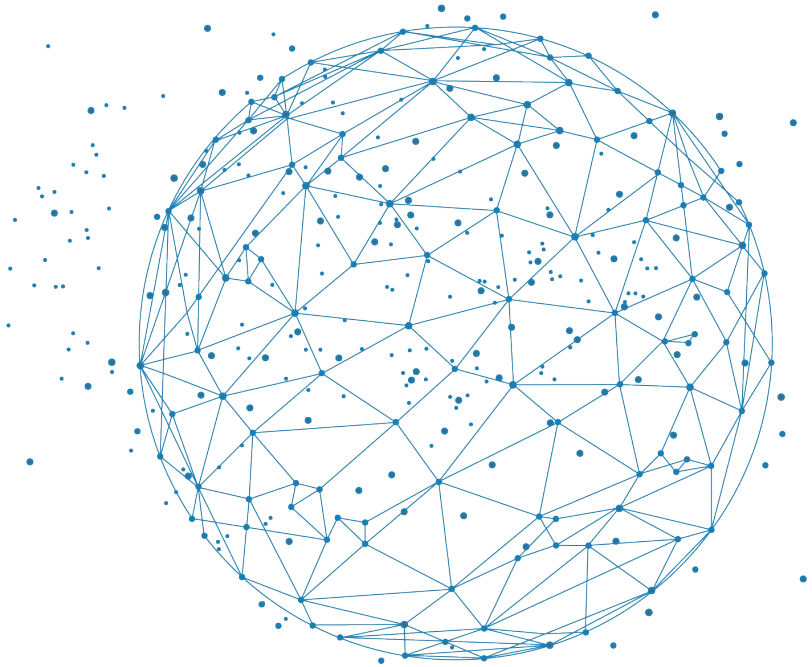


Gestion des données au cours du projet de recherche

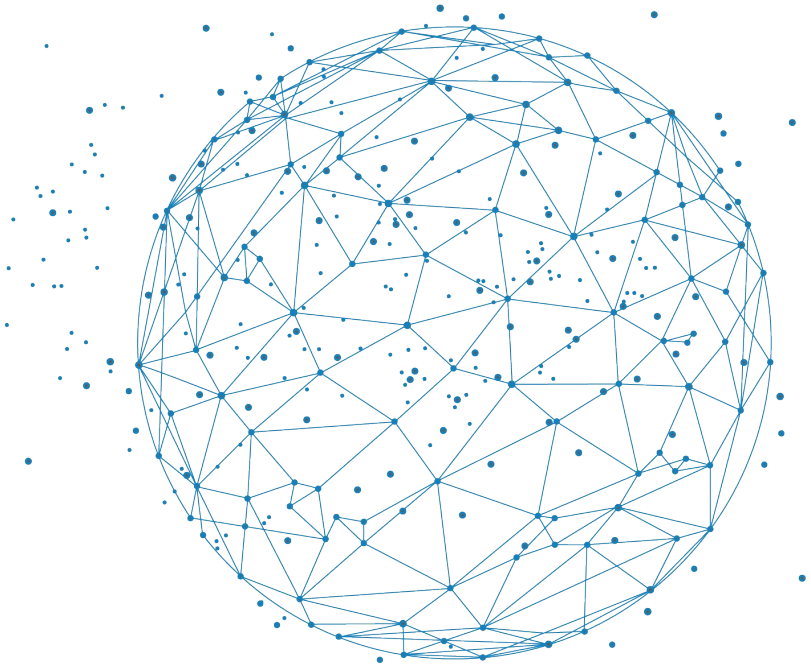
Thomas Denecker

https://orcid.org/0000-0003-1421-7641

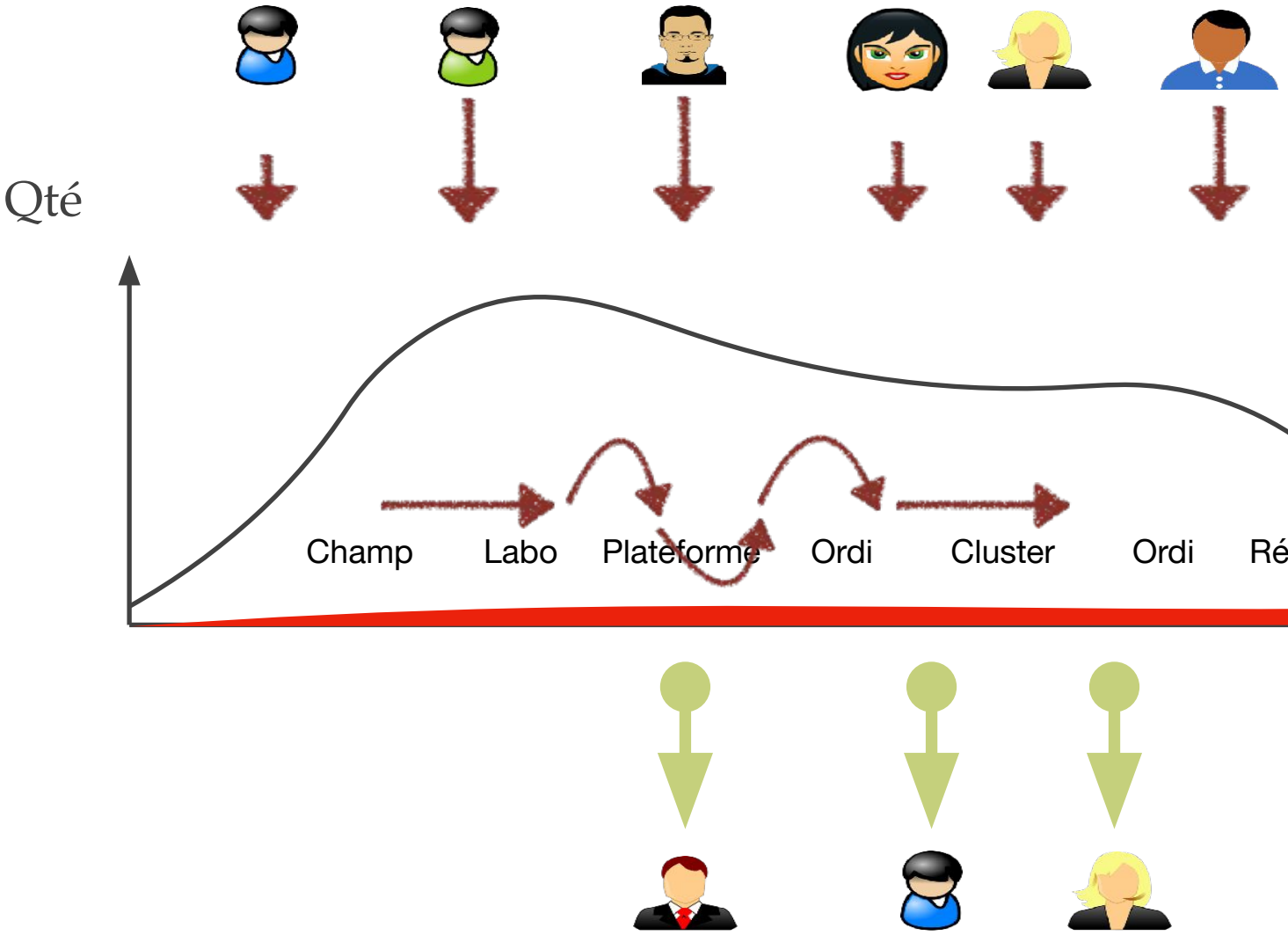
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

ELIXIR FRANCE

La vie des données



Un projet sur la durée




Cycle de vie des données scientifiques
01
IMAGINER
02
CONCEVOIR & PLANIFIER
PGD
PROTOCOLE
03
RÉALISER
MÉTADONNÉES
04
PUBLIER
DATA PAPER
ARTICLE scientifique
05
CONSERVER /PRÉSERVER
06
DIFFUSER
Chercheur producteur
LOI
07
RÉUTILISER
Chercheur réutilisateur ou producteur

Bonnes pratiques dans la gestion des données

Plusieurs personnes

Plusieurs techniques

Plusieurs lieux

Plusieurs années

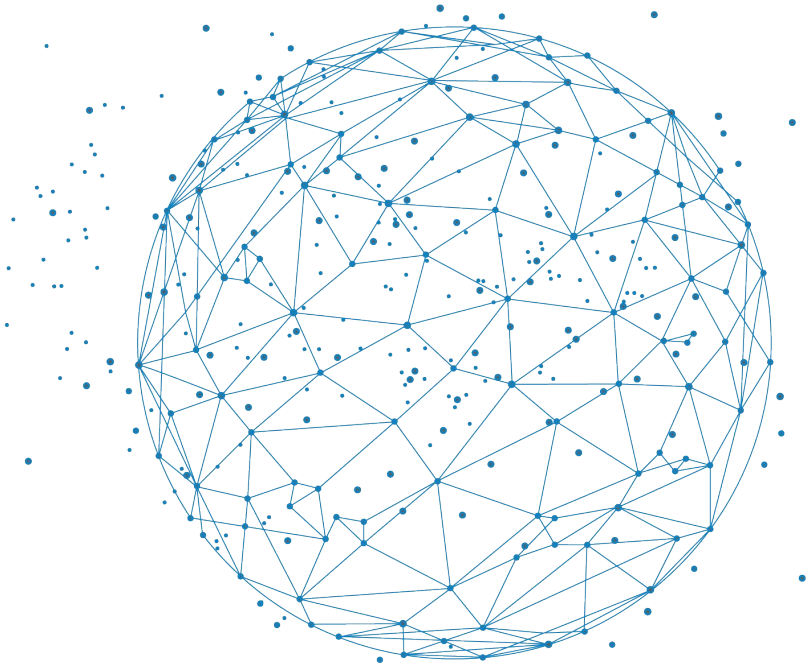
Ne rien perdre

Pouvoir retrouver

Pouvoir réanalyser

Pouvoir partager

Un environnement de travail sûr



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 chiffrées (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 ?

Vos mots de passes (au pluriel)

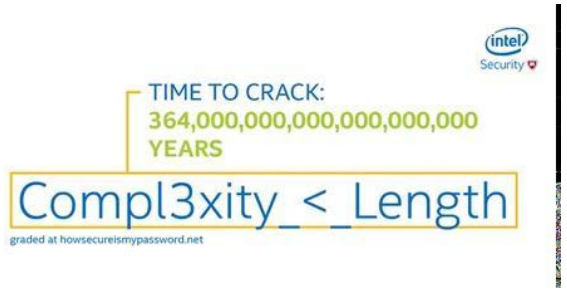
Temps pour le craquer
0.0001 secondes

Comp

Temps pour le craquer
364 000 000 000 000 000 000 ans

Compl3xity_<_Length

Source:



Un gestionnaire de mots de passe : Bitwarden

Bitwarden est un service en ligne qui vous permet de créer un coffre fort dans lequel vous allez pouvoir enregistrer tous vos mots de passe.

OpenSource

et donc pérenne

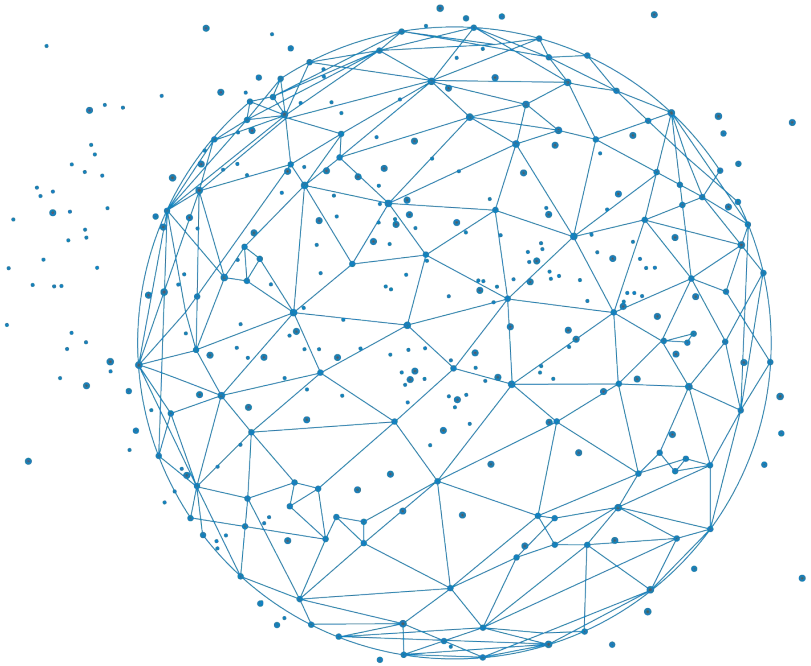
Gratuit

mais n'hésitez pas à payer la souscription Premium pour soutenir le projet

Accessible

Application Mac, Windows, Linux, Web, iPhone et Androïd

Stockage



Comparatif de systèmes de stockage des données

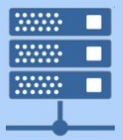
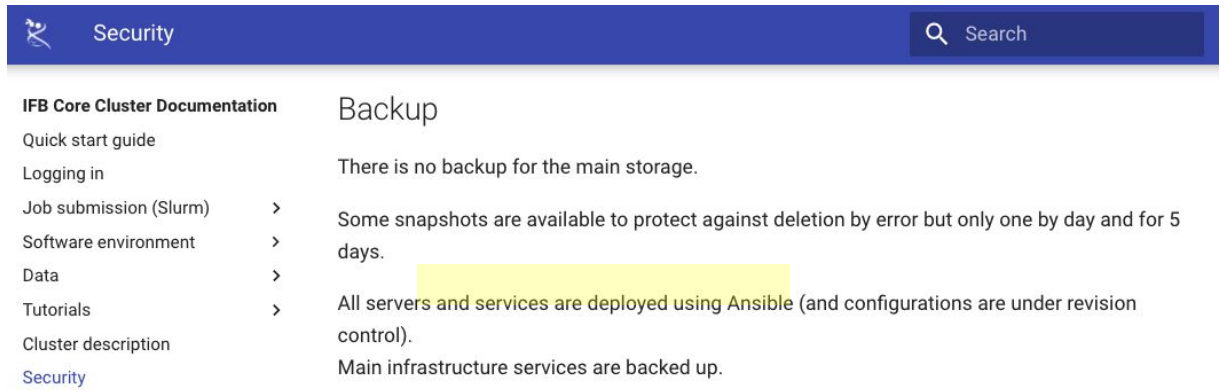
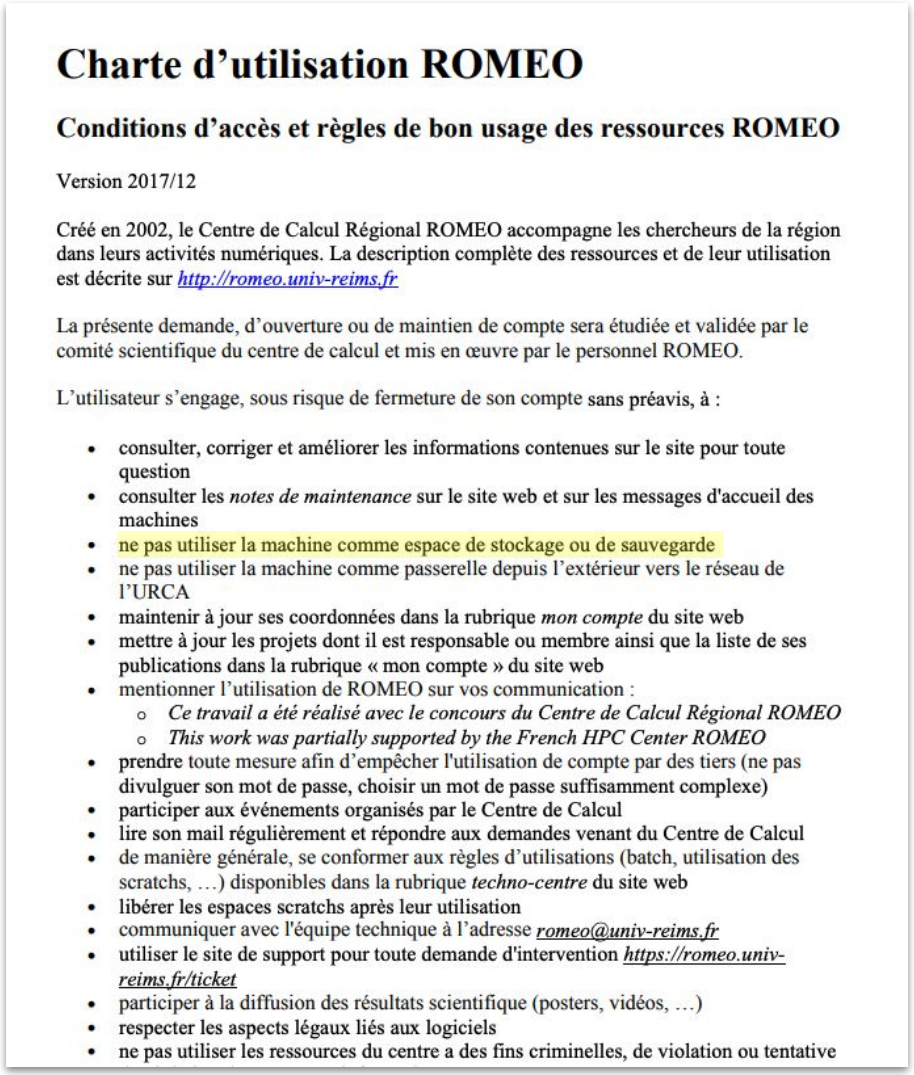
Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 Ordinateur professionnel	★★☆☆ Sujet au piratage informatique, aux détériorations et pannes	★☆☆☆ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud…)	★★★★ Pas de coût supplémentaire ou coût peu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

Attention !

Infrastructure de calcul ne rime pas toujours avec infrastructure de stockage





Charte d'utilisation ROMEO

Conditions d'accès et règles de bon usage des ressources ROMEO

Version 2017/12

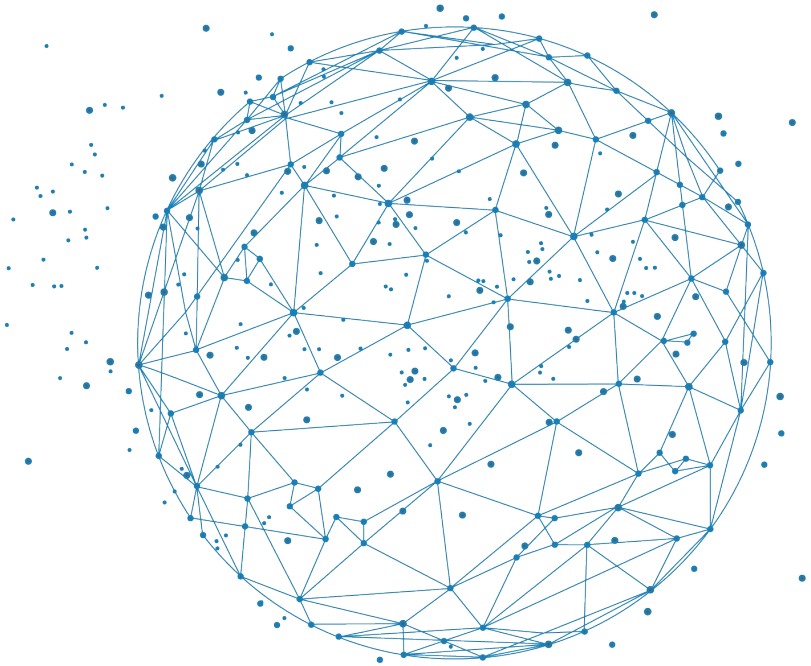
Créé en 2002, le Centre de Calcul Régional ROMEO accompagne les chercheurs de la région dans leurs activités numériques. La description complète des ressources et de leur utilisation est décrite sur *http://romeo.univ-reims.fr*

La présente demande, d'ouverture ou de maintien de compte sera étudiée et validée par le comité scientifique du centre de calcul et mis en œuvre par le personnel ROMEO.

L'utilisateur s'engage, sous risque de fermeture de son compte sans préavis, à :

- consulter, corriger et améliorer les informations contenues sur le site pour toute question
- consulter les *notes de maintenance* sur le site web et sur les messages d'accueil des machines
- ne pas utiliser la machine comme espace de stockage ou de sauvegarde
- ne pas utiliser la machine comme passerelle depuis l'extérieur vers le réseau de l'URCA
- maintenir à jour ses coordonnées dans la rubrique *mon compte* du site web
- mettre à jour les projets dont il est responsable ou membre ainsi que la liste de ses publications dans la rubrique « mon compte » du site web
- mentionner l'utilisation de ROMEO sur vos communication :
  - *Ce travail a été réalisé avec le concours du Centre de Calcul Régional ROMEO*
  - *This work was partially supported by the French HPC Center ROMEO*
- prendre toute mesure afin d'empêcher l'utilisation de compte par des tiers (ne pas divulguer son mot de passe, choisir un mot de passe suffisamment complexe)
- participer aux événements organisés par le Centre de Calcul
- lire son mail régulièrement et répondre aux demandes venant du Centre de Calcul
- de manière générale, se conformer aux règles d'utilisations (batch, utilisation des scratchs, …) disponibles dans la rubrique *techno-centre* du site web
- libérer les espaces scratchs après leur utilisation
- communiquer avec l'équipe technique à l'adresse *romeo@univ-reims.fr*
- utiliser le site de support pour toute demande d'intervention *https://romeo.univ-reims.fr/ticket*
- participer à la diffusion des résultats scientifique (posters, vidéos, …)
- respecter les aspects légaux liés aux logiciels
- ne pas utiliser les ressources du centre a des fins criminelles, de violation ou tentative

Échange de données



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

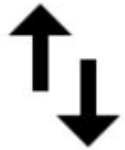
Envoi d'un disque

Dropbox, Drive, etc

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

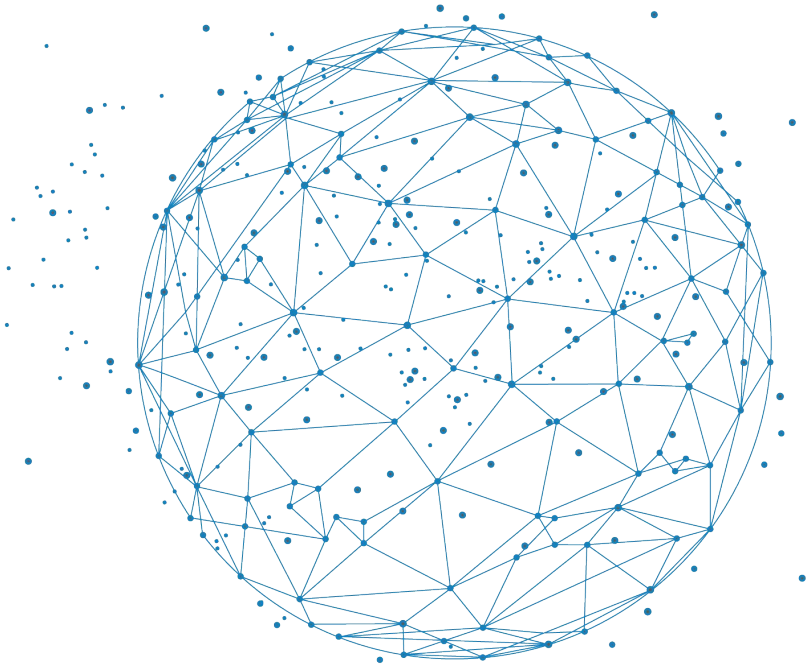
Cloud privé

Service d'un consortium



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

Protéger ses données



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, …)
- un transfert de données défectueux
- un plantage d'un disque dur
- ...

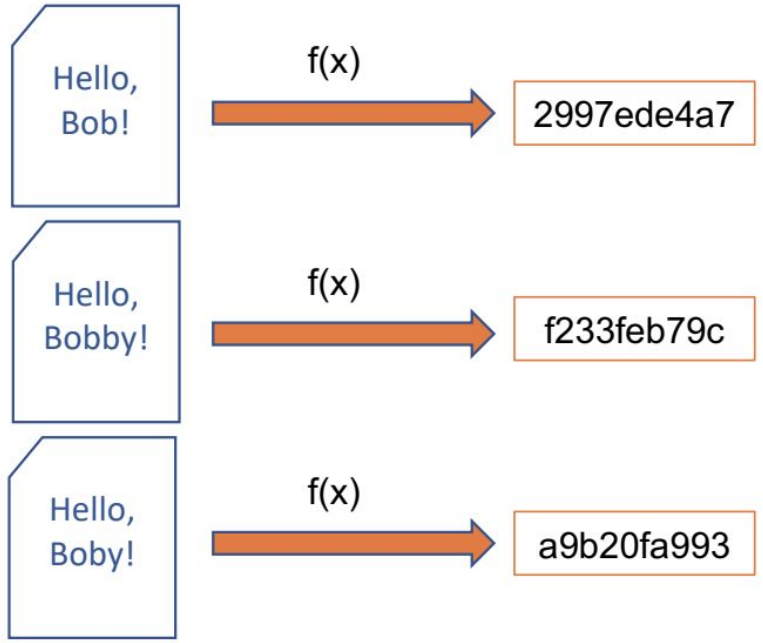
Intégrité des données

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

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

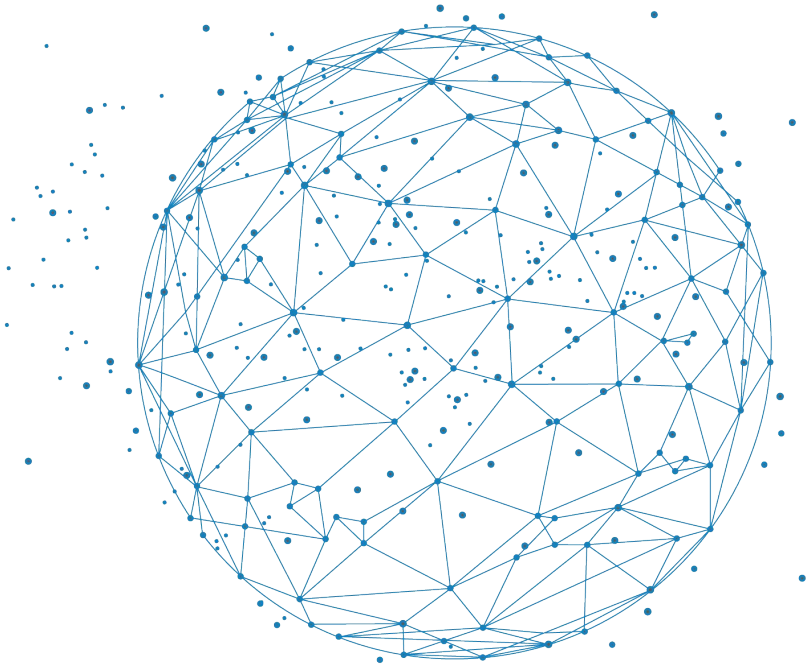
L'accès aux données sensibles doit être documenté

Copie des données

Limitez les copies au maximum !

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

La suppression des données



Suppression des données

Est-ce que ces données peuvent être supprimées ?

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)

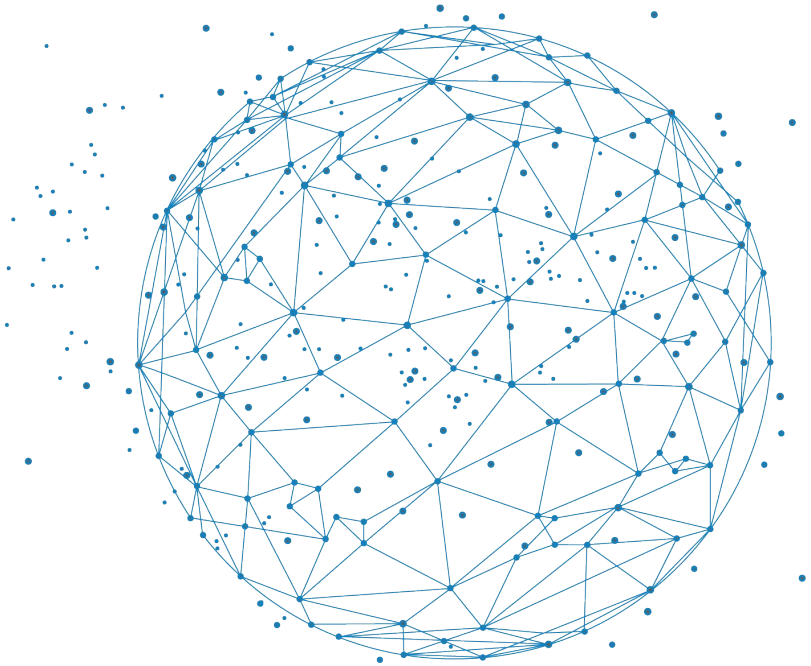
Des questions pour la suite (FAIR data by IFB)

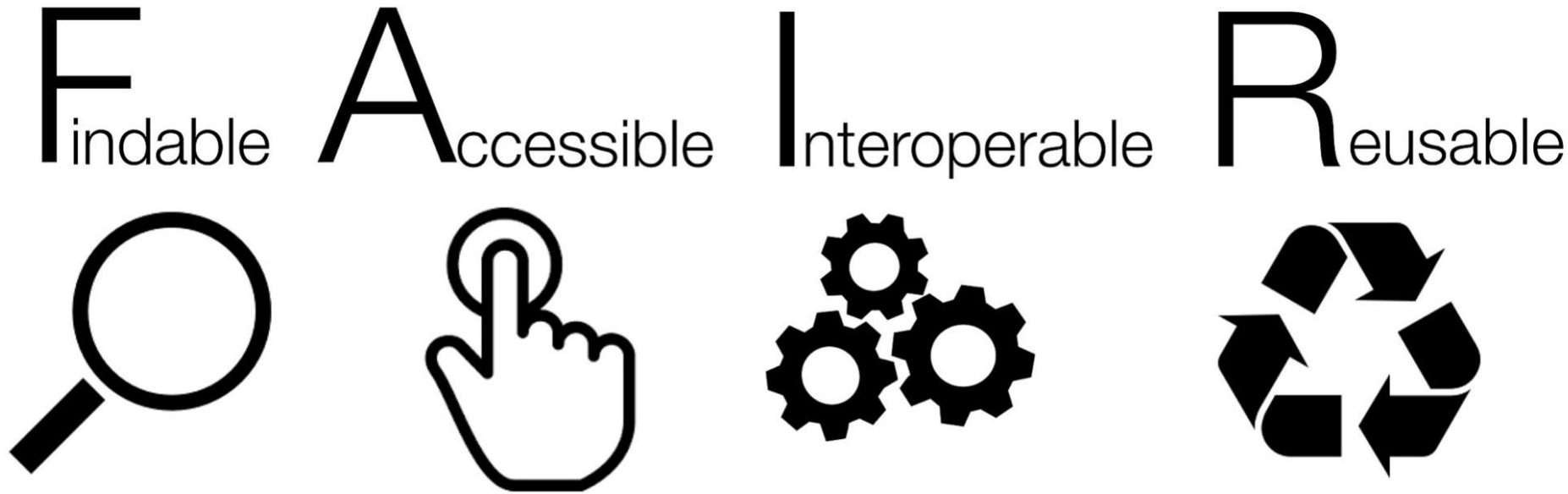
- Quels sont vos obligations en terme de rétention de données
- Dans quelles conditions allez-vous les archiver ?
- Avez-vous documenté clairement vos données ?
- Que se passera-t-il si vous partez (pour l'éternité) ?

Les infrastructures de stockage sont vos amies

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

Le nommage des fichiers





Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

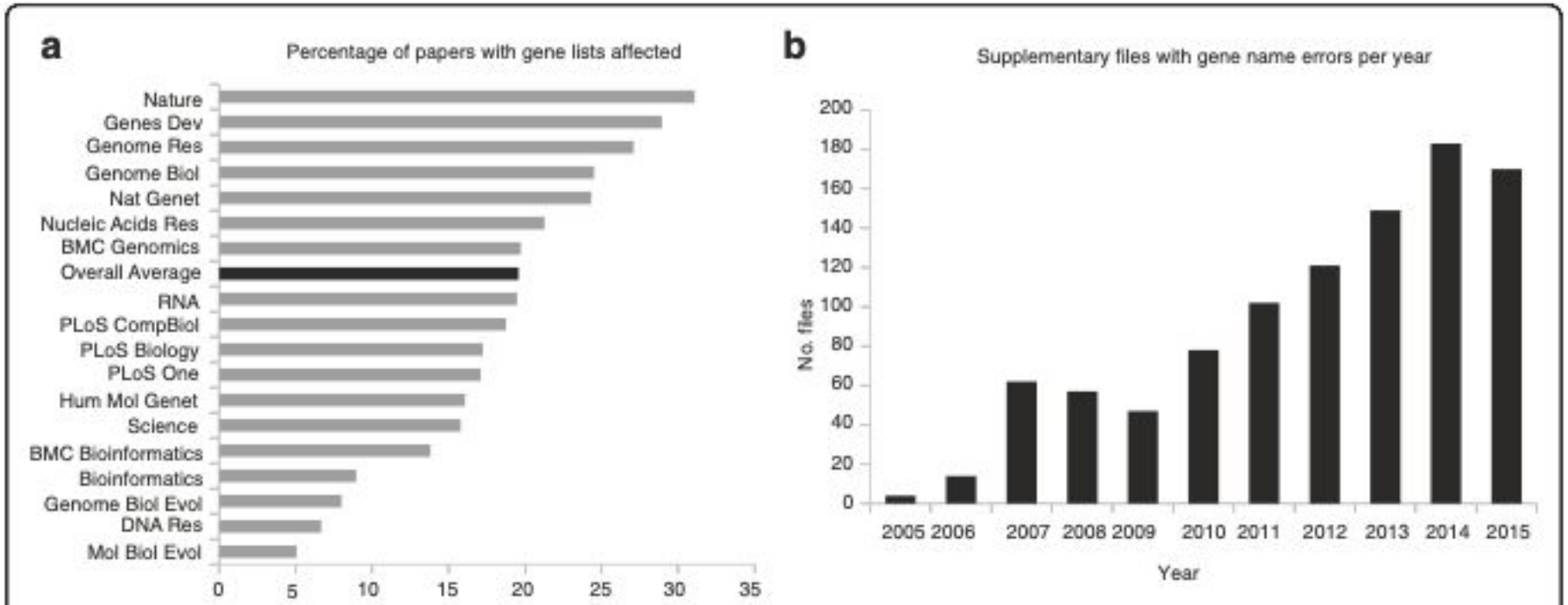


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

DONNER UN NOM BREF ET EXPLICITE et …

Pas d'espace Ni de caractères spéciaux (& / + > : ? % …)

Règles dénomination fichiers ✗

ReglesDenominationFichiers ✓

Dates au **format AAAAMMJJ** (année, mois, jour)

20150405_CR 20160310_CR 20160515_CR

Versionnez

Convention_V01 Convention_V02 Convention_VF

Rangez

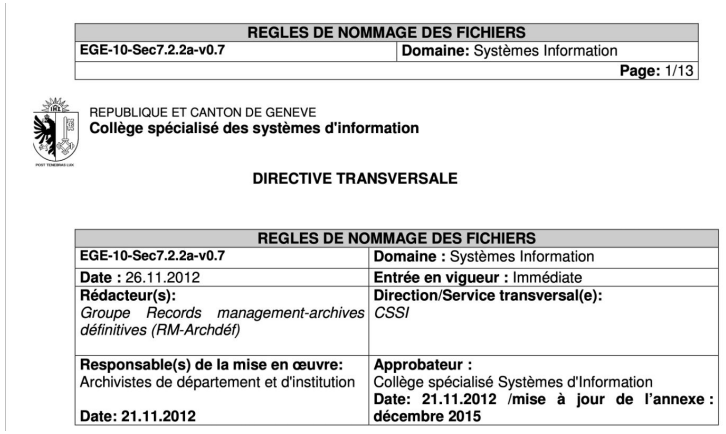
Reunion

20150407_CR

20150407_Minutes

20150407_OJ

Et documentez vos règles !



REGLES DE NOMMAGE DES FICHIERS	
EGE-10-Sec7.2.2a-v0.7	**Domaine:** Systèmes Information
	Page: 1/13

REPUBLIQUE ET CANTON DE GENEVE
Collège spécialisé des systèmes d'information

DIRECTIVE TRANSVERSALE

REGLES DE NOMMAGE DES FICHIERS	
EGE-10-Sec7.2.2a-v0.7	**Domaine :** Systèmes Information
Date : 26.11.2012	**Entrée en vigueur :** Immédiate
Rédacteur(s): *Groupe Records management-archives définitives (RM-Archdéf)*	**Direction/Service transversal(e):** *CSSI*
Responsable(s) de la mise en œuvre: Archivistes de département et d'institution **Date: 21.11.2012**	**Approbateur :** Collège spécialisé Systèmes d'Information **Date: 21.11.2012 /mise à jour de l'annexe : décembre 2015**

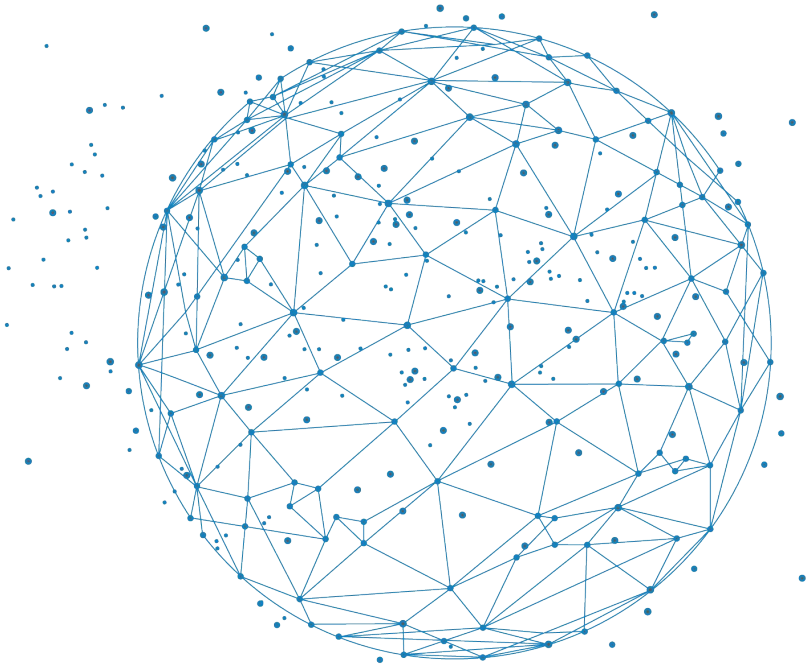
Eléments	Règle	Exemple
Sujet	**Obligatoire** Il s'agit du sujet principal traité au sein du document. Utiliser des noms communs, écrits en lettres minuscules non accentuées.	 projet formation évaluation
Séparateur	Les espaces sont interdits. Utiliser l'underscore (touche 8 du clavier) pour remplacer les espaces	 « _ »
Type de document	**Facultatif** Qualifie la nature du document. Toute abréviation sera en lettres majuscules.	 (CR) compte rendu (OJ) ordre du jour
Date	**Obligatoire** Date de création du document, date de l'événement. Format à l'américaine : AAAAMMJJ. Nommage d'une période : utilisation d'un séparateur «_» ou « -».	 20180122 201608 2010 201501_07 ou 201501-07
Version du document	**Obligatoire** Distingue les différentes versions d'un document, signalées par un « V » majuscule suivi de deux chiffres ; version provisoire (VP) et la version finale (VF), version validée (VV). Un nouveau document créé à partir d'une version finale doit être sauvegardé sous un nouveau nom de manière à ne pas écraser la version précédente.	 CR_CFVU_V0.0 CR_CFVU_V0.1 CR_CFVU_VP, VF ou VV
Extension	**Obligatoire** L'extension est ajoutée automatiquement par le système et n'apparaît peut-être pas sur vos écrans.	 .txt (fichier texte) .doc (fichier Word) .xls (fichier Excel)

Norme ISO 8601

Mahdi Yusuf / @myusuf3 https://twitter.com/myusuf3/status/865722106071453696

XKCD, ISO 8601 https://xkcd.com/1179/

Organisation des données



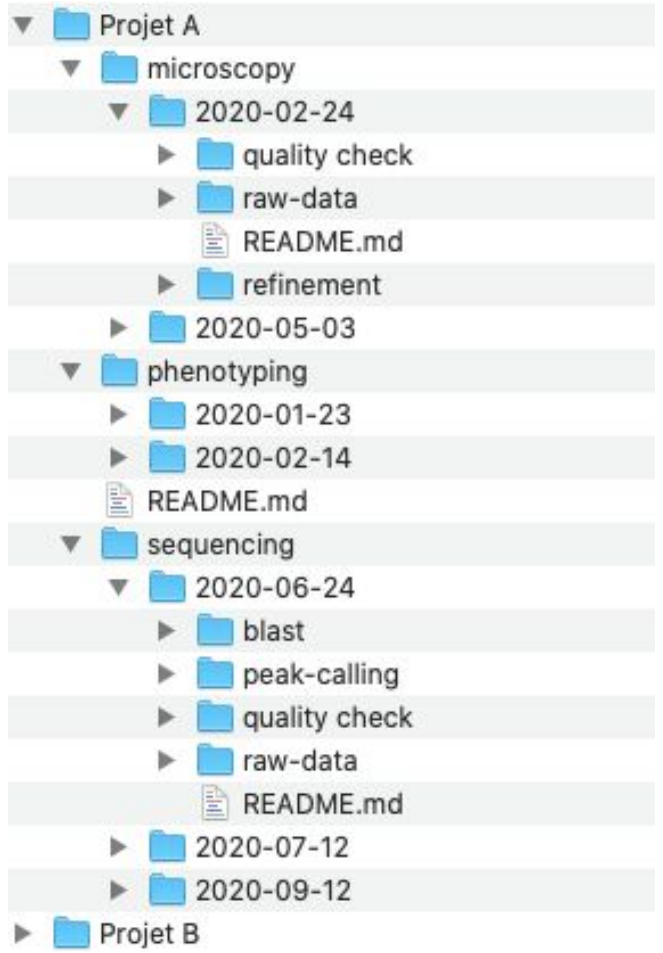
Une proposition d'organisation

Organisation des dossiers

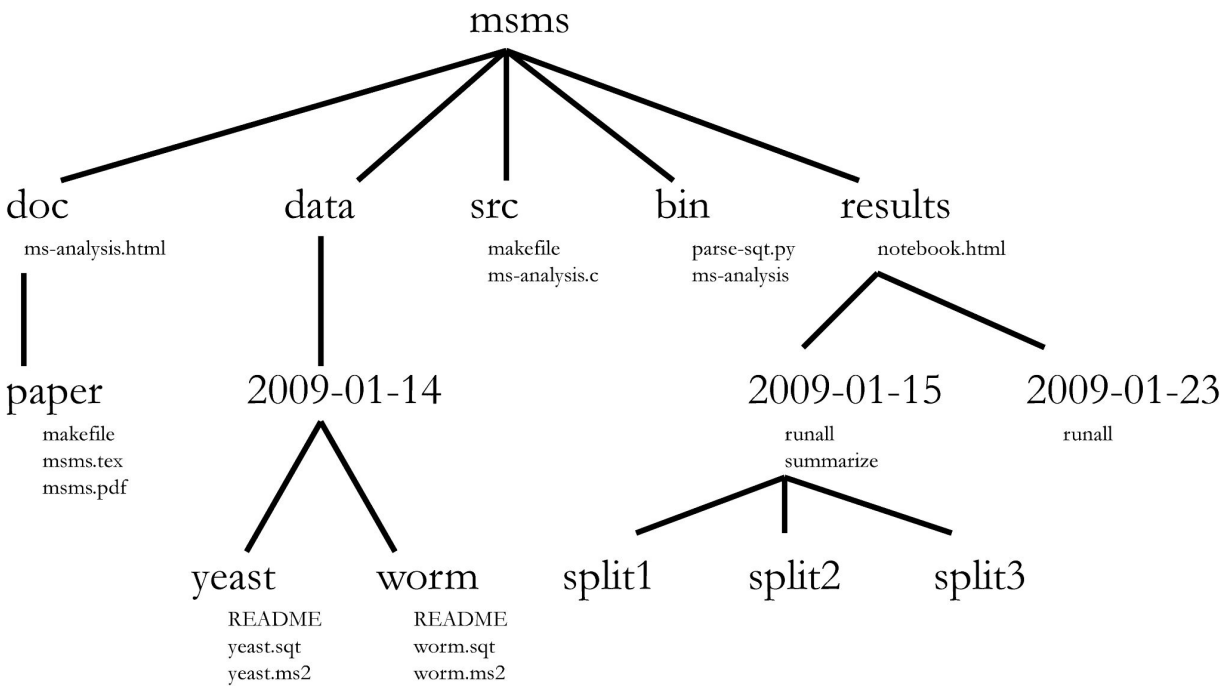
- Limitez le nombre de dossiers par niveau (5 ou 6 max)
- Allez du général au spécifique
- Choisissez des noms de dossiers explicites

Pour chaque dossier, ajoutez un fichier README:

- Choisissez un format simple et ouvert (par exemple Markdown ou TXT)
- Indiquez un minimum de métadonnées 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



D’autres exemples



Noble, PLoS Comput Biol, 2009

DOI 10.1371/journal.pcbi.1000424

Box 3. Project layout

```
.
|--CITATION
|--README
|--LICENSE
|--requirements.txt
|--data
|    |--birds_count_table.csv
|--doc
|    |--notebook.md
|    |--manuscript.md
|    |--changelog.txt
|--results
|    |--summarized_results.csv
|--src
|    |--sightings_analysis.py
|    |--runall.py
```

Wilson, PLoS Comput Biol, 2017

DOI 10.1371/journal.pcbi.1005510

Deux références

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble[1,2]*

1 Department of Genome Sciences, School of Medicine, University of Washington, Seattle, Washington, United States of America, 2 Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America

Noble, PLoS Comput Biol, 2009

DOI 10.1371/journal.pcbi.1000424

PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

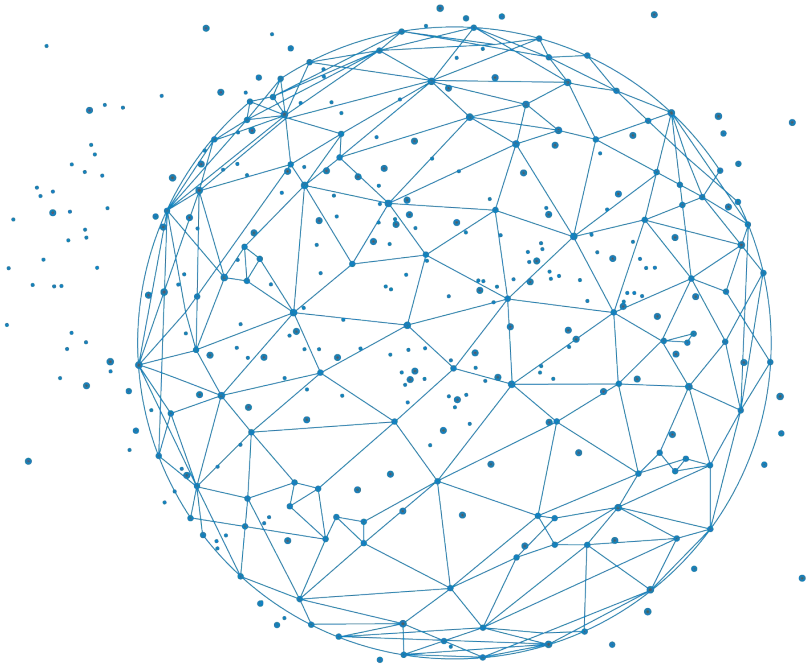
Good enough practices in scientific computing

Greg Wilson[1☯*], **Jennifer Bryan**[2☯], **Karen Cranston**[3☯], **Justin Kitzes**[4☯], **Lex Nederbragt**[5☯], **Tracy K. Teal**[6☯]

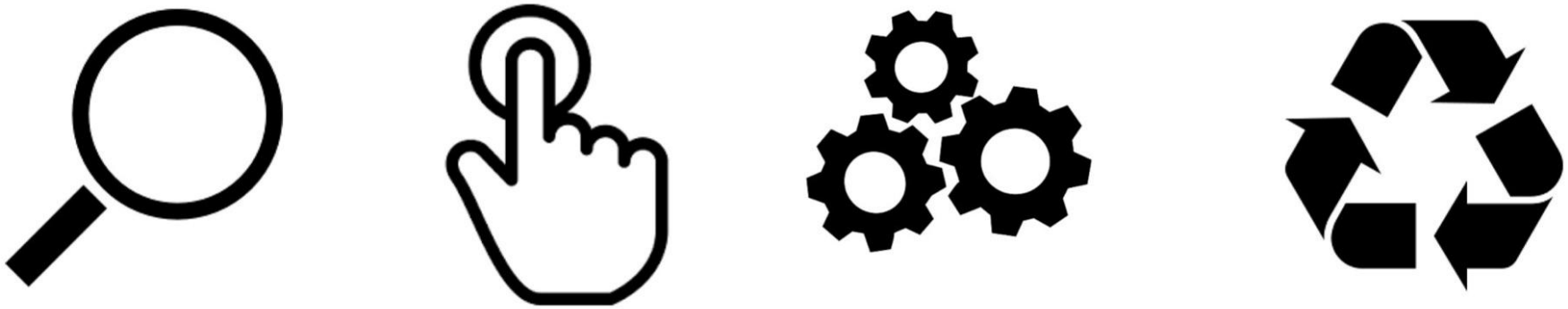
Wilson, PLoS Comput Biol, 2017

DOI 10.1371/journal.pcbi.1005510

Format de fichier



Findable Accessible Interoperable Reusable



Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- ☐ Data is provided in commonly understood and preferably open formats
- ☐ The metadata provided follows relevant standards
- ☐ Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- ☐ Qualified references and links are provided to other related data

1990 **2022**

Recommandations sur le format des fichiers

Privilégiez les formats ouverts afin de faciliter le partage des données
Définition légale du **format ouvert** en France (loi no 2004-575 du 21 juin 2004) :
On entend par standard ouvert tout protocole de communication, d'interconnexion ou d'échange et tout format de données ***interopérable*** *et dont les* ***spécifications techniques sont publiques*** *et* ***sans restriction d'accès*** *ni de* ***mise en œuvre***.
-> format bien documenté et utilisable sans demander d'autorisation

Format ouvert	Format fermé
Spécifications publiques et gratuites	Spécifications non publiques
Aucune restriction légale pour l'utiliser	Des restrictions légales s'opposent à son utilisation (droit d'auteur, copyright, brevet)
Format indépendant du logiciel utilisé qui assure l'interopérabilité des données	Format lisible qu'avec un logiciel particulier
Maintenu par une organisation à but non lucratif	Format propriétaire

Type	Format conseillé	Format non conseillé
Document texte	PDF, TXT, ODT	MS Word, RTF
Feuille de calcul	ODS, CSV	MS Excel, PDF, OOXML
Base de données	SQL, SIARD, DB tables (.CSV)	MS Access, dBase (.dbf), HDF5
Données statistiques	SPSS Portable, STATA, XML, CSV, TXT	SAS et R
Images	JPEG, TIFF, PNG	DICOM
Audio	BWF, MXF, Matroska (.mka), FLAC, OPUS	WAVE, MP3, AAC, AIFF, OGG
Video	MXF, MKV	MPEG-4, MPEG-2, AVI, QuickTime (.mov, .qt)
Information géographique	GML, MIF/MID	ESRI Shapefiles, MapInfo, KML
Images géoréférencées Raster	GeoTIFF (.tif, .tiff) ASCII GRID (.asc, .txt)	TIFF World File ESRI GRID

https://facile.cines.fr/ service de validation des formats

En pratique, on peut souvent travailler avec un format fermé populaire et le **convertir** en format ouvert.
Mais il faut vérifier si la conversion altère les informations, et prendre des mesures de compensation si nécessaire.
Ex : la conversion XLSX -> CSV perd les mises en forme.

http://dotwhat.net

0% → 100%

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

Format
Propriétaire fermé | Propriétaire ouvert | Ouvert

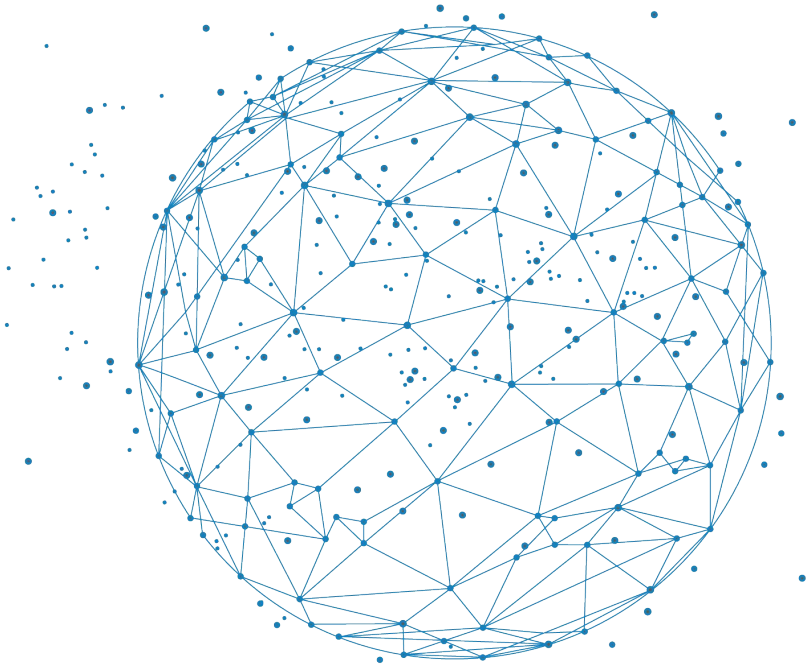
Format
En évolution | Stable

Description
Pas de schema(.org) | schema

Langage du format
Propriétaire | Norme

Formalisation
Pas de norme | Norme propriétaire | Norme Iso

Gestion des codes sources et des packages



Avec Claire !

Conclusion

Un projet c’est long !

Bonnes pratiques à garder en tête

- Stockage des données
- Un environnement de travail sûr
- Le nommage et le formats des fichiers
- Organisation des données
- Protéger ses données

Outils et solutions