

# Usecase 1

## Organiser son projet

H. Chiapello et T. Denecker

Slides inspirées de : F.de Lamotte (Montpellier), J. Seiler (Strasbourg), C. Midoux & V. Loux (Jouy-en-Josas), D. Benaben (Bordeaux)



**Bien s'organiser:** organisation d'un projet, nommage des fichiers,...

Choisir les bons **Formats de fichiers**

Assurer la **Sécurité** : sauvegarde, archivage,...



# Bien s'organiser



Plusieurs personnes

Ne rien perdre

Plusieurs techniques

Pouvoir retrouver

Plusieurs lieux

Pouvoir réanalyser

Plusieurs années

Pouvoir partager



- **1 dossier par projet**
- Limitez le nombre de dossiers par niveau (5 ou 6 max)
- Allez du général au spécifique
- Choisissez des noms de dossiers explicites et non redondants

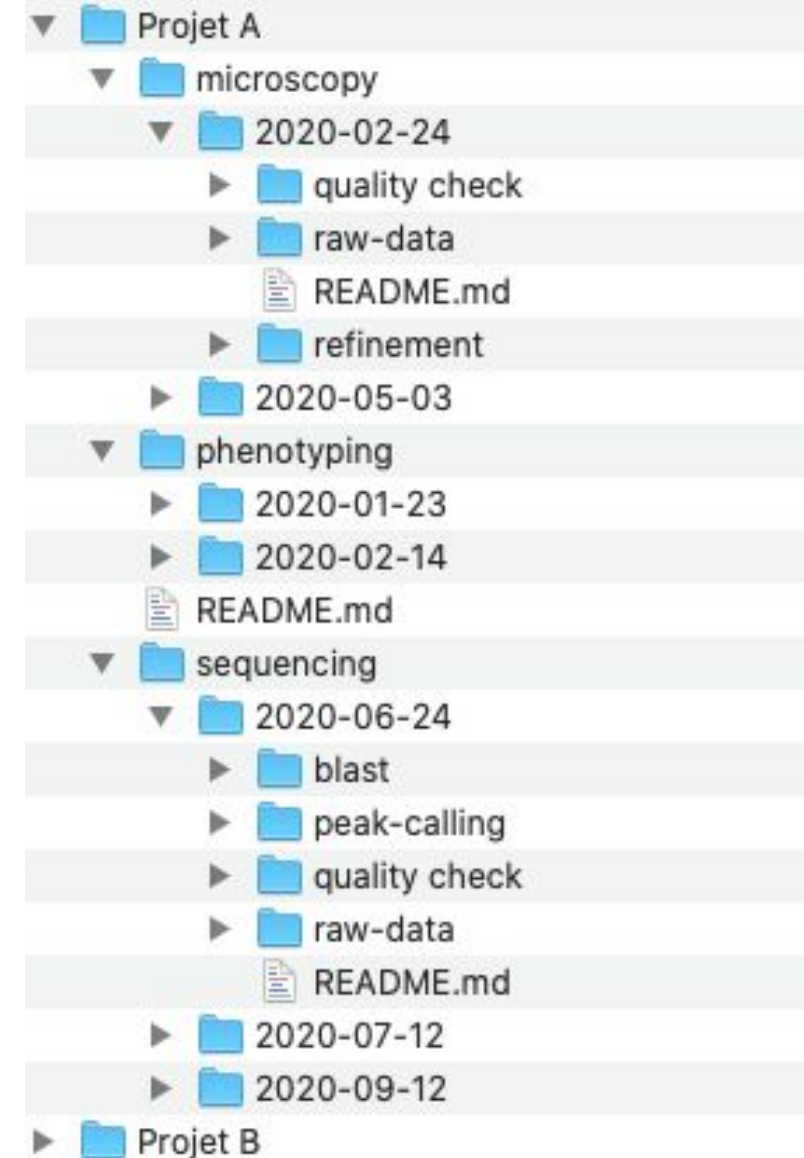
Pour chaque dossier, ajoutez un fichier **README**:

- Format simple et ouvert (par exemple Markdown ou TXT)
- Documentez le contenu du dossier :
  - 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

1. Un dossier par projet
  - a. Un sous-dossier par type de manip (microscopie, séquençage, phénotypage)
    - i. Un sous-dossier par date (2020-02-24, 2020-05-03)
      1. Un sous-dossier pour les données brutes
      2. Un sous-dossier par analyse (contrôle qualité, nettoyage statistique, raffinement)
    - ii. Un sous-dossier par publication
      1. Un lien symbolique vers chaque dossier données ou analyse associé à la publication

- Dispo sur zenodo :

<https://zenodo.org/record/4410128#.YjiRpDXjJD9>



# Un autre exemple

Il n'y a pas de solution miracle mais il est important de se mettre d'accord au sein du projet et de le documenter dans le **Plan de Gestion de Données**

- Exemple de générateur d'arborescence

<https://www.tiesdekok.com/folder-structure-generator/#>



## Data management tips



GOAL of good data management  
→ optimise the discovery & reuse of data

### Questions to ask yourself

Are my files organised in a way that I can easily find what I am searching for?  
What information would I need to understand and use my data in 20 years?  
Could others understand and use my data?

**Folder structure**

01\_Project\_name

- 01\_Organisation
  - 01\_Grant
  - 02\_Shipping
- 02\_Raw\_data
  - inhouse
    - 01\_Experiment\_X
      - 2021\_01\_22
      - 2021\_02\_12
    - public
      - database\_A
  - 03\_Analysis
    - 01\_Experiment\_X
      - 01\_Filtering
      - 02\_Figures
    - 02\_Pipeline\_A
      - 01\_Code → 01\_load\_data.ipynb
      - 02\_Pipeline → 01\_load\_data
        - tmp
        - store
        - out
      - 03\_Output
  - 04\_Literature
  - 05\_Writing
    - 01\_Abstract
    - 02\_Introduction
  - 06\_Archive

Readme.txt project description

experimental data can be sorted by date

number your main folders

keep your raw data separately

ref 1: pipeline structure

separate ongoing & complete work

**File naming**

major change minor change YYYY-MM-DD YYYY-YYYY

P05\_RNAseq-bat3\_v03-02\_20210121\_KH.csv

project number/ acronym\* describing name version\* date initials creator \*if applicable

**General naming tips for folders & files**

  - use unique, meaningful names
  - not too long (not >30-40 characters)
  - no spaces, dots, or special characters (\$%!&^()+=[:;~@)
  - hyphens (-) & underscores (\_) to separate elements

**Friendly Reminder**

Comment your code! 😊

**Metadata**

Which information is necessary to interpret, understand, and use a given dataset?

**readme.txt files**

can be used to describe projects, folders, and files

**Important information**

  - Who has created the data?
  - What is the content of the data?
  - Which questions have been answered?
  - When were the data created?
  - How were the data developed (methods)?
  - Why were the data developed?
  - With whom can the data be shared?

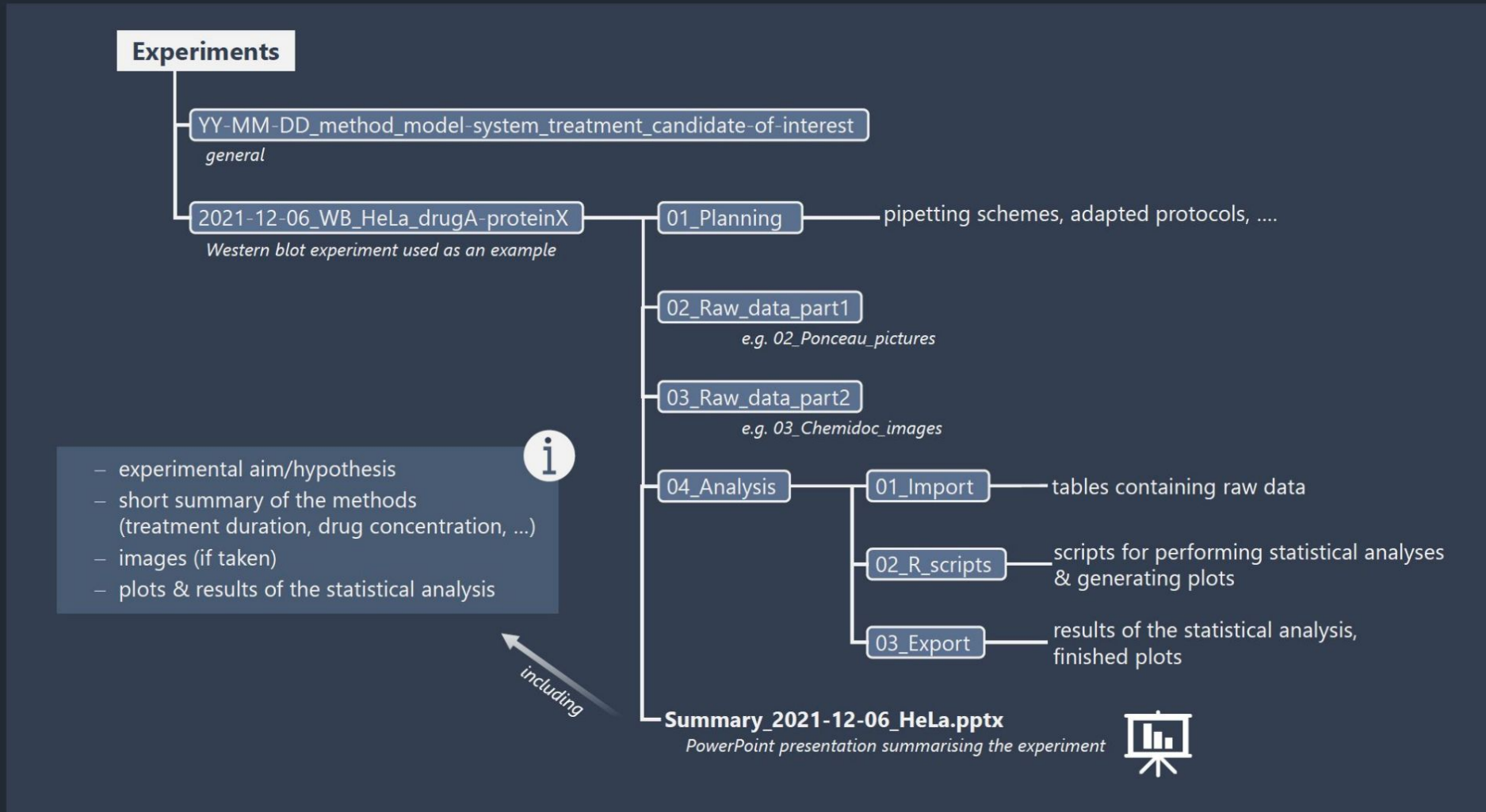
**References**

  - <https://towardsdatascience.com/how-to-keep-your-research-projects-organized-part-1-folder-structure-10bd56034d3a>
  - <https://www.wur.nl/en/Value-Creation-Cooperation/WDC/Data-Management-WDC/Doing/Organising-files-and-folders.htm>
  - <https://www.massey.ac.nz/massey/research/library/library-services/research-services/manage-data/organise.cfm>
  - <https://library.bath.ac.uk/research-data/working-with-data/organising-data>
  - <https://www.helsinki.fi/en/research/organizing-data-folders-with-5sdata-method>
  - <https://mantra.edina.ac.uk>
  - <https://old.dataone.org/education-modules>

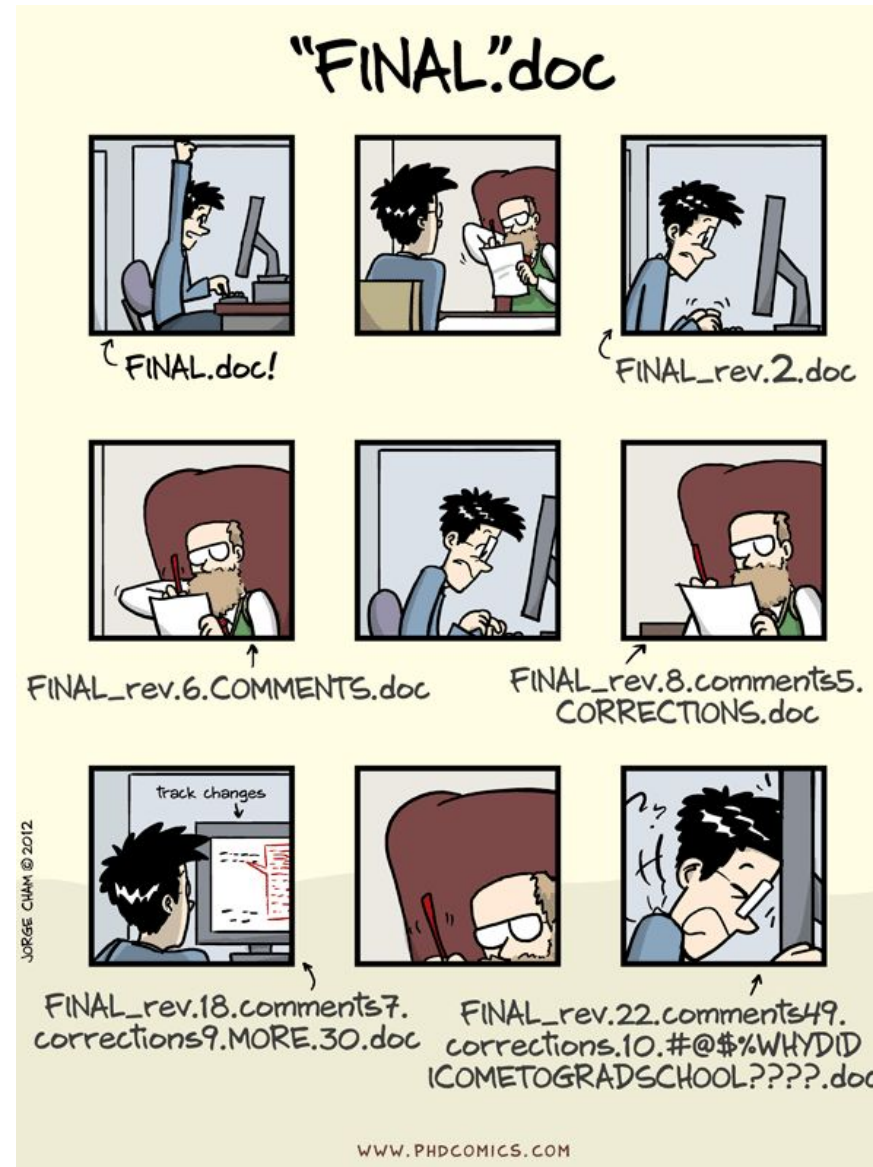
shallow deep



## Organising experimental data









Question : comment écrivez-vous une date ?



# Format ISO AAAA-MM-JJ

(2022-11-08)  
(2022-11-08T15:32:04)  
(2022-11-08T15:32:04-05:00)


## PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13  
20130227 2013.02.27 27.02.13 27-02-13  
27.2.13 2013. II. 27. 27/2-13 2013.158904109  
MMXIII-II-XXVII MMXIII <sup>LVII</sup>/<sub>CCCLXV</sub> 1330300800  
 $((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$  2013  
10/11011/1101 02/27/20/13  $\begin{matrix} 0 & 1 & 2 & 3 & 4 \\ & & 5 & 6 & 7 & 8 \end{matrix}$  

- **Pas d'espace, pas de point, ni de caractères spéciaux** (& / + > : ? % \* ...)

Utiliser des tirets (-) ou underscores (\_) pour séparer les éléments



Règles dénomination fichiers ❌



ReglesDenominationFichiers ✅

- **Dates au format AAAAMMJJ ou AAAA-MM-JJ** (année, mois, jour)



20150405\_CR



20160310\_CR



20160515\_CR

- **Versionnez** (cf GitLab)



Convention\_V01



Convention\_V02



Convention\_VF

- **Pas trop long**



cenomestpeutetreexplicitemaisprobablementbeau  
couptroplongetencorecapourraitetrepirecarilmanqu  
eladateetlalistedetouslesauteurs\_vf\_2.txt

- **Rangez**



Reunion



20150407\_CR



20150407\_Minutes



20150407\_OJ

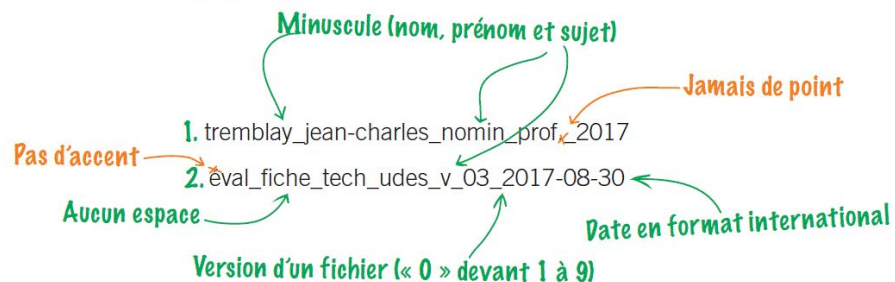
- **Et documentez vos règles**

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-Archdét)	Direction/Service transversal(e): CSSI
Responsable(s) de la mise en œuvre: Archivistes de département et d'institution	Approbateur : Collège spécialisé Systèmes d'Information
Date: 21.11.2012	Date: 21.11.2012 /mise à jour de l'annexe : décembre 2015

# Exemple de règles de nommage



## EXEMPLES TYPES DE NOM D'UN FICHIER



Éléments	Règle	Exemple
<b>Sujet</b>	<b>Obligatoire</b> Il s'agit du sujet principal traité au sein du document. Utiliser des noms communs, écrits en lettres minuscules non accentuées.	projet formation évaluation
<b>Séparateur</b>	Les espaces sont interdits. Utiliser l'underscore (touche 8 du clavier) pour remplacer les espaces	« _ »
<b>Type de document</b>	<b>Facultatif</b> Qualifie la nature du document. Toute abréviation sera en lettres majuscules.	(CR) compte rendu (OJ) ordre du jour
<b>Date</b>	<b>Obligatoire</b> Date de création du document, date de l'événement. Format à l'américaine : AAAAMMJJ. Nomme d'une période : utilisation d'un séparateur « _ » ou « - ».	20180122 201608 2010 201501_07 ou 201501-07
<b>Version du document</b>	<b>Obligatoire</b> 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_CFUU_V0.0 CR_CFUU_V0.1 CR_CFUU_VP, VF ou VV
<b>Extension</b>	<b>Obligatoire</b> 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)

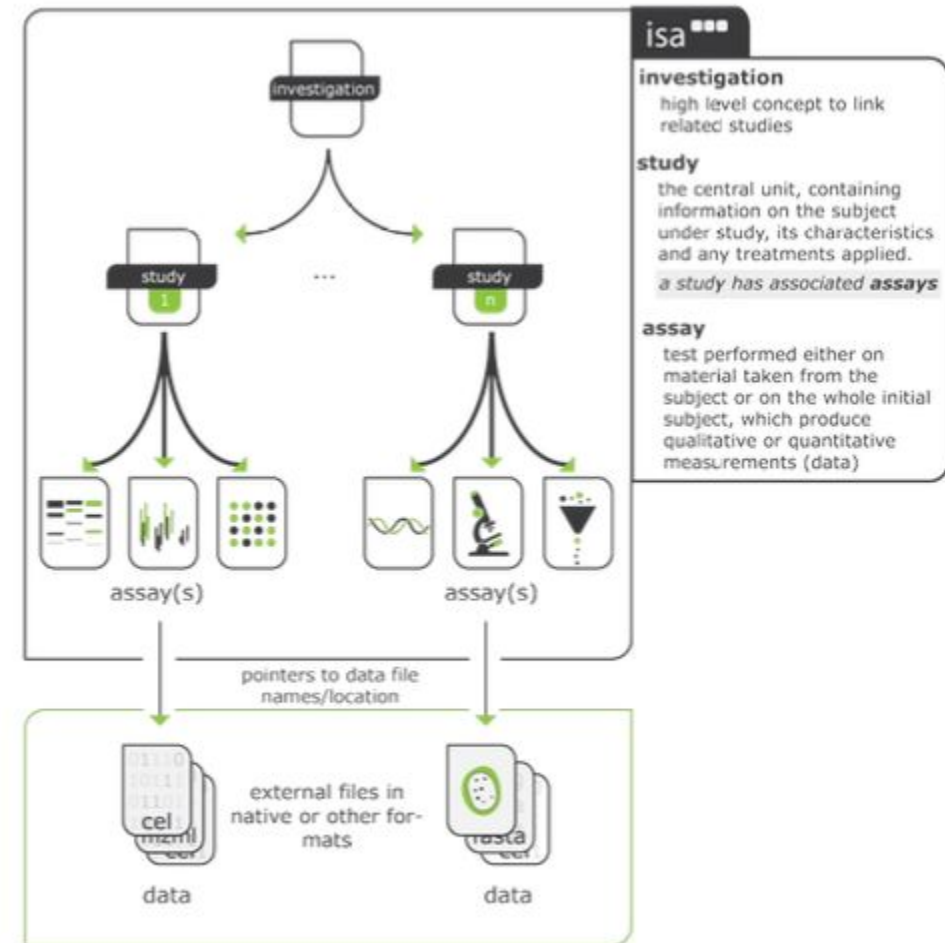


## ISA

### A standard for Life Science Data

Capture experimental metadata through 3 core entities:

- **Investigation:** the project context
- **Study:** an experimentation in one location
- **Assay:** a specific measurement that targets a trait with a method and a scale



# Choisir les bons formats de fichier





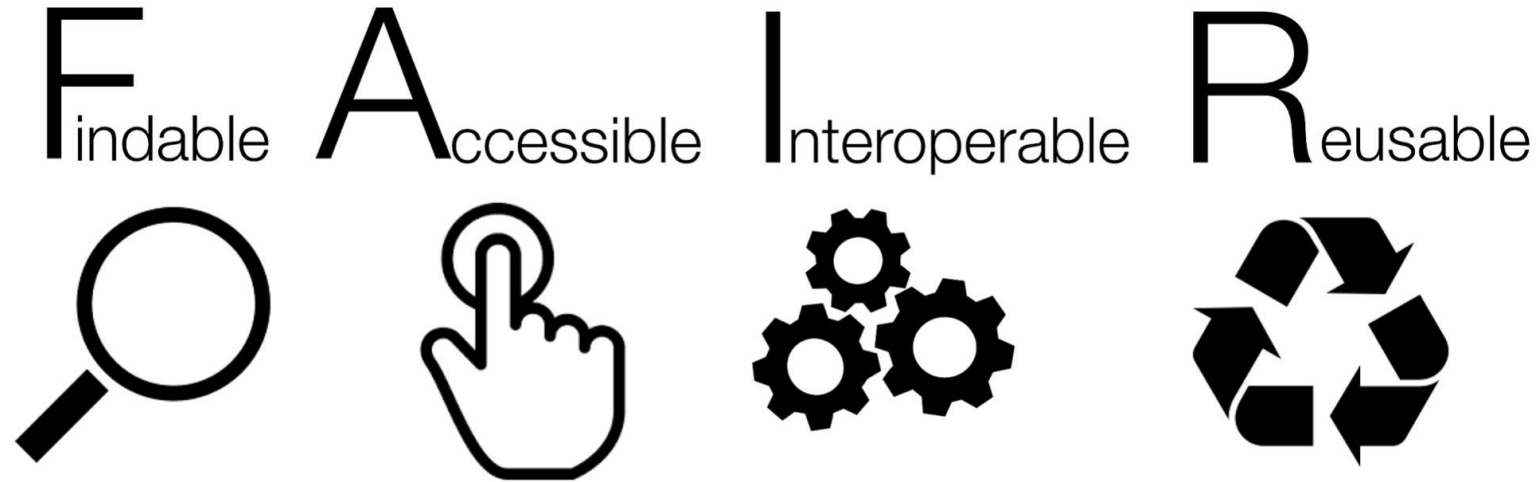
## Formats et logiciel ?

Allez à : [http://scrumblr.ca/fair\\_data\\_format](http://scrumblr.ca/fair_data_format)

et listez les formats que vous connaissez dans les colonnes appropriées

Texte brute  
Texte formaté  
Données numériques  
Tableurs & Stats  
Image (matriciel, vectoriel, 3D)  
Son & Vidéo  
Cartographie  
Bioinfo  
Code & Script  
Archive & compression





## 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



Des logiciels spécifiques sont-ils nécessaires pour traiter les formats cités ?

Fonctionnent-ils en ligne ou après installation sur un ordinateur ?

Fonctionnent-ils avec un système d'exploitation particulier (Windows, Mac, Linux) ?

Sont-ils liés à un type d'ordinateur ou à un instrument particulier (ex : microscope) ?

Sont-ils gratuits ou payants ? Qui paye ?

S'ils n'existaient plus ou si vous n'y avez plus accès, pourriez-vous continuer à travailler ?

L'éditeur du logiciel (ou la communauté) est-il en bonne santé ?

Le logiciel sera encore disponible dans 20 ans et sera-t-il encore capable d'interpréter correctement le fichier ?

**⇒ Que proposez-vous pour garantir la pérennité de l'accès à vos données ?**



Enjeu pour la préservation et l'exploitation des données

## Formats « textuels »

- Suite d'octets représentant des caractères imprimables et affichables à l'écran
- Peuvent être lus dans un éditeur de texte
- Mais souvent besoin d'un logiciel spécifique pour interpréter la structure interne, matérialisée par certains caractères, et en donner une représentation informatique exploitable
- Caractères ordinaires + caractères ayant une valeur spéciales : < > / \* etc.



**HTML** (avec des “balises”) conçu principalement pour représenter les pages web.

Contenu lisible dans un éditeur texte :

```
<html>
<head><head>
<body>
<p>Bonjour <span style='color:red'>tout le monde</span></p>
</body>
</html>
```

Mais « interprétable » par un logiciel dédié (navigateur web) :

Bonjour **tout le monde**



## Contenu lisible dans un éditeur texte

```
{\rtf1\adeflang1025\ansi\ansicpg1252\uc1\adef0\deff0\stshfdbch37\stshf1och37\stshfhich37\stshfbi0\deflang1036\deflangfe1036\themelang1036\thelangfe0\themelangcs0{\fonttbl{\f0\fbidi \froman\fcharset0\fpqrq2{\*\panose 02020603050405020304}Times New Roman;}{\f34\fbidi \froman\fcharset0\fpqrq2{\*\panose 02040503050406030204}Cambria Math;}\mlMargin0\mrMargin0\mdefJc1\mwrapIndent1440\mintLim0\mnaryLim1}{\info{\author Mathieu Saby}{\operator Mathieu Saby}{\creatim\yr2018\mo6\dy10\hr13\min44}{\revtim\yr2018\mo6\dy10\hr13\min44}{\version2}{\edmins1}{\nofpages1}{\nofwords3}{\nofchars19}\fs24\lang1036\langfe1033\loch\af37\hich\af37\dbch\af37\cgrid\langnp1036\langfenp1033 {\rtlch\fcs1 \af0 \ltrch\fcs0 \insrsid16651434 \hich\af37\dbch\af37\loch\af37 Bonjour }{\rtlch\fcs1 \af0 \ltrch\fcs0 \cf6\insrsid16651434\charrsid16651434}\hich\af37\dbch\af37\loch\af37 tout le monde }{\rtlch\fcs1 \af0 \ltrch\fcs0 \insrsid16651434
```

Mais uniquement interprétable avec Word, Libre office ou autre traitement de texte



- Langage de balisage léger
- Facile à lire et à écrire
- Facilement interprétable avec de nombreux éditeurs
- Peut être lu en l'état sans donner l'impression d'avoir été balisé ou formaté par des instructions particulières

```
# Avec `markdown`  
  
Avec `markdown` on peut simplement mettre un mot  
en italique ou en gras,  
voir même le barrer  
  
Wikipédia a une  
[page] (https://fr.wikipedia.org/wiki/Markdown) sur  
`Markdown`.  
  
Bien sur, d'autres langages de balisage léger  
existent comme :  
  
- `reStructuredText` ;  
- `Org-mode` ;  
- `Wikitexte`.
```

export



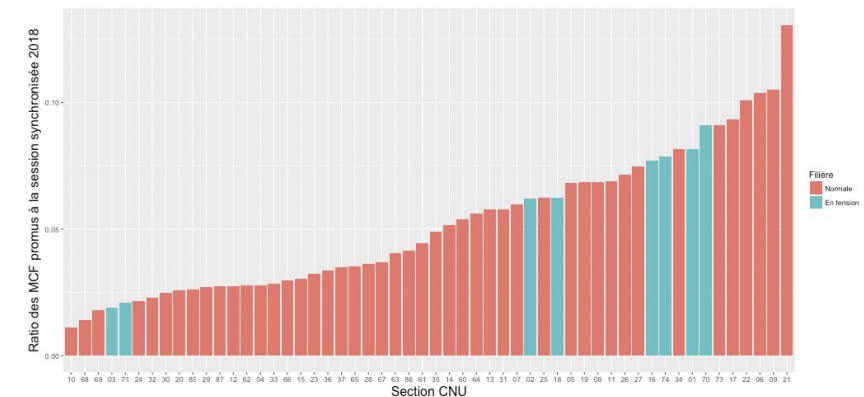
```
markdown.md - Typora  
Fichier Éditer Paragraphe Format Présentation(V) Thèmes Aide(H)  
  
Avec markdown  
  
Avec markdown on peut simplement mettre un mot en italique ou en gras,  
voir même le barrer  
  
Wikipédia a une page sur Markdown.  
  
Bien sur, d'autres langages de balisage léger existent comme :  
  
• reStructuredText ;  
• Org-mode ;  
• Wikitexte.  
  
41 Mots
```

- Suite d'octets non interprétables comme des caractères imprimables ou affichables
- Structure interne opaque
- Besoin de logiciel spécifique pour les lire et les interpréter

## Exemple de format binaire : **PNG** (image)

- Contenu illisible dans un éditeur texte (à part «?PNG » au début)
- Uniquement lisible et interprétable avec un visionneur d'images :

```
?PNG  
  
?V4?????6n?I6?"?d??θ??83???OEP|1?L?? (??>?/?  
%?? (>???P苦?;3?i???e?|??{?g?蹟X????-2?s???=+?????WQ+]?L6O  
  
w?[?C?{_???????F qb??  
  
????U?vz?????Z?b?1@?/z??c??s>~?if?,?HUS  
  
j???????F
```





- 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) :
  - 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

Note : pour que ces organisations non-lucratives puissent perdurer, il est nécessaire et important de contribuer régulièrement





Type	Formats conseillés	Formats non conseillés
Document texte	PDF, TXT, ODT	MS Word, RTF
Feuille de calcul	CSV, ODS	MS Excel, PDF, OOXML
Base de données	SQL, SIARD, DB tables (.CSV)	MS Access, dBase (.dbf), HDF5
Données statistiques	CSV, XML, TXT, RData (suivant les versions)	SAS, RData (suivant les versions)
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	GeoTIFF (.tif, .tiff)	TIFF World File
Raster	ASCII GRID (.asc, .txt)	ESRI GRID



## File formats for digital content: Probability for full long-term preservation



Content type	High	Medium	Low
Text	<ul style="list-style-type: none"> <li>Plain text (encoding: USASCII, UTF-8, UTF-16 with BOM)</li> <li>XML (includes XSD/XSL/XHTML, etc.; with included or accessible schema)</li> <li>PDF/A-1 (ISO 19005-1) (*.pdf)</li> </ul>	<ul style="list-style-type: none"> <li>Cascading Style Sheets (*.css)</li> <li>DTD (*.dtd)</li> <li>Plain text (ISO 8859-1 encoding)</li> <li>PDF (*.pdf) (embedded fonts)</li> <li>Rich Text Format 1.x (*.rtf)</li> <li>HTML (include a DOCTYPE declaration)</li> <li>SGML (*.sgml)</li> <li>Open Office (*.sxw/*.odt)</li> <li>OOXML (ISO/IEC DIS 29500) (*.docx)</li> <li>Microsoft Word 2007 or newer (*.docx)</li> </ul>	<ul style="list-style-type: none"> <li>PDF (*.pdf) (encrypted)</li> <li>Microsoft Word 2003 or older (*.doc)</li> <li>WordPerfect (*.wpd)</li> <li>DVI (*.dvi)</li> <li>All other text formats not listed</li> </ul>
Raster image	<ul style="list-style-type: none"> <li>TIFF (uncompressed)</li> <li>JPEG2000 (lossless) (*.jp2)</li> <li>PNG (*.png)</li> </ul>	<ul style="list-style-type: none"> <li>BMP (*.bmp)</li> <li>JPEG/JFIF (*.jpg)</li> <li>JPEG2000 (lossy) (*.jp2)</li> <li>TIFF (compressed)</li> <li>GIF (*.gif)</li> <li>Digital Negative DNG (*.dng)</li> </ul>	<ul style="list-style-type: none"> <li>MrSID (*.sid)</li> <li>TIFF (in Planar format)</li> <li>FlashPix (*.fpx)</li> <li>PhotoShop (*.psd)</li> <li>RAW</li> <li>JPEG 2000 Part 2 (*.jpf, *.jpx)</li> <li>All other raster image formats not listed</li> </ul>
Vector graphics	<ul style="list-style-type: none"> <li>SVG (no Java script binding) (*.svg)</li> </ul>	<ul style="list-style-type: none"> <li>Computer Graphic Metafile (CGM, WebCGM) (*.cgm)</li> </ul>	<ul style="list-style-type: none"> <li>Encapsulated Postscript (EPS)</li> <li>Macromedia Flash (*.swf)</li> <li>All other vector image formats not listed</li> </ul>
Audio	<ul style="list-style-type: none"> <li>AIFF (96kHz 16bit PCM) (*.aif, *.aiff)</li> <li>WAV (96kHz 24bit PCM) (*.wav)</li> </ul>	<ul style="list-style-type: none"> <li>SUN Audio (uncompressed) (*.au)</li> <li>Standard MIDI (*.mid, *.midi)</li> <li>Ogg Vorbis (*.ogg)</li> <li>Free Lossless Audio Codec (*.flac)</li> <li>Advance Audio Coding (*.mp4, *.m4a, *.aac)</li> <li>MP3 (MPEG-1/2, Layer 3) (*.mp3)</li> </ul>	<ul style="list-style-type: none"> <li>AIFC (compressed) (*.aifc)</li> <li>NeXT SND (*.snd)</li> <li>RealNetworks 'Real Audio' (*.ra, *.rm, *.ram)</li> <li>Windows Media Audio (*.wma)</li> <li>Protected AAC (*.m4p)</li> <li>WAV (compressed) (*.wav)</li> <li>All other audio formats not listed</li> </ul>
Video	<ul style="list-style-type: none"> <li>Motion JPEG 2000 (ISO/IEC 15444-4)??*.mj2)</li> <li>AVI (uncompressed/native, motion JPEG) (*.avi)</li> <li>QuickTime Movie (uncompressed/native, motion JPEG) (*.mov)</li> </ul>	<ul style="list-style-type: none"> <li>Ogg Theora (*.ogg)</li> <li>MPEG-1, MPEG-2 (*.mpg, *.mpeg, wrapped in AVI, MOV)</li> <li>MPEG-4 (H.263, H.264) (*.mp4, wrapped in AVI, MOV)</li> </ul>	<ul style="list-style-type: none"> <li>AVI (others) (*.avi)</li> <li>QuickTime Movie (others) (*.mov)</li> <li>RealNetworks 'Real Video' (*.rv)</li> <li>Windows Media Video (*.wmv)</li> <li>All other video formats not listed</li> </ul>



La documentation d'un format peut devenir une norme officielle nationale ou internationale ou un standard de facto

- une norme est établie par un organisme reconnu, comme l'ISO ou l'AFNOR), qui fournit des règles. On a donc une garantie de stabilité et de pérennité
- un standard est établi par un groupe privé, pour assurer une cohérence des échanges a un moment donné.

Il existe des cas de standards qui sont devenus des normes, par exemple :

- PDF, standard Adobe devenu PDF/A1 norme ISO (ISO 19005)
- Les formats LibreOffice (ODS, ODT...) sont standardisés (ISO/IEC 26300)
- PNG, standard W3C devenu norme ISO (ISO 15948)
- Le format CSV est décrit dans la RFC 4180 (mais il s'agit plus d'un document indicatif que d'une norme, plusieurs versions existent)
- Les formats bureautique Microsoft Office (XLSX, DOCX...) sont standardisés (ISO/IEC 29500). Mais les logiciels semblent parfois s'écarter du standard
-



- Utiliser des **formats ouverts** afin de faciliter le partage et l'interopérabilité
- Si on utilise des formats fermés, il faut vérifier s'assurer de la **pérennité** et si la conversion altère les informations
- Le format doit être **documenté** (standards)
- Format textuel en **UTF-8** répond à la majorité des besoins :
  - excel → csv
  - word → txt/md/html
- Mais aussi :
  - zip → tar.gz
  - images → png/svg

[https://fr.wikipedia.org/wiki/Format\\_ouvert#Les\\_principaux\\_formats\\_ouverts](https://fr.wikipedia.org/wiki/Format_ouvert#Les_principaux_formats_ouverts)

# Sécurité 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



## Les **besoins courants** pour la gestion de données lors d'un projet de recherche

- Des espaces de stockages adaptés à vos données (données scientifiques, documents bureautiques, bases de données, code source)
- Des outils adaptés à la gestion des droits des collaborateurs
- Des solutions de publication et d'archivage des données



## 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'utilisateur	- 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





## Performance vs Sécurité :

- Une infrastructure de calcul nécessite une solution de stockage **performante** :
  - accès massivement parallèle aux données
  - disques rapides
- Pour gagner en performance, on désactive les mécanismes de **sécurité** :
  - Moins voir pas de snapshots
  - Pas de réplication
  - Pas de sauvegarde
- Pour gagner en sécurité, on réduit la performance

A capacité identique, le coût d'une infrastructure performante et d'une infrastructure sécurisé est le même





## Infrastructure de calcul ne rime pas toujours avec infrastructure de stockage

Security Search

**IFB Core Cluster Documentation**

- Quick start guide
- Logging in
- Job submission (Slurm) >
- Software environment >
- Data >
- Tutorials >
- Cluster description
- Security

**Backup**

There is no backup for the main storage.

Some snapshots are available to protect against deletion by error but only one by day and for 5 days.

All servers and services are deployed using Ansible (and configurations are under revision control).

Main infrastructure services are backed up.

### 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 :

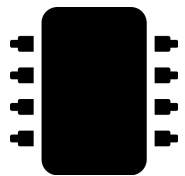
- consulter, corriger et améliorer les informations contenues sur le site pour toute question
- consulter les *notes de maintenance* sur le site web et sur les messages d'accueil des machines
- ne pas utiliser la machine comme espace de stockage ou de sauvegarde
- ne pas utiliser la machine comme passerelle 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 *scratches*, ...) disponibles dans la rubrique *techno-centre* du site web
- libérer les espaces *scratches* après leur utilisation
- communiquer avec l'équipe technique à l'adresse [romeo@univ-reims.fr](mailto: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 à des fins criminelles, de violation ou tentative



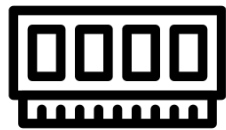
## National Network of Computing Resources

Une offre de service  
cloud et cluster couvrant  
l'ensemble du territoire Français

Cluster	Localisation du Data center	Coeurs	RAM (Go)	Stockage (To)
IFB Core	IDRIS - Orsay	5 042	26 542	2 000
Genotoul	Toulouse	6 128	34 304	3 000
ABiMS	Roscoff	2 608	10 600	2 500
GenOuest	Rennes	1 824	7 500	2 300
Migale	Jouy en Josas	2481	11 000	800
BiRD	Nantes	560	4 000	500



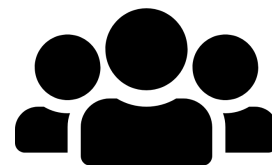
4300 coeurs



20 To RAM



2 Po



Une communauté  
d'entraide



Plus de 400  
outils



SSH  
Jupyter  
RStudio  
Galaxy



## Allocation des espaces de stockage par projet :

- 250 Go par projet, extensible sur demande argumentée
- Un projet peut être accessible à plusieurs utilisateurs
- Un utilisateur peut demander plusieurs espaces projet
- Pas de sauvegarde
- Stockage pour le calcul, **pas archivage**

## Bientôt disponible :

- Mise à disposition d'un espace scratch avec un quota plus important pour des besoins ponctuel (suppression automatique des fichiers les plus anciens)
- Sauvegarde des espaces projets



## Comment transmettre vos données de recherche ?

Pas adapté

Adapté

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



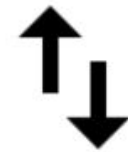
- Risque de perte
- Risque d'accès non autorisés
- Acceptable si les données sont chiffrées

Cloud privé



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

Service d'un  
consortium





**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 ?

[source](#)



## Vos ~~mots~~ phrases de passes (au pluriel)

HOW PASSWORD LENGTH WINS THE INTERNET

Passwords **102**



Password Length	Numerical 0-9	Upper & Lower case a-Z	Numerical Upper & Lower case 0-9 a-Z	Numerical Upper & Lower case Special characters 0-9 a-Z %\$
1	instantly	instantly	instantly	instantly
2	instantly	instantly	instantly	instantly
3	instantly	instantly	instantly	instantly
4	instantly	instantly	instantly	instantly
5	instantly	instantly	instantly	instantly
6	instantly	instantly	instantly	20 sec
7	instantly	2 sec	6 sec	49 min
8	instantly	1 min	6 min	5 days
9	instantly	1 hr	6 hr	2 years
10	instantly	3 days	15 days	330 years
11	instantly	138 days	3 years	50k years
12	2 sec	20 years	162 years	8m years
13	16 sec	1k years	10k years	1bn years
14	3 min	53k years	622k years	176bn years
15	26 min	3m years	39m years	27tn years
16	4 hr	143m years	2bn years	4qdn years
17	2 days	7bn years	148bn years	619qdn years
18	18 days	388bn years	9tn years	94qtn years
19	183 days	20tn years	570tn years	14sxn years
20	5 years	1qdn years	35qdn years	2sptn years

<https://www.thesecurityfactory.be/password-cracking-speed/>

At a current rate of 25\$ per hour, an AWS p3.16xlarge [...] we're capable of trying a whopping 632.000.000.000 different password combinations per second!

- Utilisez-vous un mot de passe différent pour chaque fournisseur de service ?
- Utilisez-vous un gestionnaire de mot de passe ?
  - bitwarden (Open Source)
  - Dashlane, LastPass, 1password...
- Renouvelez-vous vos mots de passe régulièrement ?
- Utilisez-vous une procédure sécurisée pour communiquer un mot de passe à vos collègues ? (par exemple pastebin.com)



## Optional Paste Settings

Syntax Highlighting:	<input type="text" value="None"/>
Paste Expiration:	<input type="text" value="Burn after read"/>
Paste Exposure:	<input type="text" value="Unlisted"/>
Folder:	<input type="text"/>
Password <span>NEW</span>	<input checked="" type="checkbox"/> Enabled <input type="text" value="iif5zL8zErFBehs6hfhjGr7djcbvhjre34v!"/>
	<input checked="" type="checkbox"/> Burn after read <span>NEW</span>
Paste Name / Title:	<input type="text" value="The root password"/>
<input type="button" value="Create New Paste"/>	





**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"  
mots de passe exportables*

## Gratuit

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

## Accessible

*Application Mac, Windows, Linux, Web, iPhone  
et Android*

1. Créer votre compte sur <https://bitwarden.com/>
2. Choisissez votre mot de passe maître (size matters)
3. Installer les applications sur vos appareils et les extensions de vos navigateurs
4. Enregistrer vos mots de passes dans votre coffre fort Bitwarden
5. **Activez l'authentification à 2 facteurs**

En plus :

- Générateur de mot de passe robuste intégré
- Analyse de vos mots de passes et reporting
- Partage de mots de passe entre collègue

[Quel gestionnaire de mots de passe choisir ? \(Les numériques\)](#)



## Situation :

Une partie de vos données est considérée comme données sensibles.

*Que mettez vous en place pour protéger l'accès à ces 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 peu e\*\*~~ peu soigneux, une erreur dans un script, ...)
  - un **transfert de données défectueux**
  - un plantage d'un disque dur
  - ...



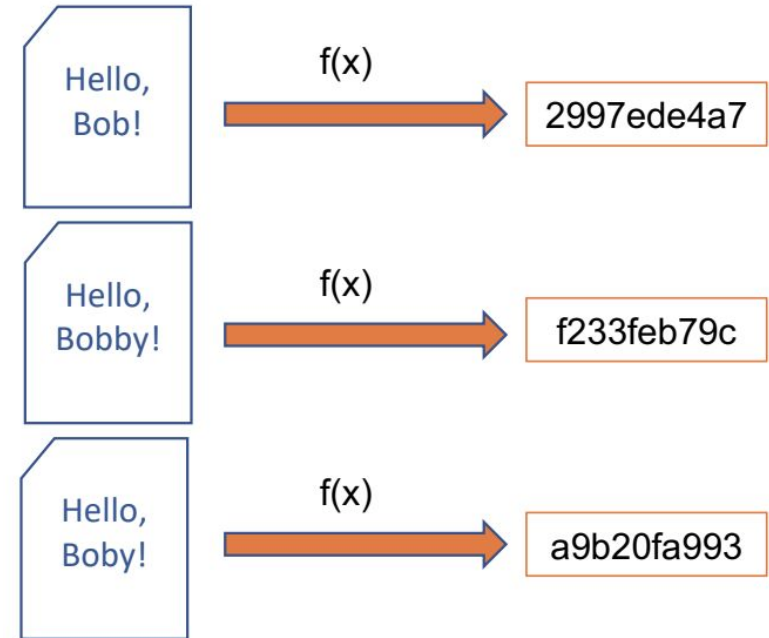
## Recommandation 1 : générer des sommes de contrôles

### Comment ?

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

### Quand ?

- **Avant un transfert de données**
  - Lorsqu'on réceptionne un nouveau jeu de données d'un collaborateur, d'une plateforme
  - Lorsqu'on transfert des données sur un stockage distant (exemple : EMBL ENA)
- **Stockage à long terme**
  - La version principale de chaque dataset
  - Les extraits de données utilisés dans les publications





## Recommandation 2 : utilisez le contrôle d'accès

### N'accordez que les permissions d'accès nécessaire :

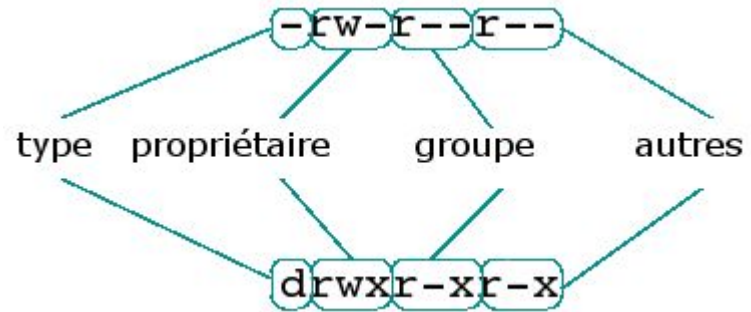
- Limitez le nombre d'utilisateurs ayant accès à vos données
- Limitez la visibilité des données (réseau interne vs internet)
- N'utilisez jamais de partage public sans chiffrement des données !

### Mettez les données brutes en lecture seule !!

L'accès aux données sensibles doit être **documenté**



## Gestion fine des droits au niveau UNIX



- Droits POSIX
- ACL

### Project mytest

#### Management console

New project member

Remove project member

exemple sur IFB



Limitez les copies au maximum !

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



## **Le stockage des données a un coût financier et écologique.**

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

## **Je veux garder mes données pour l'éternité**

- 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





## Où se situe mon fichier ?

0% ←—————→ 100%

	Moi	Mon équipe	Ma communauté	D'autres communautés	Le monde entier
<b>Lisible par</b>					
<b>Format</b>	Propriétaire fermé		Propriétaire ouvert		Ouvert
<b>Format</b>	En évolution				Stable
<b>Description</b>	Pas de schema(.org)				schema
<b>Langage du format</b>	Propriétaire				Norme
<b>Formalisation</b>	Pas de norme		Norme propriétaire		Norme ISO



Des questions ?





# De nombreux outils sont dispo :

- Google drive (ne faites pas ça !!)
- OSF
- Alfresco



# Avoir une GED c'est bien. Bien s'en servir, c'est mieux !

- Où sont localisé les serveurs ?
- Sauvegardes ?
- Qui gère les droits ?
- Comment est organisé la GED ?
- Edition synchrone et gestion des conflits ?
- Partage avec l'extérieur ?
- gestion des métadonnées
- workflow de validation

OSFHOME

My Quick Files My Projects Search Support Donate Frédéric de Lamotte

Gérez vos données de la recherche F... Files Wiki Analytics Registrations Contributors Add-ons Settings

Gérez vos données de la recherche - Formation

Public P C ...

Contributors: Jean-François Martin, alexandre dehne garcia, Frédéric de Lamotte, Victor REYS, Ionaz Vasquez-Villegas, Je'R'Rou, Urcei Kalenga, Jalbard Swann, Alexandre Benzart, GILLES Andre, Manam BARRO, Charlotte, Johanna Girodella, Germain Valentin Falty, Liyan OUYANG, chayma ben maamer

Date created: 2020-01-06 05:03 PM | Last Updated: 2020-06-26 03:29 PM

Category: Project

Description: Add a brief description to your project

Wiki

Page d'accueil

Par ordre alphabétique écrivez vos initiales suivies de votre prénom et nom

- ADG : Alexandre Dehne Garcia
- GVF : Germain Valentin Falty
- JFM : Jean-François Martin
- FZL : Frederic de Lamotte
- LKM : Charlotte Kinowski-Moysan

Read More

Citation

Components

Add components to organize your project.

Tags

academic training x data management k open data x open science x Add a tag

Recent Activity

- Frédéric de Lamotte linked Mendeley folder: DataViz to Gérez vos données de la recherche - Formation on 2020-06-26 03:29 PM
- Frédéric de Lamotte authorized the Mendeley addon for Gérez vos données de la recherche - Formation on 2020-06-26 03:29 PM
- Frédéric de Lamotte added addcn Mendeley to Gérez vos données de la recherche - Formation on 2020-06-25 01:54 PM
- Germain Valentin Falty updated wiki page Home to version 6 of Gérez vos données de la recherche - Formation on 2020-02-06 04:45 PM
- Frédéric de Lamotte updatec wiki page j'ai des problèmes ! to version 11 of Gérez vos données de la recherche - Formation on 2020-02-04 01:55 PM
- Frédéric de Lamotte updatec wiki page j'ai des problèmes ! to version 10 of Gérez vos données de la recherche - Formation on 2020-02-03 05:11 PM

Files

Click on a storage provider or drag and drop to upload

Filter i

Name	Modified
Gérez vos données de la recherche - Formation	
USB Storage (Germany - Frankfurt)	
02_AtelierIRcuqe.pptx	2020-01-22 04:49 PM
01-Cn route vers l'open science.pptx	2020-01-22 02:08 PM
02- les données de la recherche et l'open data.pptx	2020-01-22 02:06 PM
03- la Vie Des Donnees.pptx	2020-01-22 03:30 PM
04-PanGestionDonnees.pptx	2020-01-22 04:49 PM
05- gestion des données pendant le projet(2).pptx	2020-01-23 01:58 PM
07_Nommage_format.pptx	2020-01-23 01:59 PM
08_Metadonnées.pptx	2020-01-23 01:59 PM
09- Diffusez et partagez les données de recherche.pptx	2020-01-23 01:59 PM
10- Droit des données - Cas pratique.pptx	2020-01-24 09:13 AM
4_1_dmpLifeCycleMatrix.xlsx	2020-01-22 04:33 PM
entrepot_doc_parametres_synop_168.pdf	2020-01-23 11:29 AM

Mendeley

Enter citation style (e.g. "APA")

Citation

Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498(7453), 255-...



- Conseils d'hygiène numérique de l'ANSII
  - <https://www.ssi.gouv.fr/administration/bonnes-pratiques/>
- 7 recommandations cyber INRAE
  - <https://intranet.inrae.fr/cybersecurite>
- Ten simple rules for :
  - Digital Data Storage [10.1371/journal.pcbi.1005097](https://doi.org/10.1371/journal.pcbi.1005097)
  - providing effective bioinformatics research support [10.1371/journal.pcbi.1007531](https://doi.org/10.1371/journal.pcbi.1007531)
  - to make your computing more environmentally sustainable [10.1371/journal.pcbi.1009324](https://doi.org/10.1371/journal.pcbi.1009324)
  - 5 reasons why researchers should learn to love the command line [10.1038/d41586-021-00263-0](https://doi.org/10.1038/d41586-021-00263-0)
- <https://happygitwithr.com/>

# 7 recommandations

pour sécuriser sa pratique informatique

INRAE

## 1 Mots de passe : faites preuve d'imagination et diversifiez

Un bon mot de passe contient au minimum 12 caractères (minuscule, majuscule, chiffres et caractères spéciaux), et doit être propre à chaque compte utilisé. Pour les mémoriser, utilisez un gestionnaire de mots de passe.

- ✓ Choisir avec soin ses mots de passe  
<https://www.cnil.fr/fr/les-conseils-de-la-cnil-pour-un-bon-mot-de-passe>
- ✓ Utiliser le gestionnaire KeePass

## 2 Mises à jour : n'attendez plus

Les mises à jour corrigent les vulnérabilités appréciées des attaquants. Dès que cela vous est demandé, procédez à la mise à jour de sécurité de vos équipements : systèmes d'exploitation, applications mais aussi les composants.

Assurez-vous que l'Anti-Virus fonctionne et qu'il est bien à jour.

- ✓ Rapprochez-vous de votre informaticien ou de votre équipe informatique de centre pour plus d'informations en cas de besoin

## 3 Sauvegardes : l'atout sérénité

Veillez à ce que des sauvegardes régulières de vos données soient effectuées. Une fois les données perdues ou détruites, sans sauvegarde, il est trop tard...

- ✓ Utilisez les solutions de stockage ou de synchronisation fournies par l'Institut ou effectuez des sauvegardes régulières sur un support externe déconnecté comme par exemple un disque dur externe

## 4 Attention au clic de trop

**Messagerie** : méfiez-vous des apparences... Les pièces jointes ou les liens contenus dans certains courriels réservent parfois de mauvaises surprises. Les incohérences de fond ou de forme, comme les requêtes indiscrètes sont à prendre avec des pincettes

- ✓ Passez le curseur de votre souris sur le lien hypertexte, sans cliquer, l'adresse apparaîtra. Si le nom de domaine vous est inconnu, c'est probablement une tentative d'escroquerie

**Téléchargement** : Doutez de toute demande d'installation impromptue : mise à jour de flash, ajout d'extensions, plugins, mise à jour du lecteur vidéo... Ces sollicitations spontanées ne sont jamais sans arrière-pensée.

- ✓ Téléchargez le logiciel ou l'application uniquement depuis le site de l'éditeur du logiciel ou demandez conseil à votre informaticien



## 5 Séparation des usages : évitez la contagion

Pour limiter la propagation d'une action malveillante, séparez vos usages professionnels et personnels (messagerie, équipements...).

- ✓ N'utilisez pas votre messagerie professionnelle à des fins personnelles, pensez à créer des adresses différentes, en fonction de chaque usage
- ✓ N'utilisez pas de services de stockage en ligne personnels type « Cloud gratuit » à des fins professionnelles, ou du moins pas sans autorisation de l'Institut et sans avoir pris les mesures de sécurité qui s'imposent
- ✓ Ayez une utilisation appropriée d'Internet au travail et, d'une manière générale, soyez attentifs à vos propos sur les réseaux sociaux. Une fois sur Internet, vos données vous échappent et font le bonheur des adeptes de « l'ingénierie sociale » (usurpation d'identité, espionnage...). Pour en savoir plus et réagir en cas de problème <https://www.cnil.fr/fr/maitriser-mes-donnees>

## 6 Nomadisme : faites rimer mobilité et sécurité

En déplacement, attention et discrétion doivent guider vos usages : gardez vos appareils, supports et fichiers avec vous. Évitez d'utiliser les services en libre-service (bornes électriques, wifi gratuit d'hôtel...) ou les objets qui vous sont offerts (clés USB, goodies connectés...), ils peuvent avoir été configurés à des fins malveillantes.



- ✓ Lors de vos missions à l'étranger, utilisez de préférence et si possible un ordinateur dédié
- ✓ Plutôt que de se connecter sur la borne wifi d'un hôtel, partagez la connexion 4G de votre forfait mobile et donnez l'accès seulement à votre ordinateur
- ✓ Utilisez le service VPN de l'Institut qui permet de se raccorder à une connexion sécurisée et chiffrée et d'accéder aux ressources internes de l'INRA



## 7 En cas d'incident, quelques réflexes à avoir...

- si la machine a un « comportement étrange » pouvant laisser croire à un piratage ;
- si vous vous êtes fait voler votre matériel ou si vous l'avez perdu ;
- s'il y a eu un clic de trop, si vous avez donné votre mot de passe trop vite, si vous avez un doute...

- ✓ Rapprochez-vous immédiatement de votre informaticien ou de votre équipe informatique de centre
- ✓ Quoi qu'il arrive par précaution : déconnectez la machine du réseau pour stopper l'attaque, n'éteignez pas votre ordinateur et changez vos mots de passe

Pour en savoir plus ...

<https://intranet.inrae.fr/cybersecurite/>

