CC BY SA

Module 2 Séquence 4

Stockage et accès

Fred de Lamotte - Montpellier
https://orcid.org/0000-0003-4234-1172

Julien Seiler - Strasbourg
@julozi

Stockage et accès

Stockage des données

Fonction fondamentale : **la conservation des données**

Stockage :

- désigne des méthodes et des technologies permettant de conserver des données
- concerne tous les types de supports de stockage de masse (DD, Clé USB…) ou support de stockage dématérialisé (cloud)
- intègre des problématiques d'usage collaboratif : dépôt, partage.

Critères de sélection pour choisir un support de stockage :

- la fréquence d'utilisation des données,
- les besoins en capacité de stockage (taille),
- la sécurité des données,
- la vitesse d'accès à la donnée
- la fiabilité et le coût du support

Un environnement de travail sûr

Comprendre l'environnement de travail que vous utilisez avant de démarrer votre projet :

Votre poste de travail :

- Y'a-t-il des sauvegardes (stratégie 3-2-1) ?
  - **3** copies sur au moins **2** systèmes différents dont au moins **1** est distant = **0** inquiétude
    Par exemple : stockage en RAID (copie locale) + sauvegarde sur un disque externe qui reste au labo
- Votre environnement est-il mis à jour régulièrement ?
- Disposez-vous d'un antivirus (à jour) ?
- Vos données sont-elles chiffrés (en cas de vol) ?

Vos solutions de stockage :

- Y'a-t-il des sauvegardes (stratégie 3-2-1) ?
- Est-ce que la pérennité est en phase avec vos besoins ?
- L'environnement est-il mis à jour régulièrement ?

Un environnement de travail sûr

Vos mots de passes (au pluriel)

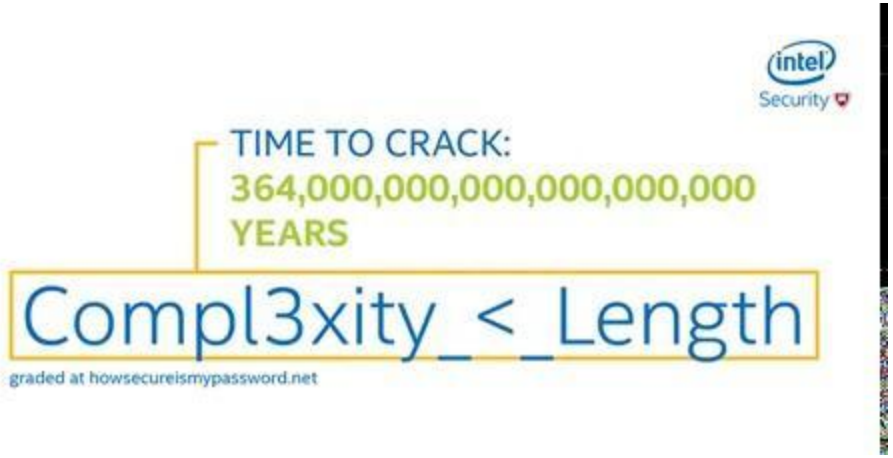
- Utilisez-vous des mots de passe robustes ?

Type de mot de passe	Taille de clé équivalente	Force	Commentaire
Mot de passe de **8 caractères** dans un alphabet de **70 symboles**	49	Très faible	Taille usuelle
Mot de passe de **10 caractères** dans un alphabet de **90 symboles**	65	Faible	
Mot de passe de **12 caractères** dans un alphabet de **90 symboles**	78	Faible	Taille minimale recommandée par l'ANSSI pour des mots de passe ergonomiques ou utilisés de façon locale.
Mot de passe de **16 caractères** dans un alphabet de **36 symboles**	82	Moyen	Taille recommandée par l'ANSSI pour des mots de passe plus sûrs.
Mot de passe de **16 caractères** dans un alphabet de **90 symboles**	104	Fort	
Mot de passe de **20 caractères** dans un alphabet de **90 symboles**	130	Fort	Force équivalente à la plus petite taille de clé de l'algorithme de chiffrement standard AES (128 bits).

Exemple : N,cn'eplr.2lMcb! (16 caractères, alphabet de 90 symboles)

Un environnement de travail sûr

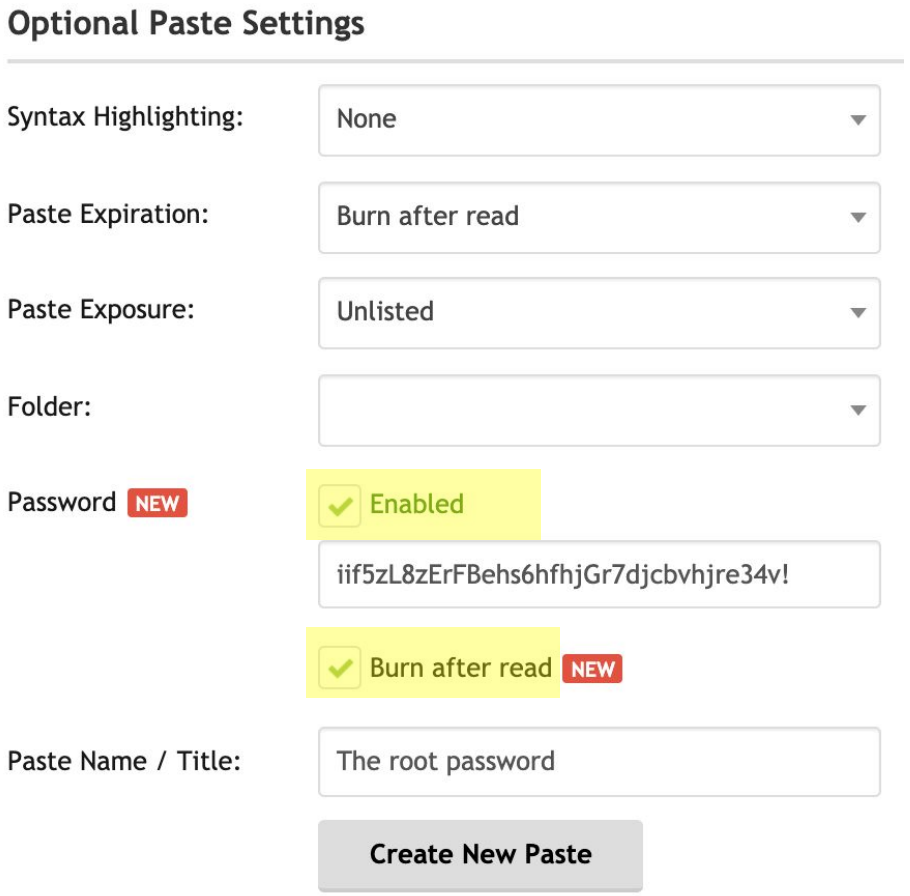
Vos mots de passes (au pluriel)



Un environnement de travail sûr

Vos mots de passes (au pluriel)

- Utilisez-vous un mot de passe différent pour chaque fournisseur de service ?
- Utilisez-vous un gestionnaire de mot de passe ?
- Renouvelez-vous vos mots de passe régulièrement ?
- Utilisez-vous une procédure sécurisé pour communiquer un mot de passe à vos collègues ? (par exemple pastebin.com)



Stocker et sécuriser : quels compromis ?

Comparatif de systèmes de stockage des données

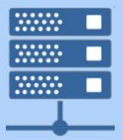
Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 Ordinateur professionnel	★★☆☆ Sujet au piratage informatique, aux détériorations et pannes	★☆☆☆ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud…)	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter les données confidentielles et sensibles
 Support externe	★☆☆☆ - Sujet au vol, à la perte du support - Durée de vie limitée (dégradation du matériel)	★★★☆ Facilement transportable, il permet de transférer les données vers un autre ordinateur	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter ou de sécuriser physiquement les données confidentielles et sensibles
 Serveur institutionnel	★★★★ Stockage fiable, durable et sécurisé (contre le vol, le piratage, les incendies…)	★★☆☆ La connexion au serveur institutionnel ne facilite pas le travail avec des personnes extérieures	★★☆☆ Coût assez important mais pas forcément répercuté sur l'usager	- Pour un stockage plus pérenne - Adapté pour le stockage de données sensibles et des versions « stables » de vos données - Toutes les institutions ne proposent pas ce service
 Serveur Cloud	★★☆☆ On ne sait pas vraiment où sont stockées les données, ni ce qu'elles deviennent	★★★★ Permet un travail synchronisé avec toutes les personnes ayant été autorisées au partage	★★☆☆ Payant à partir d'une certaine limite de stockage	- Pour un partage avec des personnes externes à l'institution - Ne pas y mettre de données sensibles ou confidentielles - Pas de contrôle sur la procédure de sauvegarde des données

Tableau tiré de http://doranum.fr/le-stockage-des-donnees/

Transfert de vos données de recherche

Comment transmettre vos données ?

Pas bien → Bien

Messagerie instantanée

Email

- Pas conçu pour le transfert de données
- Les communications peuvent être interceptées
- Localisation du stockage et durée de rétention inconnues

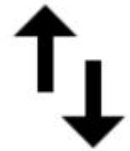
Envoi d'un disque

Dropbox, Drive, etc

- Risque de perte
- Risque d'accès non autorisés
- Acceptable si les données sont chiffrées

Cloud privé

Service d'un consortium



- Optimisé pour le transfert de données scientifiques
- Sécurisé
- Support gratuit

Transfert de vos données de recherche

Connaissez-vous ces outils ?

Petits jeux de données :

- scp
- rsync
- (s)ftp
- wget (https)

Gros jeux de données :

- Globus
- bbcp
- fdt

La vitesse de transfert dépend de :

- L'outil utilisé
- L'infrastructure source et destination
- Le réseau
- La granularité des données

N'oubliez pas le chiffrement !!!!

Entretien des données

Organisation

- Définissez une politique d’organisation de vos données pour chaque projet
- Documentez et diffusez votre politique au sein de l’équipe
- La cohérence prime sur la préférence personnelle

Entretien des données

Organisation des dossiers

- Limitez le nombre de dossiers par niveau (5 ou 6 max)
- Allez du général au spécifique
- Choisissez des noms de dossiers explicites

Pour chaque dossier, ajoutez un fichier README:

- Choisissez un format simple et ouvert (par exemple Markdown ou TXT)
- Indiquez un minimum de méta-données concernant le dossier et son contenu :
  - Titre
  - Date de création / réception des données
  - Origine/Source des données
  - Version
  - Propriétaire/responsable des données
  - Organisation des données
  - Méthode de réception/téléchargement des données

Entretien des données

Organisation des dossiers

Exemple :

Un dossier par projet

Un sous-dossier par type de manip (microscopie, séquençage, phénotypage)

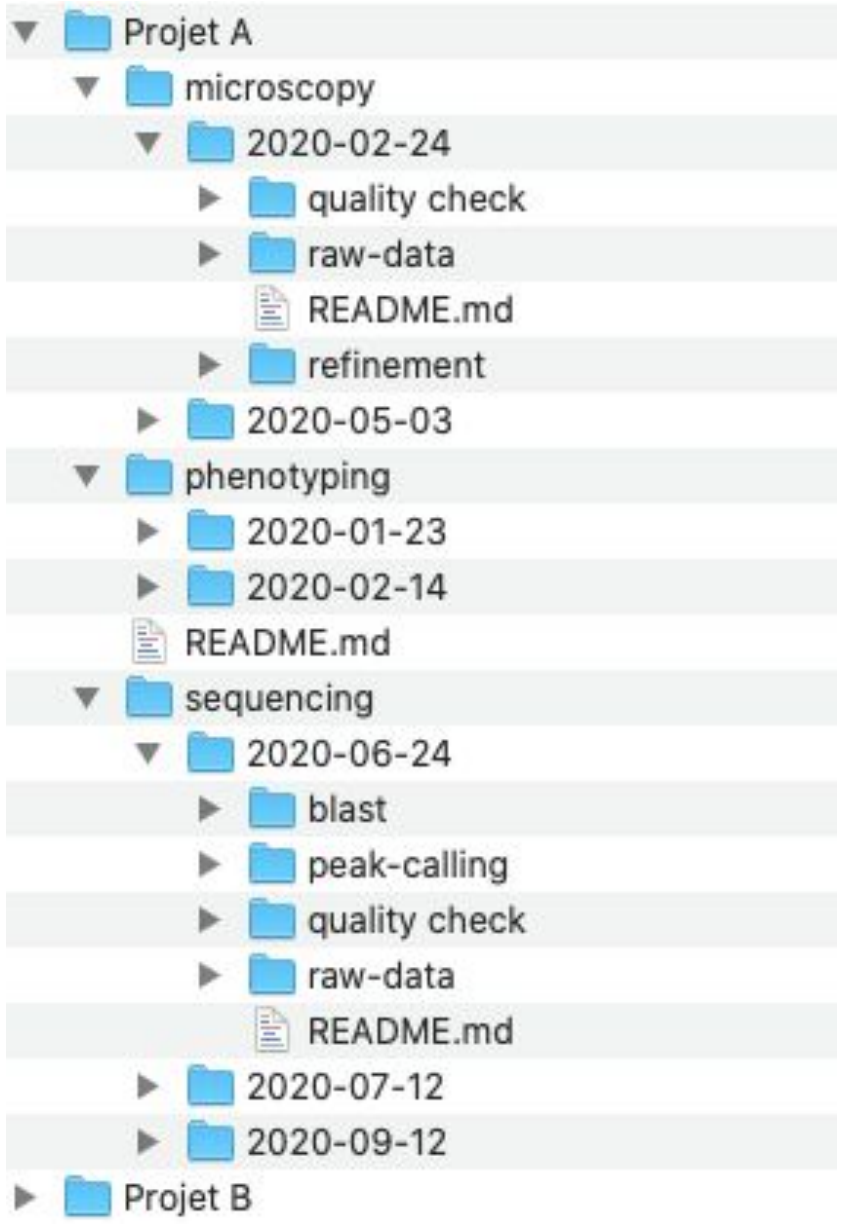
Un sous-dossier par date (2020-02-24, 2020-05-03)

Un sous-dossier pour les données brutes

Un sous-dossier par analyse (contrôle qualité, nettoyage statistique, raffinement)

Un sous-dossier par publication

Un lien symbolique vers chaque dossier données ou analyse associé à la publication



Intégrité des données

Identifier et contrôler la corruption des données

- Corruption : introduction de modifications non intentionnelles des données

Les données peuvent être corrompues par :
- des modifications non souhaités (ransomware, collègue…)
- un transfert de données défectueux
- un plantage d’un disque dur
- ...

Intégrité des données

Identifier et contrôler la corruption des données

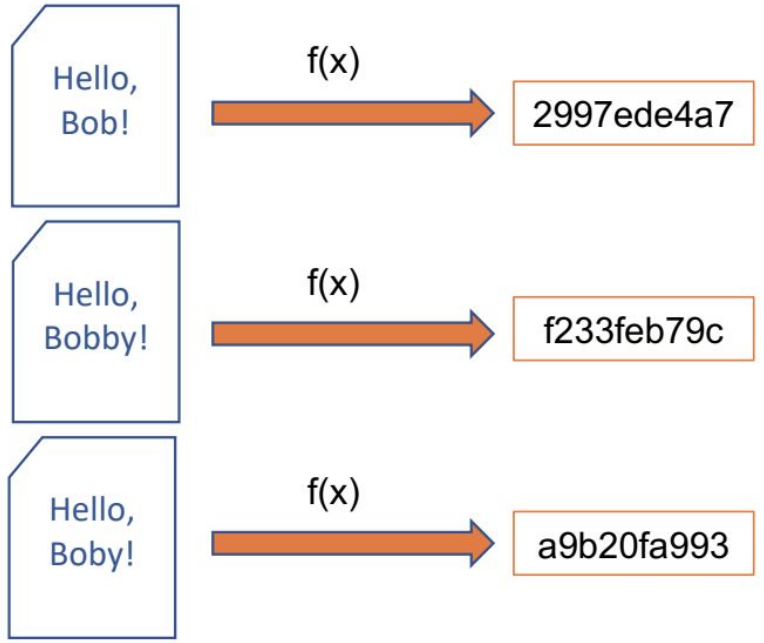
Solution 1 : générer des sommes de contrôles

Comment ?

- Linux / macOS : md5sum, sha256sum
- Windows : certutil

Quand ?

- Avant un transfert de données
  - Lorsqu'on réceptionne un nouveau jeu de données d'un collaborateur
  - Lorsqu'on transfert des données sur un stockage distant
- Stockage à long terme
  - La version principale de chaque dataset
  - Les extraits de données utilisés dans les publications



Intégrité des données

Identifier et contrôler la corruption des données

Solution 2 : utilisez le contrôle d'accès

N'accordez que les permissions d'accès nécessaire :

- Limitez le nombre d'utilisateurs ayant accès à vos données
- Limitez la visibilité des données (réseau interne vs internet)
- N'utilisez jamais de partage public sans chiffrement des données !

Mettez les données brutes en lecture seule

L'accès aux données sensibles doit être documenté

Copie des données

Limitez les copies au minimum !

- Copie principale (master)
  - Egalement appelé donnée “source” ou “brute”
  - Stratégie 3-2-1
- Copie de travail
  - A éviter au maximum
  - Utilisez des liens symbolique vers la copie principale
- Copie de sauvegarde
  - Ne travaillez jamais sur votre copie de sauvegarde

Suppression des données

Est-ce que ces données peuvent être supprimés ?

Le stockage des données a un coût financier et écologique.

- Distinguez clairement la copie principale (master) de ses dérivés
- Organisez régulièrement une revue des données
- Récupérer rapidement les données sur supports externes (disque ou clé USB)

Conservation des données

“Je veux garder mes données pour l’éternité”

Ne manquez pas le module 4...

- Quels sont vos obligations en terme de rétention de données
- Dans quelles conditions allez-vous les archiver ?
- Avez-vous documenter clairement vos données ?
- Que se passera-t-il si vous partez (pour l’éternité) ?

Les infrastructures de stockage sont vos amies

- Politique de sauvegarde professionnel et cohérente
- Nombre de copies minimum (stratégie 3-2-1)
- Gestion claires des droits d’accès
- Haute disponibilité et accessibilités
- Sécurité

Essayons de nous améliorer

Où se situe mon fichier ?

Lisible par

Moi Mon équipe | Ma communauté | D'autres communautés | Le monde entier

Format

Propriétaire fermé | Propriétaire ouvert | Ouvert

Format

En évolution | Stable

Description

Pas de schema(.org) | schema

Langage du format

Propriétaire | Norme

Formalisation

Pas de norme | Norme propriétaire | Norme Iso

Exercice

- Télécharger la matrice Excel **modèle radar.xlsx** sur osf.io
- Donnez une note de 0 à 5 pour chaque critère pour votre fichier