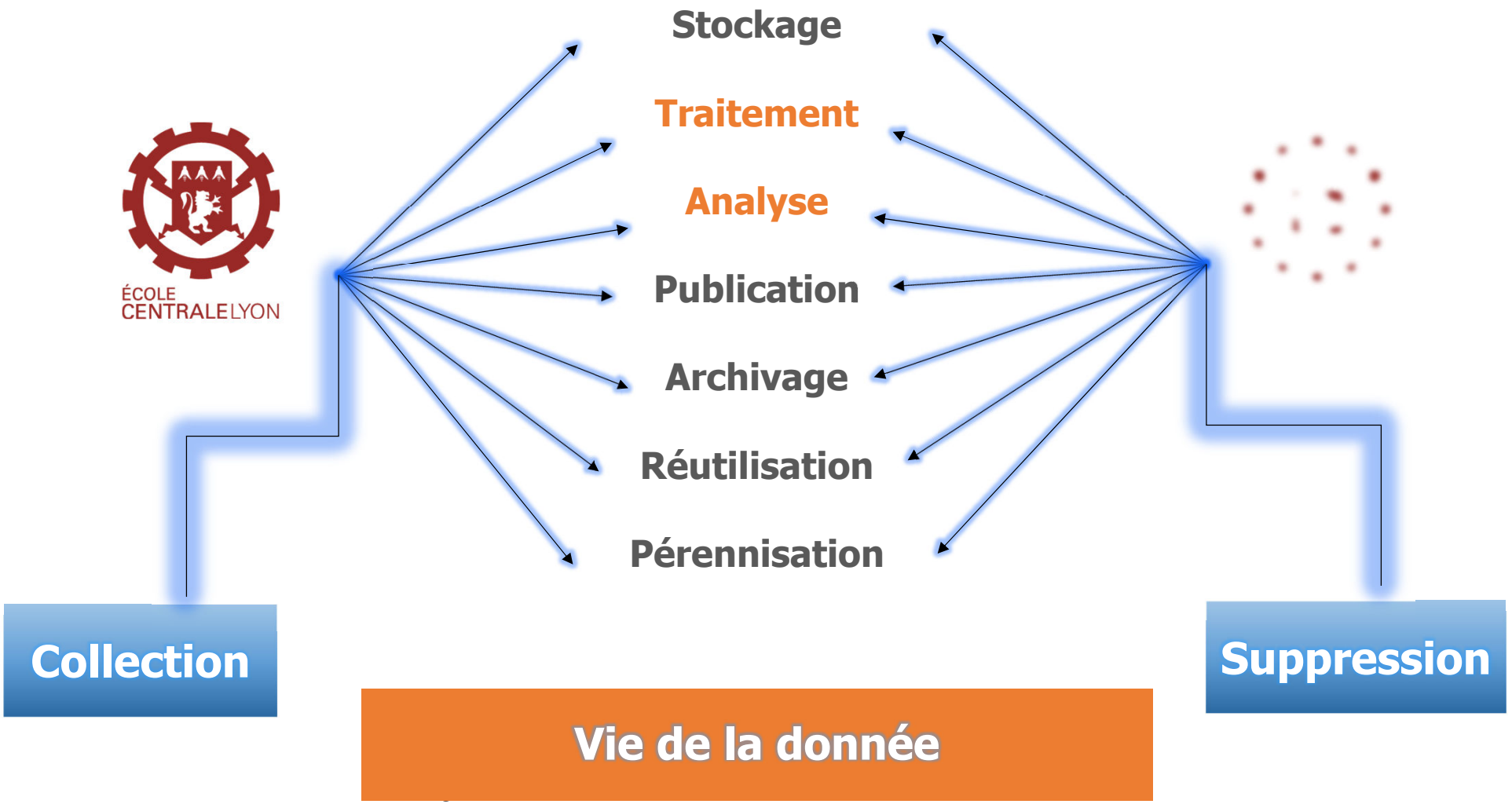
- Hébergement dans le datacentre ECL, au CRI22
- Un cluster de calcul parallèle, Newton (+3500 cœurs CPU, 8 GPU)
- Espaces de stockage partagé (+2 Po)
- Visualisation et Post-traitement des Données
- Ateliers et Séminaires, Informatiques et Calcul Scientifique



ÉCOLE CENTRALE LYON

Bienvenue au PMCS2I

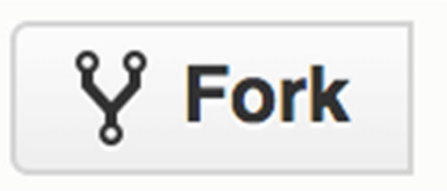
Cycle de vie des données



Avant de pouvoir être utilisée

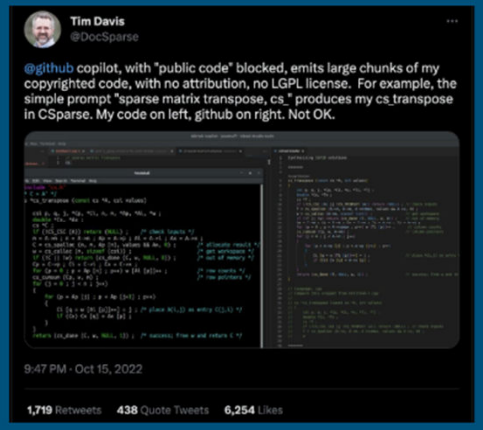
- Transcrite, traduite, vérifiée, validée & nettoyée
- Données personnelles identifiantes, anonymisées
- Code soient également accessibles
- L'analyse permet de décrire les données
- L'interprétation, donner sens à ces résultats
- Métadonnées persistantes

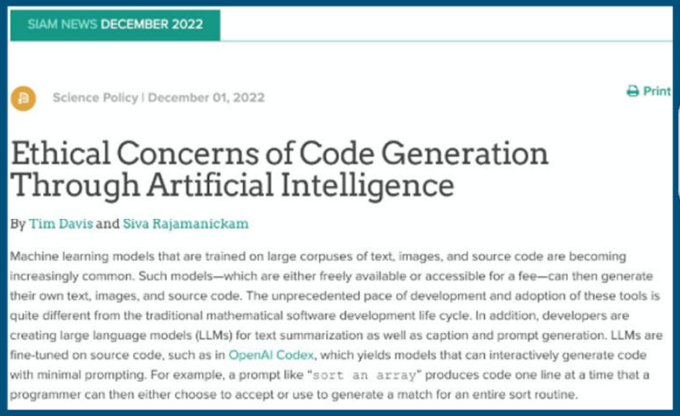
Duplication et Création de Code



- La génération des données
- L'analyse et traitement des données
- Duplication et modification des projets

I found my (L)GPL code in your dataset!





SIAM NEWS DECEMBER 2022

Science Policy | December 01, 2022 Print

Ethical Concerns of Code Generation Through Artificial Intelligence

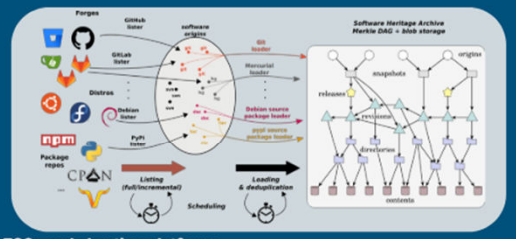
By Tim Davis and Siva Rajamanickam

Machine learning models that are trained on large corpuses of text, images, and source code are becoming increasingly common. Such models—which are either freely available or accessible for a fee—can then generate their own text, images, and source code. The unprecedented pace of development and adoption of these tools is quite different from the traditional mathematical software development life cycle. In addition, developers are creating large language models (LLMs) for text summarization as well as caption and prompt generation. LLMs are fine-tuned on source code, such as in OpenAI Codex, which yields models that can interactively generate code with minimal prompting. For example, a prompt like "`sort an array`" produces code one line at a time that a programmer can then either choose to accept or use to generate a match for an entire sort routine.

The source code opportunity

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE
Inria
unesco

Largest archive of source code
digital commons built since 2015



500+ code hosting platforms

All versions, full development history
In a single giant Merkle Graph

- 35×10^9 nodes
- 500×10^9 edges
- ~ 2 PB storage

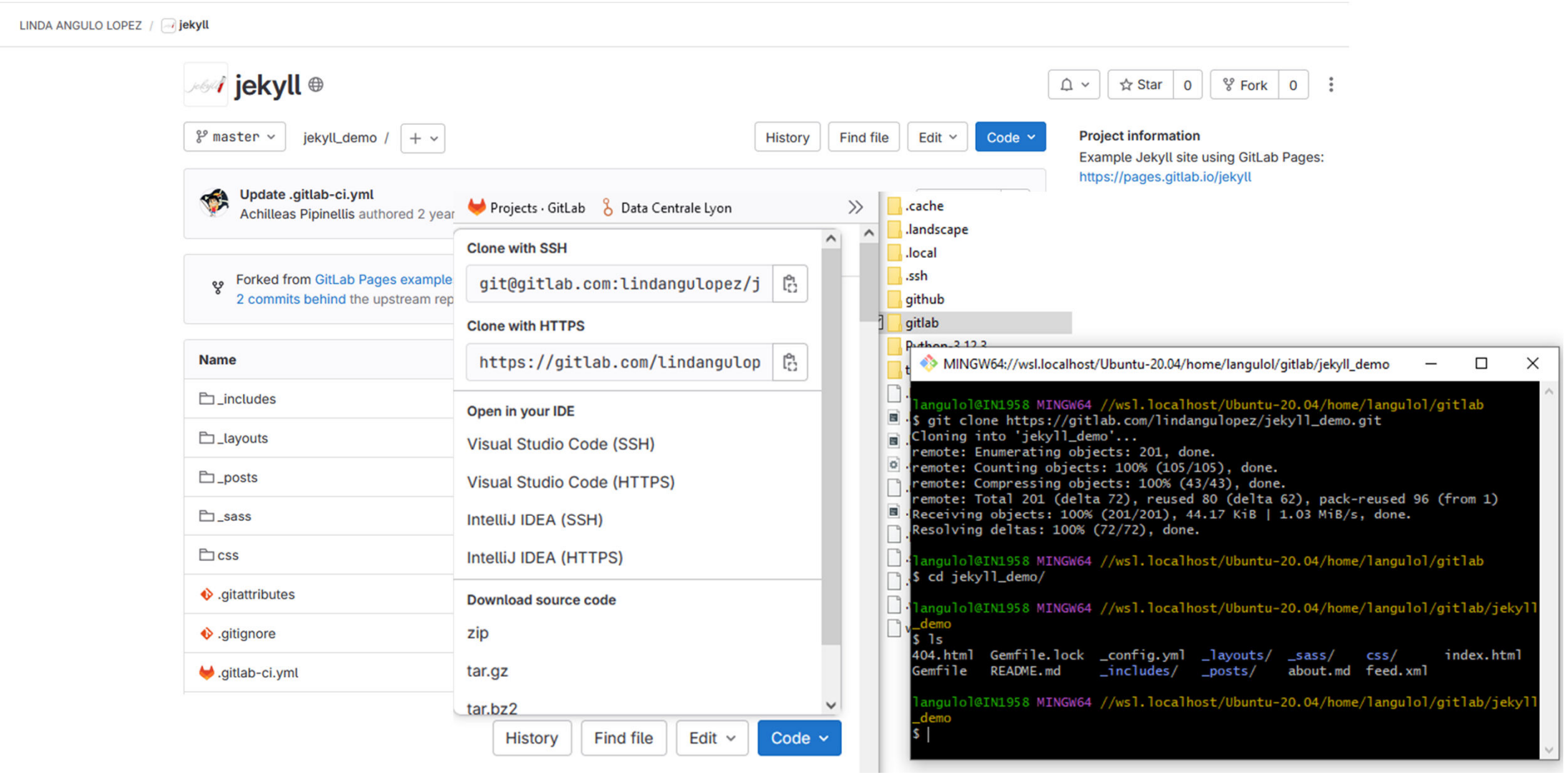
Ensures **availability**
Guarantees **integrity**
Enables **traceability**

of all source code

Unique dataset for machine learning,
an infrastructure for transparency and accountability

https://sinews.siam.org/Details-Page/ethical-concerns-of-code-generation-through-artificial-intelligence

Forkez et contribuez à un projet Git



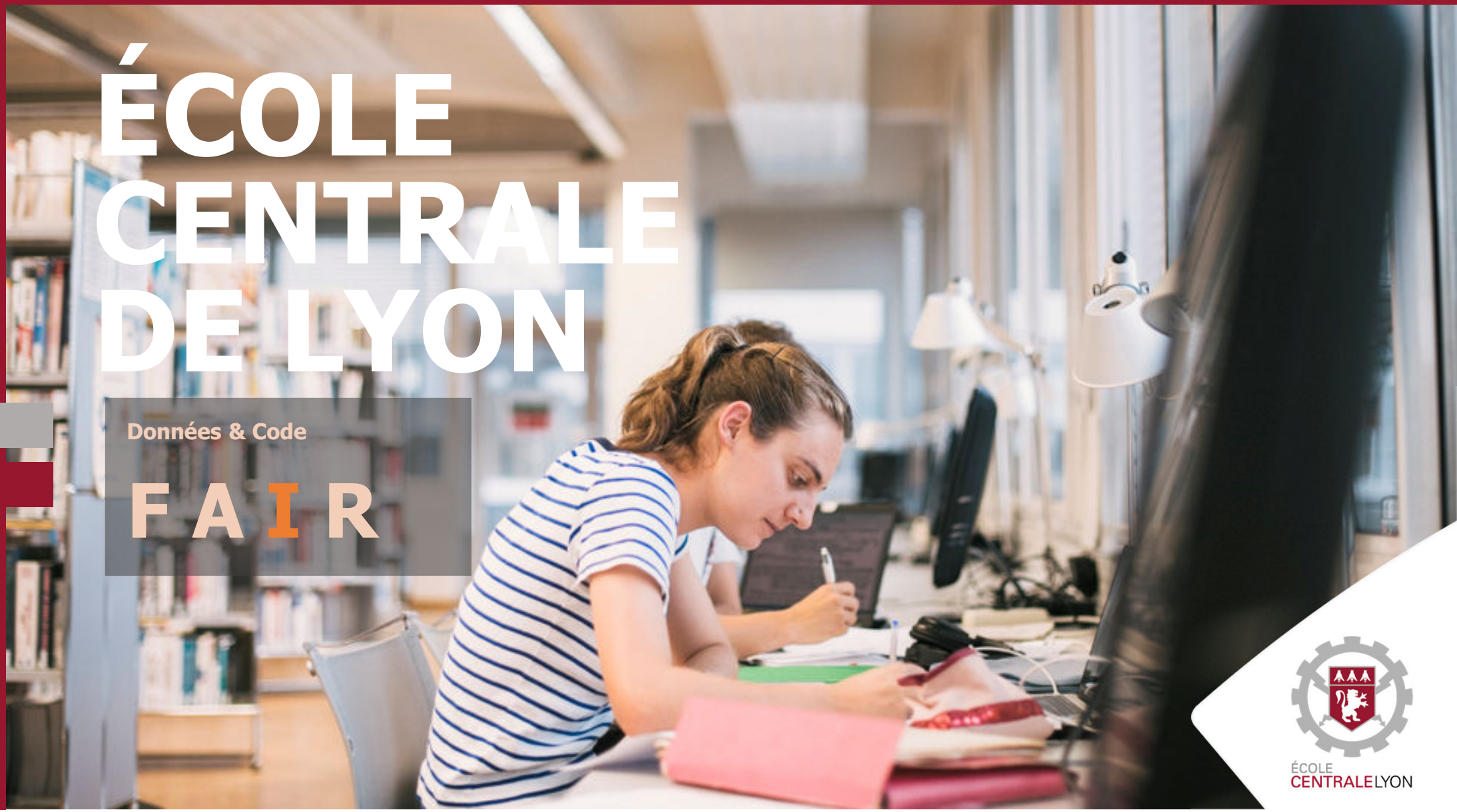
Duplication du projet, par <<Fork>>

- Création d'une branche thématique à partir de la branche master,
- validation de quelques améliorations (commit),
- poussée de la branche thématique sur votre projet Git (push),
- ouverture d'une requête de tirage sur Git (Pull request),
- discussion et éventuellement possibilité de nouvelles validations (commit).
- Le propriétaire du projet fusionne (merge) ou ferme (close) la requête de tirage.
- Synchronisation de la branche master mise à jour avec celle de votre propre dépôt.

ÉCOLE CENTRALE DE LYON

Données & Code

F A I R



ÉCOLE CENTRALELYON

Préparer les fichiers ouvert, auto documenté

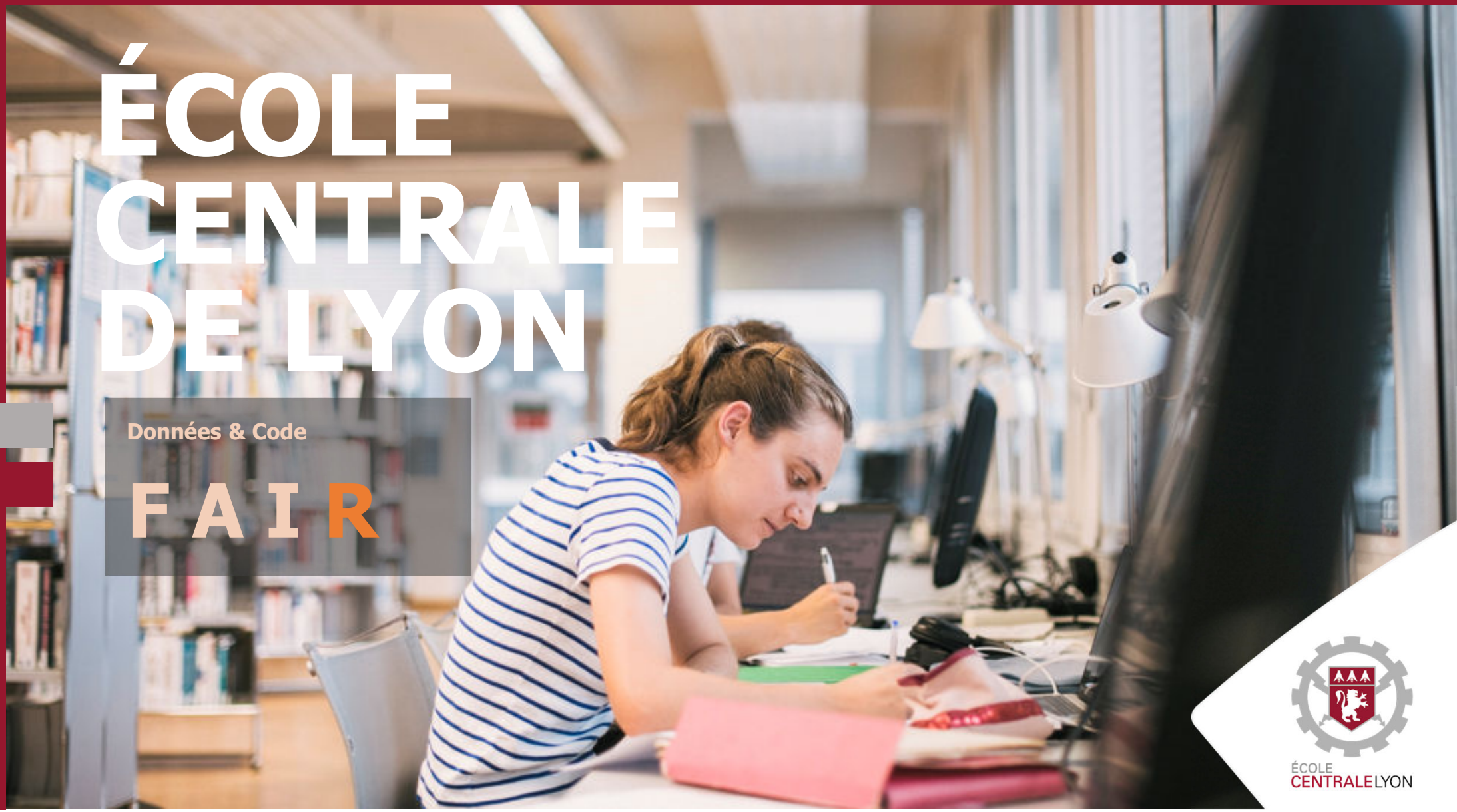
NetCDF

- Représenter et formater des données dimensionnées sous forme de tableaux
- intégrant les métadonnées directement dans l'entête du fichier

HDF5

- Type conteneur
- Structure de fichier hiérarchique
- Simuler des données grâce au calcul intensif
- compression et d'écriture/lecture parallèles

ÉCOLE CENTRALE DE LYON



Données & Code

F A I R

ÉCOLE
CENTRALELYON

Cycle de vie des données

Valoriser, promouvoir et partager

Publication d'articles scientifiques

- une petite partie des données ou métadonnées est accessible

Bases pluridisciplinaires, Recherche Data Gouv

- **Recommandations du financeur**
- **La possibilité de pérenniser l'accès**
- **Entrepôts certifiée, critères CoreTrustSeal**
- **Niveau de sécurité nécessaire**

Data-papers

- **la description détaillée des jeux, métadonnées**

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

GLOSSAIRE DONNEES DE LA RECHERCHE (1/2)

CHEATSHEET, DONNEES DE LA RECHERCHE (1/2)
Formation Doctorale Ecole Centrale de Lyon

Entrepôt de données	Espace numérique dans lequel on peut chercher ou déposer des données de recherche. Un entrepôt ne recueille que des jeux de données achevés, à la différence d'une plateforme de stockage qui doit être utilisée tant que l'on travaille encore sur ses données.
FAIR	**Findable Accessible Interoperable Reusable, ensemble de bonnes pratiques à suivre pour gérer et partager ses données au mieux.**
Horizon Europe	Programme européen pour la recherche et le développement pour 2021-2027. Le programme précédent s'intitulait « Horizon 2020 »
Inist	**Institut national de l'information scientifique et technique, unité du CNRS chargée de faciliter l'accès à l'information scientifique et aux données de la recherche ainsi que d'accompagner les chercheurs dans ces démarches.**
Loi CADA	Loi française de 1978 créant la Commission d'accès aux documents administratifs. Elle permet aux citoyens de demander l'accès aux documents administratifs rendus publics.
Loi pour une République numérique	**Loi française de 2016 mettant en place, entre autres, l'ouverture par défaut des données publiques, la publication en accès ouvert des articles de recherche financés au moins à 50% par des fonds publics (avec un embargo), l'ouverture des données de la recherche sous conditions.**
Métadonnées	Ensemble d'informations permettant de décrire une donnée. Par exemple, les métadonnées de données d'enquêtes, vont être le nom des personnes interrogées, leur âge, la date de l'entretien… Et l'enregistrement audio de l'entretien sera la donnée en elle-même.


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

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

ÉCOLE
CENTRALELYON