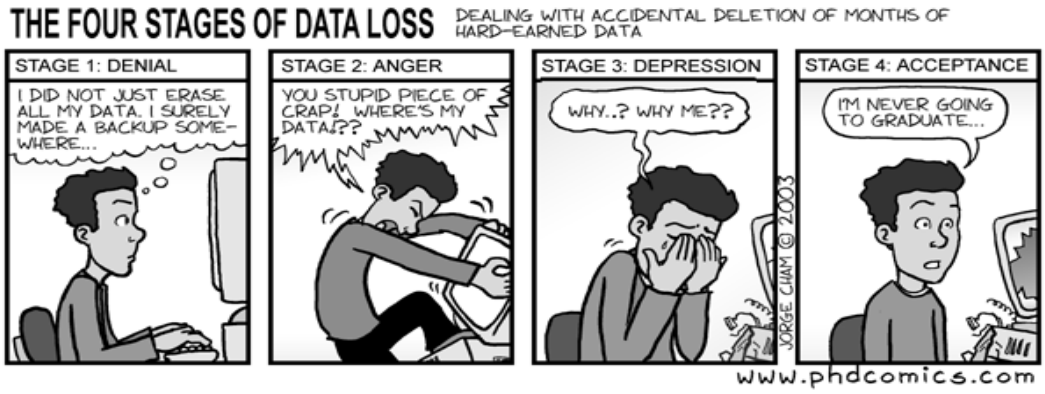
Cellule données de la recherche
donnees-recherche@agroparistech.fr
V3 – Oct 2023

Les bonnes pratiques de gestion des fichiers et des données.

Comment gagner du temps et éviter les catastrophes !



(copyright phdcomics)

I. Pourquoi adopter ces bonnes pratiques ?

Bien gérer ses données, de la conception d'un projet à l'ouverture de ces données, permet de gagner du temps en sachant où trouver les données à tout moment et en évitant au maximum les erreurs, les confusions, les redondances, les versions multiples et les catastrophes comme des pertes définitives de données par exemple !

Des données bien organisées, bien documentées et aisément traçables permettent d'attester de la fiabilité des résultats et de produire des données réutilisables par soi et par les autres.

L'organisation et la gestion des données d'un projet répondent à trois besoins fondamentaux :

- **Savoir où les trouver** : gestion des fichiers, stockage, sauvegardes
- **Savoir les lire et les relire** : choix des formats et des logiciels
- **Savoir les comprendre** : documentation, métadonnées, dictionnaire de données

Dans cette optique, il faut donc, autant que possible, anticiper quels types de données vont être produits, quel volume et dans quels formats ainsi que ce qu'il faudra stocker puis conserver à terme et pour quelle durée (par exemple, il vaut mieux prioriser les données chères, non-reproductibles, longues et difficiles ou délicates à produire (expérimentation animale par exemple), à forte valeur patrimoniale plutôt que les données issues de modélisation, simulation sauf si les processus sont longs).

Il est important également de définir les paramètres de l'utilisation de ces données durant le projet :

- **Qui va les utiliser :** Une personne ? Plusieurs ? Dans un même lieu ? Dans des lieux différents?
- **Comment :** Données requises à quelle fréquence ? A quelle vitesse ?
- **Depuis quel lieu de stockage :** local (disques durs, serveurs, etc.), sous-traité (Data Centers), entrepôt… selon le volume, les infrastructures et les moyens (logiciels, espaces de stockage, compétences) dont on dispose mais également le besoin de sécurisation et de sauvegarde des données (contrôle d'accès, cryptage, baie de sauvegarde, etc…)

Pour toutes ces considérations, il est fondamental de considérer ces points dès le montage puis le lancement des projets. Il faut se mettre d'accord au sein de l'équipe : définir qui fait quoi, les référents

pour chaque tâche, choisir une langue (FR/EN) et adopter une organisation efficace, claire (bien documentée) mais pas trop lourde à gérer et utiliser. Il faut faire au mieux mais rester réaliste !

La toute première étape consiste donc à remplir son Plan de Gestion de Données ! ☺

II. Organiser et Gérer ses fichiers

1) Choisir les types de fichiers

Pour éviter de se retrouver bloqué par des aspects techniques et financiers et assurer la pérennité des données, il faut privilégier, autant que possible, des formats et logiciels ouverts, gratuits, pérennes, non propriétaires, courants et standards.

Pour aller plus loin : Cines - FACILE (outil de vérification des formats de fichiers), Doranum (quiz sur les formats) et une bonne synthèse des formats pour tout type de fichiers.

2) Choisir un type de stockage

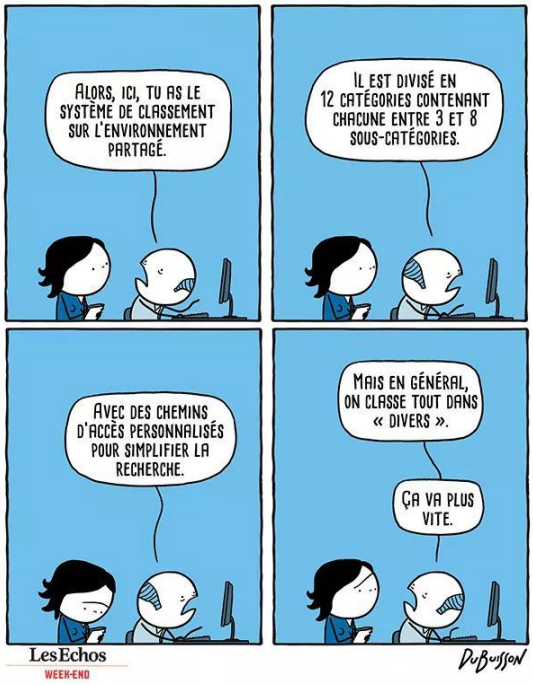
- Disque local → travail en solo
- Espace serveur partagé → travail collaboratif et rapidité des échanges dans un laboratoire, une équipe. Ce système permet un gain de place (en évitant la multiplication des copies d'un même fichier) et d'efficacité (non multiplication des versions d'un même fichier). La gestion des droits d'accès aux fichiers permet de paramétrer les actions possibles de chaque utilisateur selon son statut et son rôle dans le projet.
- Hébergement en ligne → travail collaboratif entre plusieurs laboratoires ou équipes dispersés. Les avantages sont les mêmes mais il faut absolument réfléchir à la sécurité des données. Par exemple : pas de données personnelles ou sensibles sur le Cloud ou sur un support externe, sauf si cryptées.

Dans tous les cas : définir les modalités de sauvegarde !

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
Dropbox Google Drive OneDrive 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

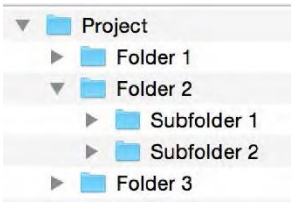
Tableau tiré de DoraNum

3) Choisir une organisation des fichiers



Deux grands types d'organisation sont envisageables :

- **Par Mots-clés :** l'organisation repose sur l'indexation des contenus et les recherches s'appuient sur le vocabulaire employé. Cette organisation est idéale pour certains types de fichiers : photos, images, publications mais nécessite l'utilisation d'un thésaurus de mots-clés et une convention à suivre par TOUS les membres du projet. C'est une organisation peu structurée et qui nécessite des logiciels particuliers de gestion et de recherche.
- **Hiérarchique :**



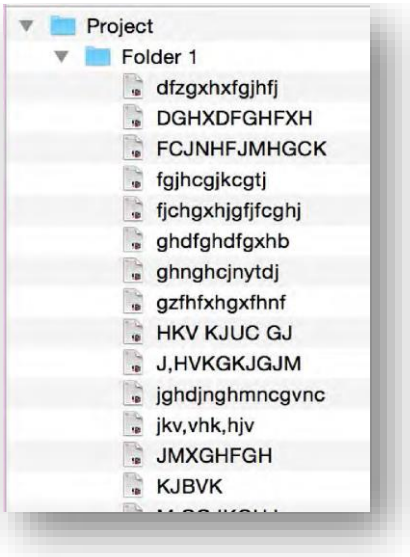
D'après Malinowski, 2017

Cette organisation implique de choisir les grandes catégories (administration, données brutes, analyses, résultats, rédaction…) en amont car il est périlleux de modifier la structure *a posteriori*. Cette structure doit répondre à différents critères :

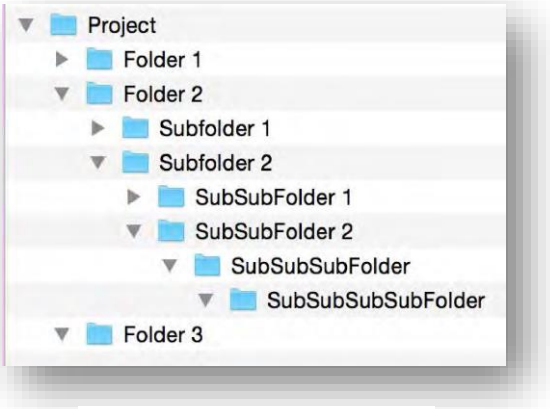
- ✓ Éviter toute redondance : un fichier=un emplacement=un chemin d'accès (utiliser des raccourcis si besoin de faire apparaitre un fichier à plusieurs emplacements)

✓ Trouver un équilibre entre largeur et profondeur

Trop large = trop détaillé DONC subdiviser !



Trop profond = pénible + incompatible avec certains systèmes d'exploitation DONC rassembler !



D'après Malinowski, 2017

4) Définir des Règles de Nommage

Afin que chacun identifie facilement les fichiers et leur contenu, il faut que les noms de fichier soient courts mais clairs et précis. Ils doivent contenir un moyen de classement rapide, par exemple une date, éventuellement le nom du projet (surtout s'il y a un risque de mélange) et l'objet du fichier.

Le format doit suivre certaines règles :

- Pas d'espaces : tirets-bas (underscore) (2019_projet_objet) ou CamelCase (2019ProjetObjet)
- Pas de caractères spéciaux, pas d'accents
- Pas de points autres que pour l'extension

Ceci est applicable aux échantillons (quoi, qui, où, comment, quand…) et certains logiciels de mesure sont paramétrables pour attribuer automatiquement les noms de fichiers.

Il existe des logiciels de renommage massif (Bulk Rename Utility, Renamer…) en cas de modifications majeures en cours de projet ou de décisions trop tardives sur la convention de nommage.

5) Définir le versionnement

Dans le même esprit que pour les règles générales de nommage, il faut être clair : éviter Final/Final1/Final2/FinalFinal… Il faut donc choisir un mode de versionnement :

- Dates : mais garder à l'esprit les cas où il peut y avoir plusieurs versions par jour, par mois…
- Numérotation simple : 01, 02…15… Anticiper le format (nombre de caractères) selon le nombre de versions prévisibles
- Numérotation par section : V1_0, V2_3, V5_1_2. Définir ce que signifient les versions de premier ordre et de second ordre voire plus

Le versionnement doit être documenté. Sur les documents rédigés (rapports, protocoles…), un historique des versions en tête de fichier permet de journaliser les modifications : ajouter une ligne pour chaque version : Auteur, Date, Détails/Modifications.

La conservation de l'intégralité des versions d'un fichier est à considérer selon leur utilité dans le projet et un bilan après la fin du projet peut permettre de diminuer la volumétrie des documents produits en choisissant soigneusement ceux à conserver.

6) Prévoir des sauvegardes

Le versionnement facilite donc le choix de ce qu'il faut sauvegarder. Il faut ensuite établir précautionneusement le processus de sauvegarde et notamment :

- Méthode :
  - Manuelle : dangereux → il faut être rigoureux et régulier
  - Automatique : beaucoup plus sûre, à privilégier dès que les infrastructures le permettent
- Règles d'or :
  - **3** copies (1 active + 2 sauvegardes)
  - **2** supports différents
  - **1** en-dehors du site

Dans une entité (projet, laboratoire, équipe) il faut absolument documenter la sauvegarde, c'est-à-dire définir qui s'en occupe, quand, comment, où ? Et désigner un référent pour cette tâche.

En synthèse :

- **Compiler toutes les règles dans un document ouvert à tous : ReadMe.txt (ou PGD!)**
- **Faire un répertoire des fichiers, surtout si l'organisation est complexe. Ceci combine les avantages des mots-clés et de l'organisation hiérarchique mais reste chronophage.**
- **Désigner un référent pour l'organisation, la sauvegarde, l'archivage.**

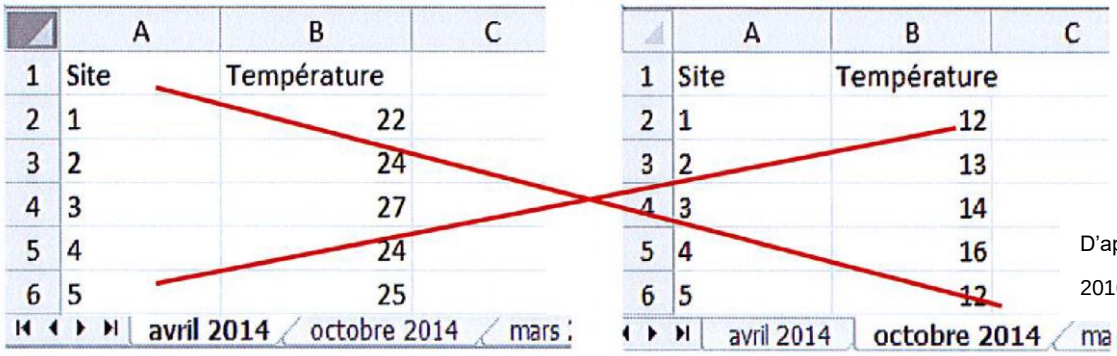
Faire au mieux mais rester réaliste en fonction des moyens dont on dispose !

III. Organiser ses données dans les fichiers, quelques exemples avec les tableaux de données

L'organisation des fichiers est un élément extrêmement important dans la bonne conduite d'un projet mais l'organisation des données dans les fichiers l'est tout autant. Quelques règles simples permettent d'optimiser le travail en solitaire ou collaboratif et d'éviter des blocages et erreurs dans les processus d'analyse et des pertes d'information. Ces règles doivent être suivies et partagées par tous les membres d'un même projet, d'une même équipe. Il est donc conseillé d'y réfléchir ensemble.

1) Structure du fichier

L'organisation des données dans les fichiers doit suivre une logique de simplicité et de non-redondance. Pour les tableurs : un seul tableau par feuille et pas de multiplication de tableaux similaires ou d'onglets dans un même fichier:



	A	B	C
1	Site	Température	
2	1	22	
3	2	24	
4	3	27	
5	4	24	
6	5	25	

avril 2014 | octobre 2014 | mars

	A	B	C
1	Site	Température	
2	1	12	
3	2	13	
4	3	14	
5	4	16	
6	5	12	

avril 2014 | octobre 2014 | ma

D'après Arnould, 2016

Si l'organisation des données devient trop complexe, il vaut mieux faire une base de données.

2) Présentation des données

Garder en tête ces trois règles d'or :

- **Une colonne** = UNE variable ; **une ligne** = UNE observation ; **une cellule** = UNE valeur
- **Pas de vides** : ni lignes, ni colonnes, ni cellules (décider d'un code unique pour les valeurs manquantes)
- **Pas de doublons**

Les en-têtes de colonnes ou noms de champ sont là pour informer clairement sur le contenu, ils doivent être sur une seule ligne et être descriptifs, clairs et homogènes d'un fichier à l'autre.

Il est donc conseillé, pour les en-têtes de colonnes/lignes autant que pour le contenu des cellules, de :

- Choisir une langue unique, un système commun d'unités
- Ne pas faire figurer les unités, ni utiliser de caractères spéciaux, d'espaces, d'accents… la lecture des fichiers en serait perturbée pour beaucoup de logiciels
- Adopter une convention d'écriture (ex: M ≠ male ≠ mâle, format des dates, choix du marqueur décimal…)

La simplicité, la lisibilité et l'intelligibilité des données les rendent transmissibles et facilement analysables. Mieux vaut donc éviter les codes couleur, les cellules fusionnées, les commentaires libres (les colonnes « Notes/Commentaires/Remarques » sont à codifier au maximum).

3) Qualité des données

Si les données sont organisées et présentées selon les règles énoncées ci-dessus, elles seront facilement utilisables mais leurs valeurs requièrent également un maximum d'attention pour en garantir la qualité.

Quelques règles simples :

- **On contrôle les valeurs :**
  - Toutes les variables qualitatives suivent une même convention d'écriture et il n'y a pas d'espaces en trop avant ou après la valeur (sinon deux valeurs apparemment égales seront considérées comme différentes dans les analyses)
  - Toutes les valeurs d'une même variable quantitative sont dans la même unité
  - Les gammes de valeurs sont cohérentes, réalistes, pertinentes (min/max, étendue…)
  - La complétude est bonne : tout est renseigné sauf si c'est justifié
- **On ne modifie pas les données brutes et on les conserve !**
- **On tient compte et on s'accorde sur la précision et l'exactitude des données :**
  - Mesures physiques : calibrage des appareils, contrôle statistique des distributions

- Temporelle : selon le contexte et les questions scientifiques
- Spatiale : un géoréférencement peut être très précis mais faux et à l'inverse, il peut être exact mais avec une très faible précision. L'idéal étant bien sûr d'être exact et précis !
- Sémantique : référentiels/thésaurus/ontologies (géographiques, taxonomiques, standards…)

Plusieurs méthodes permettent d'appliquer ces règles aux données. Par exemple, il est possible de fixer des contraintes sur certaines variables (min/max, obligatoirement positif, nulles possibles oui/non?) ; d'effectuer des contrôles logiques (ex: type_relief = montagne/altitude = 1m =>erreur) ; de confronter les valeurs à des tables de références ; d'utiliser des indicateurs de qualité : vérifié=Vrai/Faux, indice de niveau de contrôle...

Conclusions

Pour une gestion sereine et efficace des données d'un projet ou d'une équipe, il est nécessaire d'anticiper un maximum de paramètres et de se coordonner en identifiant clairement les responsabilités de chacun.

Il est important de documenter autant que possible chaque élément : dictionnaire de données (liste des champs, explication littérale du contenu, unité, type de valeurs, contraintes éventuelles (min/max…)), commentaires dans les codes informatiques, règles de gestion des fichiers et des données, modes de sauvegardes.

La combinaison de tous ces éléments permet ainsi d'envisager plus sereinement la gestion des éventuels aléas techniques en cours de projet, de s'assurer que chaque partenaire a minima possède le même niveau de compréhension sur l'ensemble des éléments, d'éviter la réponse en urgence aux demandes des financeurs concernant l'ouverture des données, et de limiter les conséquences fâcheuses !

L'idéal est de compiler le maximum d'informations et de décisions dans un Plan de Gestion des Données complet et régulièrement révisé.

Sources

Arnould, P.-Y. and M.-C. Jacquemot-Perbal (2016). Guide de bonnes pratiques : Gestion et valorisation des données de recherche, CNRS

Flamerie, F. (2018). Organiser efficacement ses données - Document de cours, Urfist Bordeaux

Malinowski, C. (2017). Data Management: File Organization, MITLibraries

Plumejeaud-Perreau, C. and N. Mandran (2018). Qualité des données. ANF « Sciences des données : un nouveau challenge pour les métiers liés aux bases de données », 5-7 novembre 2018, Sète, CNRS

Saby, M. (2019). Organiser, documenter et protéger ses données au quotidien. Formation Doctorale, Université Nice-Sophia-Antipolis

INRA Pôle Données de la Recherche IST, 2016. Nommage et organisation des fichiers.

Doranum, 2017. Comment bien nommer ses fichiers.

Quidoz, M.-C. (2018). Les principes FAIR appliqués aux sauvegardes sur le long terme. In : CNRS. Interopérabilité et pérennisation des données de la recherche - Comment « FAIR » en pratique ? Paris, 27 Novembre 2018.