Gérer ses données
Pourquoi? Comment?

Gaëlle Jaouen
UMR EcoFoG

25/11/2019

Partager les données de la recherche : pourquoi, comment ?
Guadeloupe, du 25 au 27 Novembre 2019 – CRAG INRA

Contexte

- Déroulement d'un projet:
  - Avant : Anticipation, préparation, planification
  - Pendant : Réalisation, manipes, accumulation de ***données***
  - Après : Rapport final, publications scientifiques, ouverture/archivage des ***données***

→ Les données sont le carburant de la science ☺

Objectifs

- Gérer ses données de la conception du projet à l'ouverture pour
  - Gagner du temps
  - Être efficace, éviter les erreurs
  - Éviter des catastrophes!
  - Attester fiabilité des résultats
  - Produire des données réutilisables par soi et par les autres, Être prêts à les partager

→ Faire au mieux pour soi et pour le projet

Anticiper

- Investir un peu de temps dans l'anticipation pour éviter d'en perdre beaucoup plus tard…
- Prévoir quelles données vont être produites
  - Types?
  - Volumes?
  - Format?
- Qui va les utiliser
  - Une personne?
  - Plusieurs dans le même labo?
  - Plusieurs dans des labos différents?
- Comment
  - Fréquence?
  - Flux?

Anticiper

- Que faudra-t-il conserver? Archiver?
  - Quoi?
    - données chères, non-reproductibles
    - longues et difficiles à produire, ≠ modélisation, simulation sauf si processus longs
    - Forte valeur patrimoniale
  - Volume?
  - Durée?
  - Où? Local, entrepôt…
- Protéger selon :

  risque (destruction, perte…) x probabilité

Anticiper

- Savoir de quels moyens on dispose
  - Logiciels
  - Espaces de stockage, locaux ou externes
  - Compétences
- Se mettre d'accord au sein de l'équipe
- Définir qui fait quoi, référents
- Choisir une langue (FR/EN)

→ Commencer à remplir son PGD ☺

→ Rester raisonnable! Ne pas en faire trop

→ Il n'existe pas d'organisation idéale, trouver le mieux pour VOUS

Organiser les fichiers

- Choisir un lieu de stockage
  - Disque local → solo
  - Espace serveur partagé → un labo
    - Gain de place
    - Rapidité d'échanges
    - Évite chargement de serveurs par mail lourds (+écolo)
    - Versions uniques de fichier, pas de doublons
    - Gérer droits des utilisateurs
  - Hébergement en ligne (GoogleDrive etc) → plusieurs labos
    - Mêmes avantages
    - Penser sécurité

Types de stockage

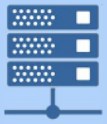
Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 Ordinateur professionnel	★★☆☆ Sujet au piratage informatique, aux détériorations et pannes	★☆☆☆ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud…)	★★★★ Pas de coût supplémentaire ou coût peu 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/

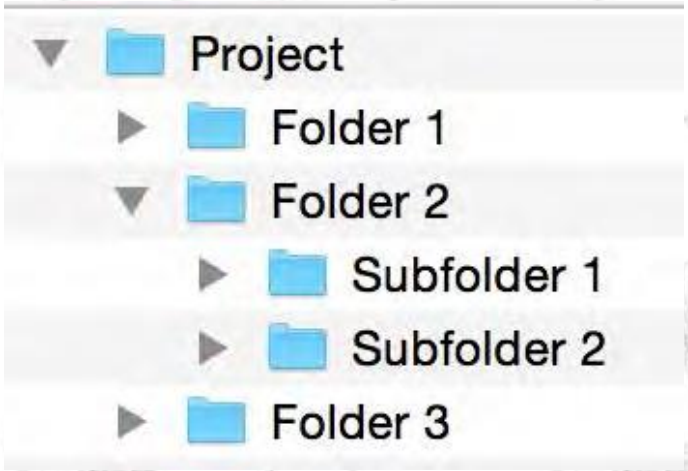
Pas de données personnelles ou sensibles sur le Cloud ou sur un support externe, sauf si cryptées

Organiser les fichiers

- Choisir une organisation
  - Par Mots-clés
    - Organisation fluide
    - Recherche facile, rapide
    - Utilisation d'un thésaurus de mots-clés/une convention à suivre par TOUS
    - Idéal pour certains types de fichiers : photos, images, publi
    - Peu structuré
    - Besoin de logiciels particuliers (TagFlow + autres exemples)

Organiser les fichiers

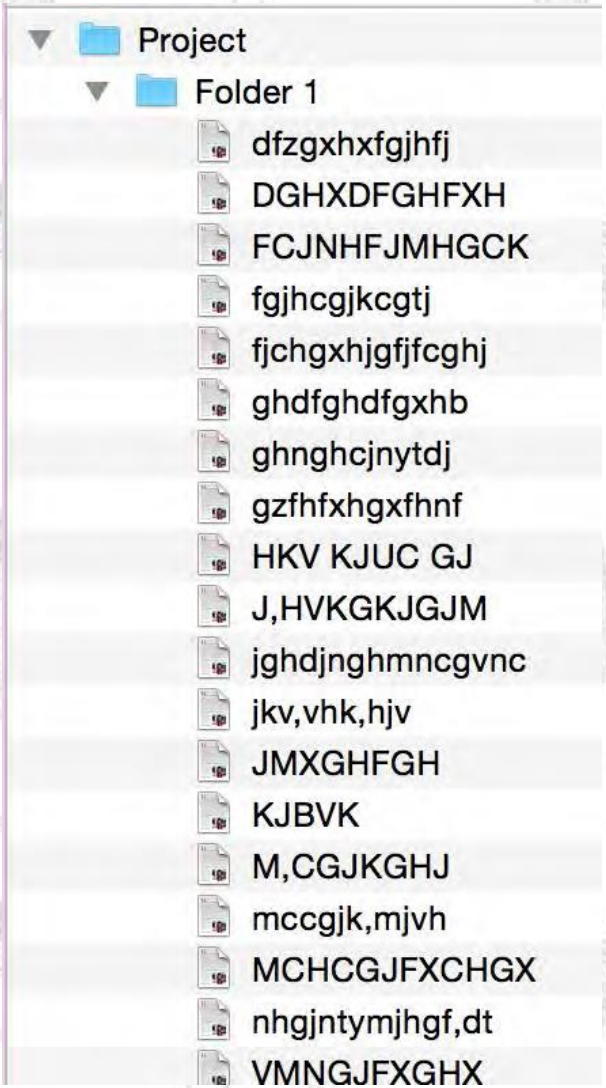
- Choisir une organisation
  - Par Mots-clés
  - Hiérarchique



Organiser les fichiers

- Organisation Hiérarchique
  - Choisir les grandes catégories (admin, données brutes, résultats, rédaction…)
  - Éviter toute redondance :
    - un fichier=un emplacement=un chemin d'accès
    - Mais utiliser des raccourcis Windows si besoin
  - Trouver un équilibre entre largeur et profondeur
    - Trop large = trop détaillé (ou pas assez!)

Organiser les fichiers

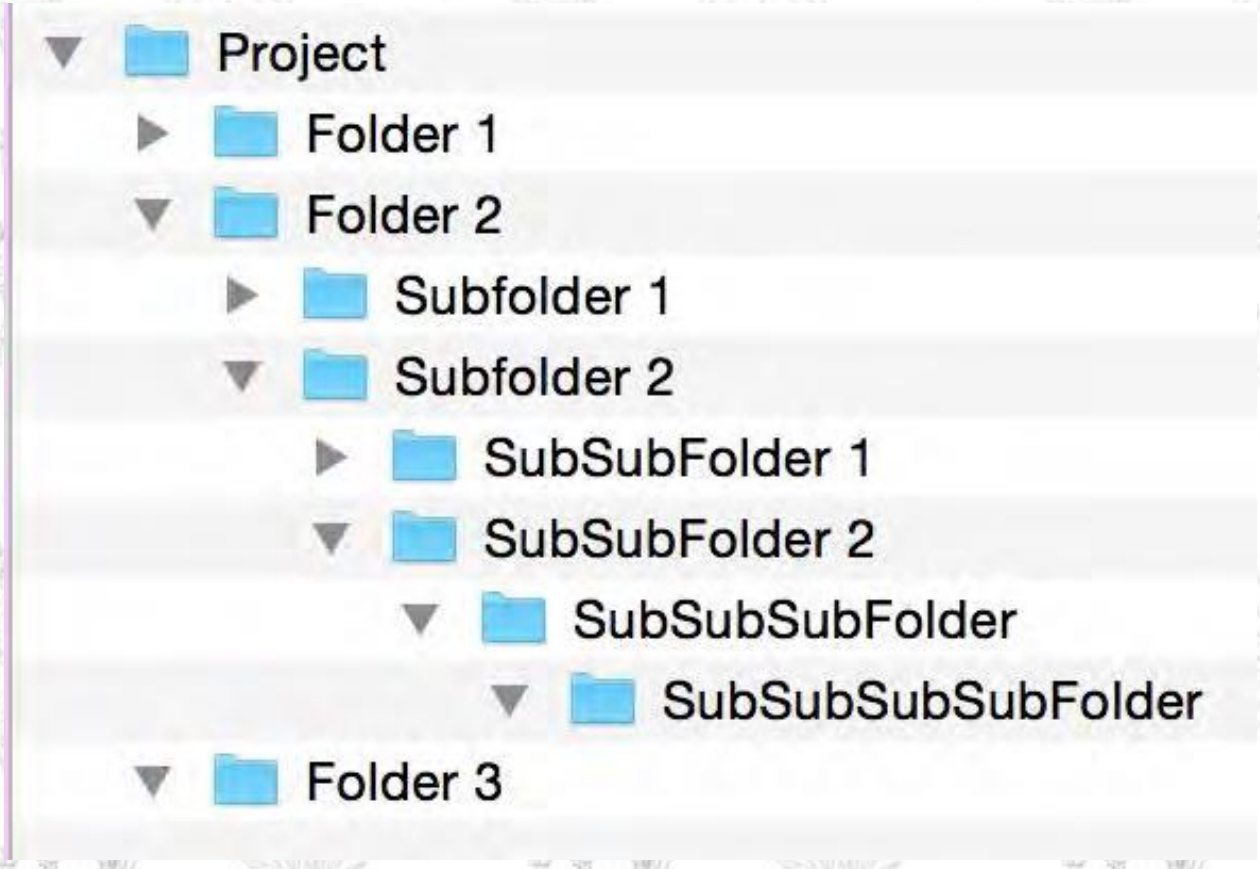


→ Organiser, Subdiviser

Organiser les fichiers

- Organisation Hiérarchique
  - Choisir les grandes catégories (admin, données brutes, résultats, rédaction…)
  - Éviter toute redondance :
    - un fichier=un emplacement=un chemin d'accès
    - Mais utiliser des raccourcis Windows si besoin
  - Trouver un équilibre entre largeur et profondeur
    - Trop large = trop détaillé
    - Trop profond = pénible + incompatible avec certains OS

Organiser les fichiers



→ Organiser, Rassembler

Organiser les fichiers

- Organisation Hiérarchique
  - Choisir les grandes catégories (admin, données brutes, résultats, rédaction…)
  - Éviter toute redondance :
    - un fichier=un emplacement=un chemin d'accès
    - Mais utiliser des raccourcis Windows si besoin
  - Trouver un équilibre entre largeur et profondeur
    - Trop large = trop détaillé
    - Trop profond = pénible + incompatible avec certains OS
  - Bien anticiper!

Organiser les fichiers

- Définir une Convention de Nommage
  - Contenu des noms
    - Court, clair et précis
    - Date : permet classement rapide
    - Projet : si risque de mélange
    - Objet
  - Format
    - Pas d'espaces : underscore (2019_project_data) ou CamelCase (2019ProjectData)
    - Date : année en premier, jour-mois-année, subdivisions?
    - Pas de caractères spéciaux, pas d'accents
    - Pas de points autres que pour l'extension

→ Applicable aux échantillons (quoi, qui, où, comment, quand…)

→ Certains logiciels de mesure sont paramétrables/noms

→ Existence de logiciels de renommage massif (Bulk Rename Utility, Renamer…)

→ Tout le monde suit la convention!

Organiser les fichiers

- Définir le **Versionnement**
  - Éviter Final/Final1/Final2/FinalFinal… être clairs
  - Mode de versionnement
    - Dates : sauf si plusieurs versions par jour!
    - Numérotation simple :
      - prévoir plusieurs caractères 01-02…15…
      - Anticiper le format selon le nombre de versions
    - Numérotation par section :
      - V1_0, V2_3, V5_1_2
      - Définir ce que signifient les versions
    - Historique des versions en tête de fichier (surtout doc)
      - Ajouter une ligne pour chaque version : Auteur, Date, Détails/Modifications

Organiser les fichiers

- Définir le Versionnement
  - Mode de versionnement
  - Faire un dossier « Old » ou « Archives » si premières versions sont à conserver
  - Si beaucoup de fichiers, utiliser logiciels type Git qui gèrent le versionnement (GitHub, GitLab, FramaGit…)

Organiser les fichiers

- Prévoir des sauvegardes
  - Versionnement facilite choix de quoi sauvegarder
  - Méthode :
    - Manuelle : dangereux → être rigoureux!
    - Automatiques : plateforme, politique de labo
  - Règles d'or :
    - **3** copies (actives + 2 sauvegardes)
    - **2** supports (DD, serveur, cloud…)
    - **1** en-dehors du site
  - Documenter! Qui, Quand, Comment, Où

Organiser les fichiers

- Conclusions
  - Compiler toutes les règles dans un document ouvert à tous : ReadMe.txt (ou PGD!)
  - Faire un répertoire des fichiers, surtout si organisation complexe
    - Combine avantages des mots-clés et de l'organisation hiérarchique
    - Un peu chronophage…
  - Attention au cryptage, ne pas perdre le mdp !
  - Désigner un référent pour l'organisation, la sauvegarde, l'archivage

Organiser les données

- Fondamentaux:
  - Savoir où les trouver : gestion des fichiers
  - Savoir les lire et les relire : choix des formats/logiciels
  - Savoir les comprendre : documentation, métadonnées, dictionnaire de données
- Choix des logiciels…
  - Non-propriétaires/ouverts, pérennes
  - Gratuits autant que possible
  - Utilisation courante dans votre communauté scientifique
- … et des formats
  - Lisibles par humains/machines
  - Standards (csv, txt, pdf, xml…)

  → Conversion possible en fin de projet!

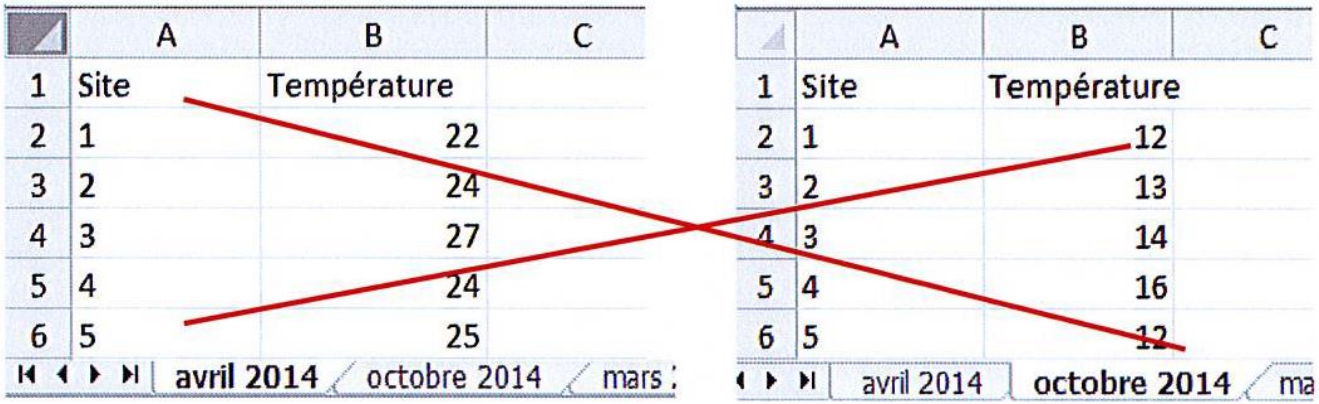
Comprendre et réutiliser les données

- Documenter:
  - Métadonnées
  - Dictionnaire de données

→ **Esther**

Bonnes pratiques pour les données

- Tableau de données
  - Un seul tableau par feuille
  - Éviter multiplication de tableaux similaires, d'onglets



	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

  - Si trop complexe : faire une base de données

Bonnes pratiques pour les données

- Tableau de données
  - Entêtes de colonnes/noms de champ:
    - Une ligne et une seule
    - Descriptifs, clairs et homogènes d'un fichier à l'autre
    - Choisir une langue
    - Autant que possible : pas d'unité
    - Pas de caractères spéciaux, d'espaces, d'accents…
  - Une colonne = UNE variable
  - Une ligne = UNE observation
  - Une cellule = UNE valeur
  - Pas de vides : ni lignes, ni colonnes, ni cellules
  - Pas de doublons!

Bonnes pratiques pour les données

- Tableau de données
  - Contenu uniforme :
    - Valeurs manquantes = un code unique
    - Convention d'écriture : M ≠ male ≠ mâle
    - Format des dates ; une ou plusieurs colonnes?
    - Choix du marqueur décimal
  - Pas d'unités de mesure dans les cellules
  - Éviter commentaires : colonne « Notes » à codifier au maximum
  - Pas de code couleur : incompréhensible, illisible, non interopérable
  - Pas de cellules fusionnées
  - Pas de graphiques

Qualité des données

- Choix d'unités : standards internationaux
- Contrôle des valeurs
  - Uniformité des variables qualitatives, des formats
  - Unités homogènes, la même partout!
  - Valeurs min/max, cohérentes, réalistes/pertinentes
  - Trouver origine si données aberrantes
- Ne jamais modifier les données brutes

Qualité des données

- Complétude : tout est renseigné (sauf si justifié)
- Précision/Exactitude
  - Mesure physique : calibrage, contrôle stat
  - Temporelle : selon le contexte scientifique
  - Spatiale

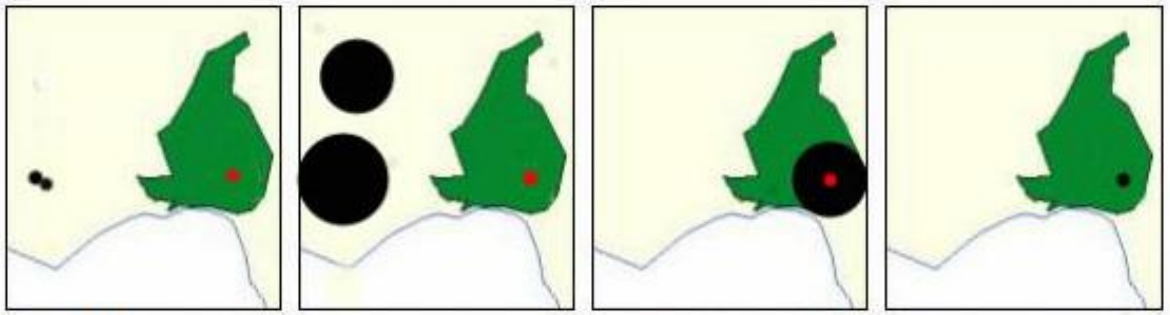


Fig. 1. *Montre les différences entre exactitude et précision dans un contexte spatial. Les points rouges indiquent les vraies localisations, les points noirs représentent les localisations rapportées par un collecteur.*

a. Haute précision, faible exactitude,
b. Faible précision, faible exactitude indiquant une erreur aléatoire,
c. Faible précision, haute exactitude,
d. Haute précision et haute exactitude.

(Chapman, 2005)

Qualité des données

- Précision/Exactitude
  - Sémantique : référentiels/thésaurus/ontologies (géographiques, taxonomiques, standards…)
- Méthodes
  - Limitations des valeurs :
    - Contraintes : min/max, null possible?
    - Contrôle logique : ex: mâle + utérus = faux/erreur
    - Tables de références dans BD
  - Utilisation d'indicateurs de qualité : Vrai/Faux, indice de niveau de contrôle

Conclusions

- Anticiper
- Se coordonner
- Gérer responsabilités et droits d'accès
- Documenter les données :
  - Dictionnaire de données
  - Commentaires dans les codes informatiques
- Documenter l'ensemble du projet : Fichier ReadMe.txt
  - Règles de nommage et d'organisation
  - Répertoire de fichiers
  - Contenu des fichiers / dictionnaire de données / Référentiels
  - Logiciels ou codes informatiques nécessaires
  - Précautions pour la réutilisation
  - Référents, responsables
- Planifier, compiler : PGD !!

Sources

- Arnould, P.-Y. and M.-C. Jacquemot-Perbal (2016). Guide de bonnes pratiques : Gestion et valorisation des données de recherche, CNRS
- Chapman, Arthur D. (2005). Les principes de qualité des données, version 1.0. Trad. Chenin, N. Copenhague: Global Biodiversity Information Facility, 76 pp. http://www.gbif.org/document/80626
- DoRANum (2017). Métadonnées, standards et formats (www.doranum.fr)
- 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