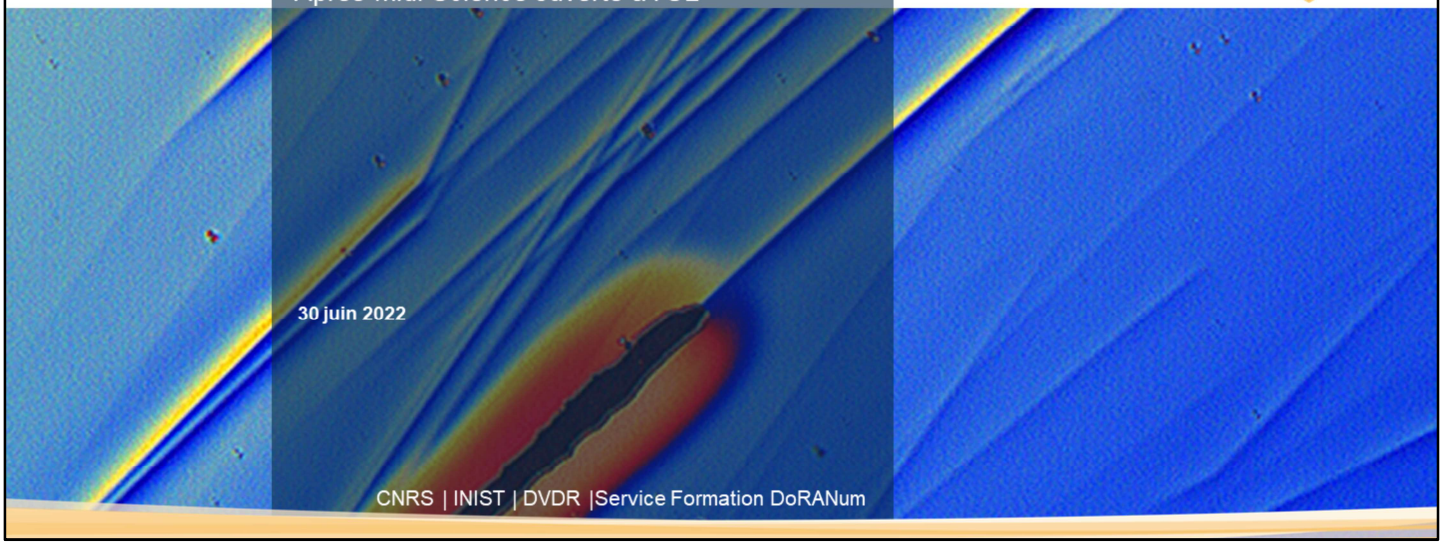
Le stockage des données

Bonnes pratiques

Après-midi Science ouverte à l'UL



30 juin 2022

CNRS | INIST | DVDR |Service Formation DoRANum

STOCKAGE DES DONNÉES – BONNES PRATIQUES
AU PROGRAMME

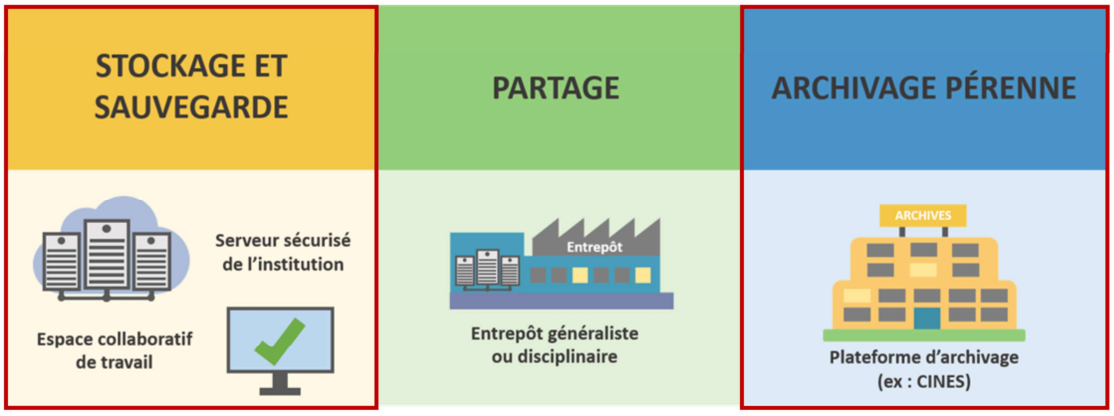
1. Stockage des données
2. Sauvegarde des données
3. Organisation des fichiers de données
4. Formats des données
5. Conservation des données
6. Ressources utiles

PRÉAMBULE

P 3

Source : NYU Health Sciences Library. Data Sharing and Management Snafu in 3 Short Acts. 19 décembre 2012. https://www.youtube.com/watch?v=N2zK3sAtr-4

FAITES-VOUS BIEN LA DISTINCTION ENTRE STOCKAGE ET ARCHIVAGE ?



P 4

Le **stockage** sécurisé et la sauvegarde des données se font durant le projet. L'objectif est de garantir la sécurité des données et d'en faciliter l'accès pour l'ensemble des collaborateurs du projet.
Plateforme de stockage : infrastructure proposant un stockage des données, avec des fonctionnalités de gestion des accès et éventuellement de sauvegarde intégrée.

L'**archivage** pérenne a pour objectif de conserver les données, d'en garantir l'accès et d'en préserver l'intelligibilité sur le long terme.
Nécessité d'assurer :

- la pérennisation des supports de stockage sur du long terme
- l'accès au contenu même quand les formats des données deviennent obsolètes
- l'intégrité des données.

Plateforme d'archivage : infrastructure qui intègre dans son fonctionnement tout le processus nécessaire à l'archivage des données.
En réalité, l'archivage pérenne concerne peu de données. Seulement celles qui présentent une grande valeur scientifique reconnue par la communauté dont elles proviennent. Soit parce qu'elles sont très couteuses, soit parce qu'elles sont uniques, non reproductibles.

STOCKAGE DES DONNÉES

- **Stockage** = enregistrement d'une information sur un support physique

- Ce support physique peut avoir des caractéristiques très variées
  - En fonction du matériel utilisé

  Ex : pour un ordinateur portable on peut choisir un disque SSD ou SATA
  - En fonction de la technologie utilisée pour accéder au support physique

  Ex : on ne peut pas ouvrir sous Windows un disque dur formaté sous Linux

cnrs

P 5

Un **SSD (Solid State Drive)** est un matériel informatique permettant le stockage de données sur de la mémoire flash (à la manière d'une clé USB).
Les SSD offrent des taux de chargement plus rapides pour les applications. Grâce à leur technologie, les SSD sont plus légers et plus à même de résister aux manipulations et aux chutes. Ils consomment moins d'énergie, ce qui limite le risque de surchauffe de l'ordinateur. Mais le SSD n'offre pas un grand espace de stockage, ce qui limite ses capacités de stockage.

Le **SATA (Serial Advanced Technology Attachement)** est un support informatique magnétique qui sert à stocker des données personnelles sur une mémoire morte qu'il intègre. Le disque dur SATA offre quelques avantages : le tout premier est sa vitesse d'accès aux données à stocker qui est d'environ 0,1 ms. Son deuxième avantage est sa vitesse de lecture et d'écriture qui est de 400 Mo/s.

On ne peut pas ouvrir sous Windows un disque dur qui a été formaté sous Linux car les technologies utilisées pour organiser le stockage ne sont pas les mêmes.

Il existe des espaces de stockage spécifiques pour les traitements intensifs de données permettant de faire des calculs efficacement.

COMPARATIF DES DIFFÉRENTS SUPPORTS DE STOCKAGE

Support de stockage	Sécurité	Accès	Coût	Remarques d'utilisation
Ordinateur professionnel	Sujet au vol, au piratage, aux détériorations et pannes ★★	Non adapté au partage Nécessite l'utilisation d'un autre support de sauvegarde ★	Pas de coût supplémentaire ou coût peu important ★★★★	Pour un stockage temporaire Nécessité de chiffrer les données confidentielles et sensibles
Clé USB ou disque dur externe	Sujet au vol, à la perte Durée de vie limitée ★	Facilement transportable Permet de transférer les données vers un autre ordinateur ★★★	Pas de coût supplémentaire ou coût peu important ★★★★	Pour un stockage temporaire Nécessité de chiffrer/sécuriser les données confidentielles et sensibles
Serveur institutionnel	Stockage fiable, durable et sécurisé ★★★★	Ne facilite pas le travail avec des collaborateurs extérieurs ★★	Coût assez important pas forcément répercuté sur l'usager ★★	Pour un stockage plus pérenne Adapté au stockage de données sensibles Toutes les institutions ne proposent pas ce service
Serveur cloud institutionnel	Stockage fiable, durable et sécurisé ★★	Accès sécurisé Permet un travail synchronisé avec tous les collaborateurs autorisés au partage ★★★★	Payant à partir d'un certain seuil de stockage ★★	Pour un partage avec des collaborateurs externes Toutes les institutions ne proposent pas ce service

cnrs

P 6

Source : DoRANum. Stocker ses données de façon sécurisée. 4 juillet 2018. https://doranum.fr/stockage-archivage/stockage-donnees/

QUEL EST LE VOLUME DES DONNÉES ?

Accessible Reusable

- 1 Po = 1 000 To ; 1 To = 1 000 Go ; 1 Go = 1 000 Mo
  - Capacité disque dur externe « moyen » : 2 To
  - Plateforme de stockage de site : quelques Po
  - Fichier texte < fichier bureautique ~ fichier audio ~ fichier image < fichier vidéo
- **Gros volumes** : quelques To à quelques dizaines de To
- Problématiques d'un volume important :
  - transfert des données
  - chargement en mémoire pour traitement / analyse / calcul
  - capacité de stockage nécessaire et coût
  - politique de sauvegarde ...

P 7

Gros volumes : quelques To à quelques dizaines de To (temps de transfert ou temps de chargement devenant prohibitif).

SÉCURISATION DU STOCKAGE DES DONNÉES

- Au niveau d'un **serveur de stockage** :
  - Répartition des données sur plusieurs disques d'un même serveur afin d'améliorer la tolérance aux pannes lorsqu'un des disques a un problème

- Au niveau d'une **plateforme de stockage** (plusieurs serveurs) :
  - Intégration de sauvegardes ou de synchronisations permettant de sécuriser les données stockées lorsque l'un des serveurs a un problème ou devient inaccessible

SÉCURISATION DE L'ACCÈS AUX DONNÉES

Accessible Reusable

- **Gestion des droits d'accès** : toutes les solutions de stockage offrent une gestion des droits d'accès plus ou moins fine, plus ou moins sécurisée

- **Chiffrement du stockage** : pour augmenter la sécurisation des accès aux données, il est possible de chiffrer les fichiers

  Ils ne seront alors lisible qu'avec la clé de déchiffrement

  Attention à ne pas perdre la clé de déchiffrement

cnrs

P 9

Gestion des droits d'accès : toutes les solutions de stockage offrent une gestion des droits d'accès plus ou moins fine, plus ou moins sécurisée.

- Depuis un accès ouvert via une URL (stockage de fichiers sur le cloud) jusqu'à la gestion de droits Unix pour des espaces de stockage associés à des machines de calcul par exemple.

Chiffrement du stockage : on peut chiffrer des fichiers, un répertoire ou un disque tout entier.
Attention à **ne pas perdre la clé de déchiffrement** car il serait alors impossible de récupérer les données !

COÛT DU STOCKAGE DES DONNÉES

- **Coût financier**
  - Possibilité d'optimiser en partageant une plateforme et les coûts humains associés
  - Plus le stockage est performant (accès rapide aux données) plus il est cher

- **Coût environnemental**
  - Coût des émissions de CO2

P 10

SAUVEGARDE DES DONNÉES

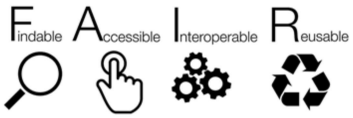
- **Sauvegarde** = duplication des données sur des supports de stockage différents, localisés dans des endroits différents
- Dans l'idéal, dupliquer et stocker les données à différents endroits sur différents supports

Règle du 3-2-1 :

- garder 3 exemplaires des données
- sur 2 supports ou technologies différents
- dont 1 se trouve hors site

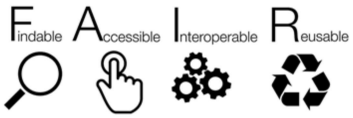
- Organiser et planifier ces sauvegardes
- Définir l'hébergement
- Gérer les versions

ARBORESCENCE DES DOSSIERS



- **Pas de règle** pour l'**organisation des dossiers** et l'**ordre des éléments** dans les noms de fichiers
- La **meilleure règle** est celle qui **convient à l'équipe**, **aux partenaires du projet** et **aux utilisateurs**, en général en plaçant l'**élément le plus important en premier**
- Il est important de **bien documenter** la construction des noms de répertoires et de fichiers, ainsi que le nommage des versions

ARBORESCENCE DES DOSSIERS

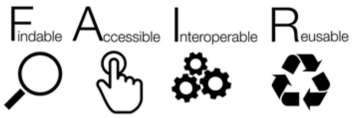


- Trouver l'équilibre dans la **profondeur de l'arborescence**
  - trop profonde → trop de clics pour atteindre le bon fichier
  - peu profonde → trop de fichiers pourraient se retrouver dans un seul dossier (les organiser en sous-dossiers)

- On peut **organiser les dossiers** par :
  - projet, expérience,...
  - date (année, mois, jour)
  - type de données (par exemple : textes, scripts, graphiques,...)

P 13

ARBORESCENCE DES DOSSIERS



- **Numéroter** éventuellement les dossiers
- **Séparer les données brutes des données analysées/traitées**
- Ne pas hésiter à mettre un **fichier Readme.txt** (ou Lisezmoi.txt) dans chaque répertoire pour expliquer le type de données qui s'y trouve

P 14

Les **fichiers Readme.txt** peuvent être utilisés pour décrire des projets, des dossiers et des fichiers.

NOMMAGE DES DOSSIERS ET FICHIERS

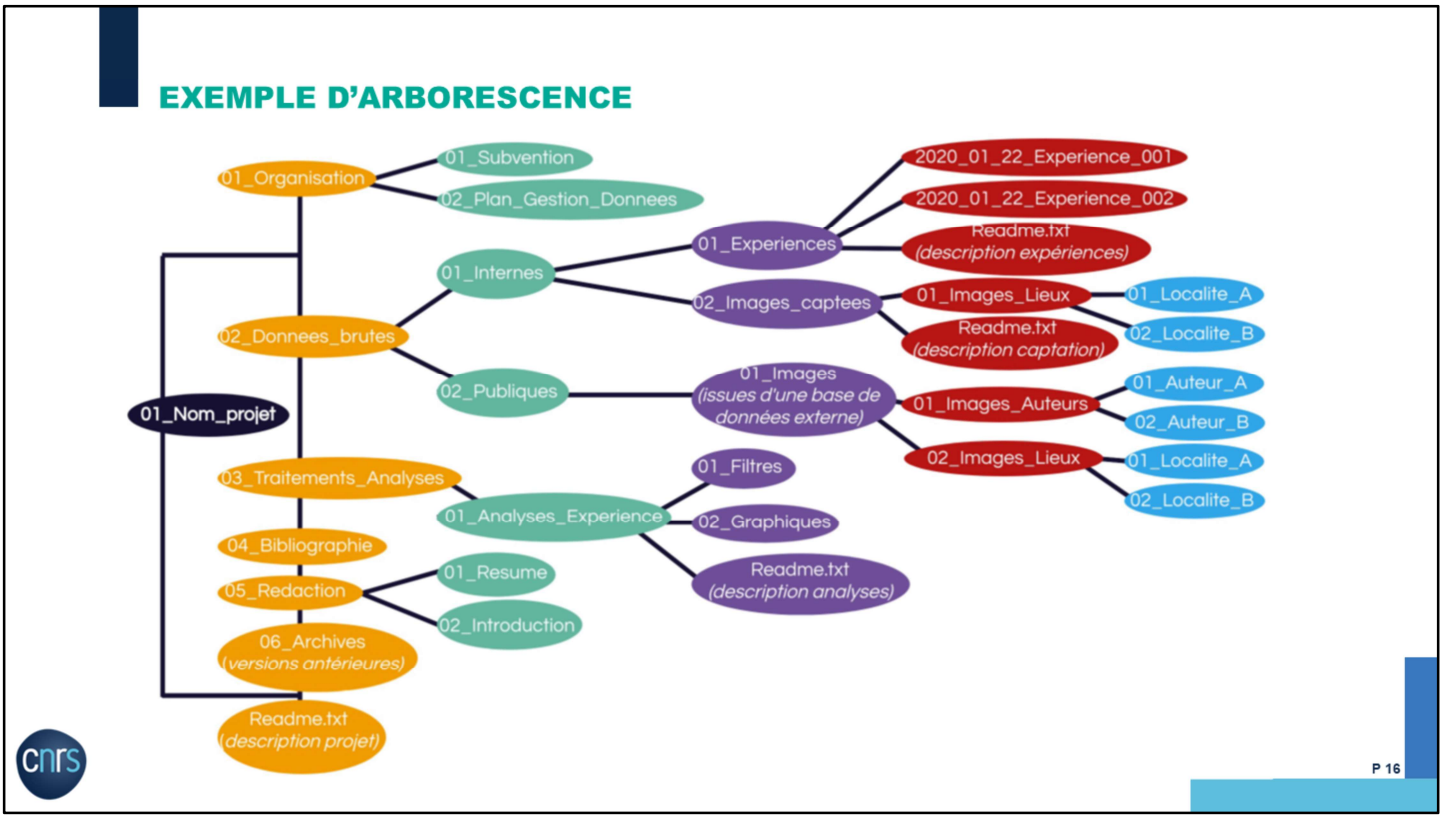
Important de **pouvoir identifier** immédiatement **le contenu des fichiers**

- Utiliser des lettres (majuscules et minuscules), des chiffres
- Utiliser des noms significatifs, des abréviations explicites
- Ecrire les chiffres avec un nombre de caractères significatifs
  (ex : séquence de 1 à 10 → 01-10 ; de 1 à 100 → 001-100)
- 30 à 40 caractères maximum
- Pas d'espaces, de points ou de caractères spéciaux (£"$%!”&*^()+=[]:{}~@)
- Ajouter traits d'union (-) et underscores (_) pour séparer les éléments

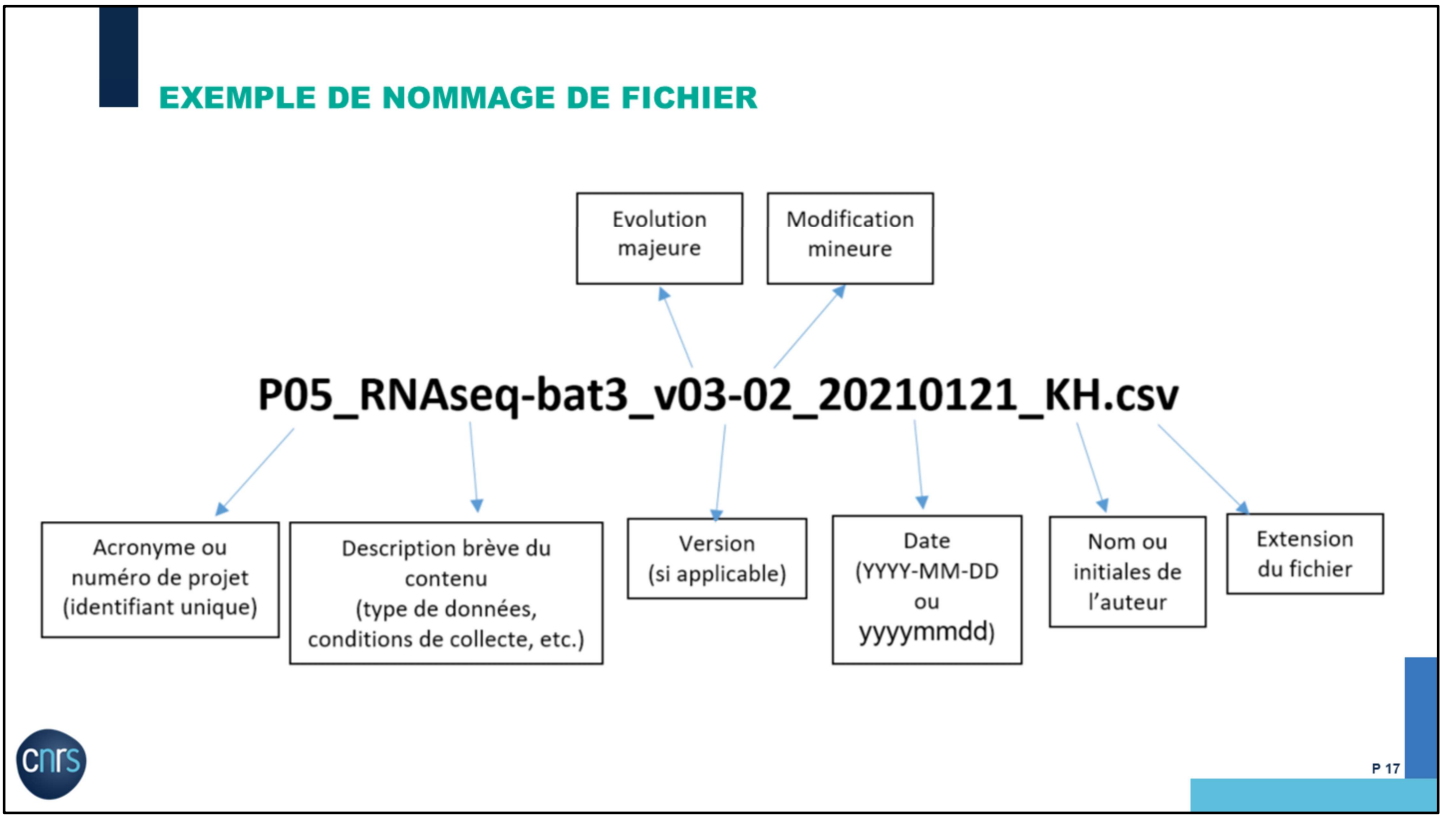
cnrs

P 15

Source : traduit et inspiré de l’infographie de Kira Höffler (https://twitter.com/KiraHoeffler/status/1367804034413920259/photo/1) et de l'intervention de Violaine Louvet lors du séminaire sur le stockage des données de la recherche (https://doranum.fr/stockage-archivage/seminaire-stockage-des-donnees-de-la-recherche/)



Source : traduit et inspiré de l'infographie de Kira Höffler (https://twitter.com/KiraHoeffler/status/1367804034413920259/photo/1) et de l'intervention de Violaine Louvet lors du séminaire sur le stockage des données de la recherche (https://doranum.fr/stockage-archivage/seminaire-stockage-des-donnees-de-la-recherche/)



Source : traduit et inspiré de l'infographie de Kira Höffler (https://twitter.com/KiraHoeffler/status/1367804034413920259/photo/1)

FORMATS DES DONNÉES

- **Format** = manière dont les données sont **structurées**, **organisées** et **encodées** sur le support physique
  - En général, défini par l'extension dans le nom du fichier

Pouvoir

- ouvrir les fichiers avec un maximum de logiciels
- le faire maintenant et dans 5, 10, 20 ans
- partager ses données avec son équipe / ses partenaires / sa communauté
- traiter et analyser les données grâce à différents logiciels
- Croiser des données de différentes sources (fichiers, formats distincts)

Le **format ouvert** est interopérable et indépendant du logiciel utilisé pour le créer, le modifier, le lire et l'imprimer

- Le **format fermé** est dépendant du logiciel développé par l'entreprise
- Ne pas se faire piéger par un format fermé qui contraindra la liberté de choix quant aux logiciels (libres ou propriétaires) qu'on voudra utiliser

Ressource supplémentaire : https://doranum.fr/stockage-archivage/quiz-format-ouvert-ou-ferme/

CONSERVATION DES DONNÉES

- **Court terme** : pendant le projet de recherche
  - Si les données sont volumineuses, évaluer le coût de re-création vs le coût de stockage

- **Moyen terme et long terme** : se poser les bonnes questions
  - Quelles sont les données nécessaires pour assurer la reproductibilité
  - Combien de temps en ai-je besoin ?
  - Il est important d'identifier les données :
    - de test
    - de validation, associées à une publication
    - difficilement reproductibles ou réutilisables dans d'autres domaines
    - non reproductibles ou de forte valeur

cnrs

P 20

Moyen terme et long terme : se poser les bonnes questions
Quelles sont les données nécessaires pour assurer la reproductibilité des résultats de mon projet ?
Combien de temps en ai-je besoin ?
Il est important d'identifier les données :

- de test → à supprimer en fin de projet
- de validation, associées à une publication → à garder jusqu'à 5 à 10 ans après publication
- difficilement **reproductibles** ou réutilisables dans d'autres domaines → à garder plus de 10 ans après publication
- non reproductibles ou de forte valeur (par exemple historique) → à archiver

RESSOURCES UTILES

- DoRANum – Stockage et archivage : https://doranum.fr/stockage-archivage/
- Séminaire Stockage des données de la recherche – Intervention de Violaine Louvet sur l'organisation des données - https://doranum.fr/stockage-archivage/seminaire-stockage-des-donnees-de-la-recherche/
- Science ouverte à l'Université de Lorraine - https://scienceouverte.univ-lorraine.fr/

Merci de votre attention

contact-formation@inist.fr

www.cnrs.fr