Gestion des données au cours du projet de recherche

3 heures

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

Julien Seiler - Strasbourg
@julozi

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Bonnes Pratiques

Un projet sur la durée (ré-intro)

La vie des données

Les principes FAIR

Stockage des données

Un environnement de travail sûr

Le nommage des fichiers

Les formats de fichiers

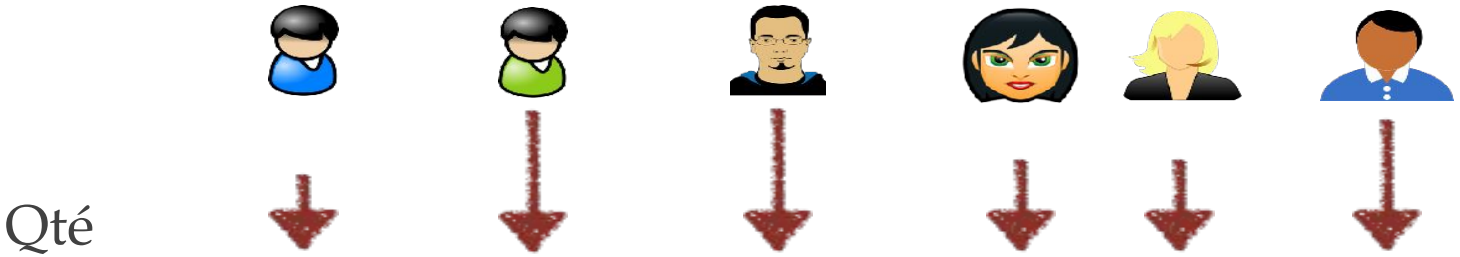
Organisation des données

Protéger ses données

La suppression des données

Outils et solutions

Rappel : un projet sur la durée

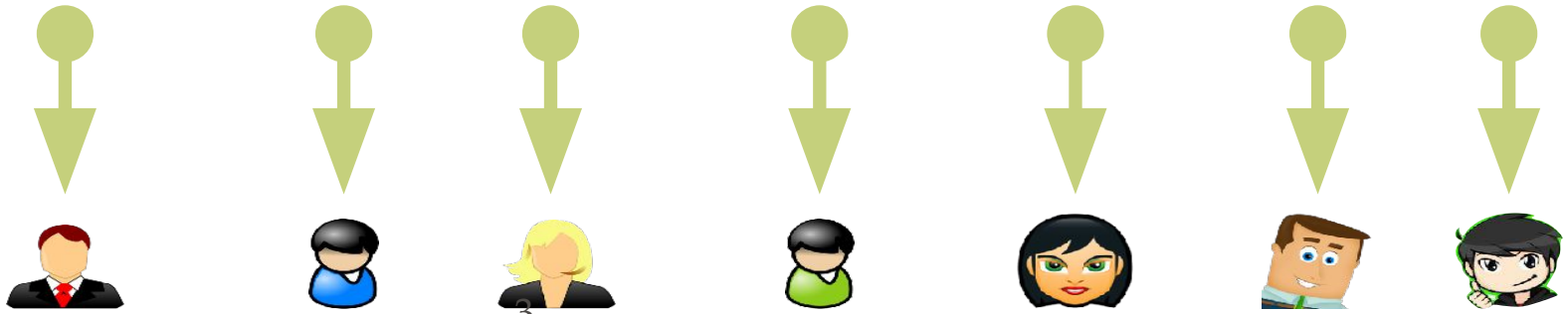


Qté

Champ Labo Plateforme Ordi Cluster Ordi Rédaction

Publication
DataPaper
Syst Info

Tps




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

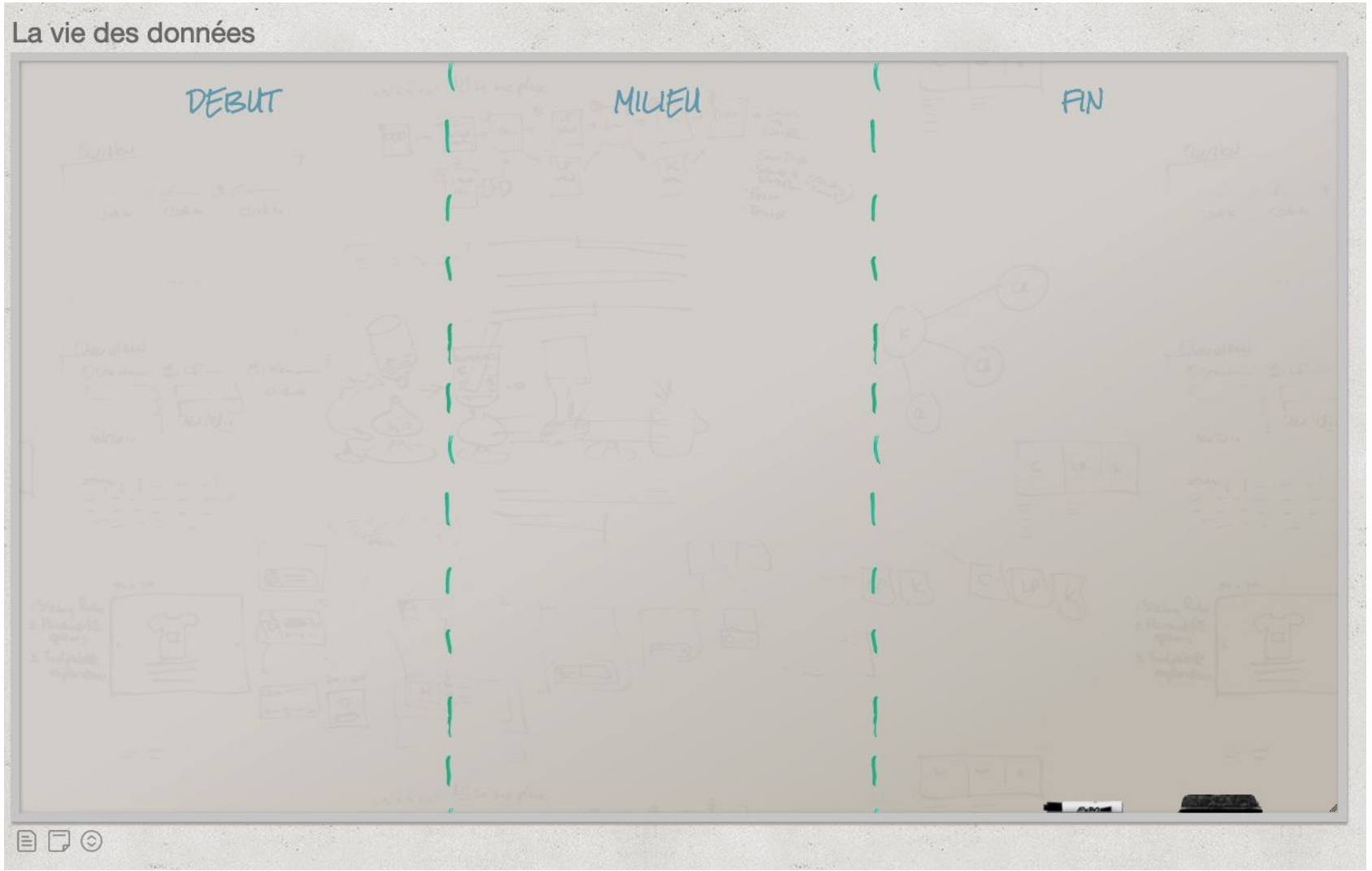
Pouvoir réanalyser

Pouvoir partager

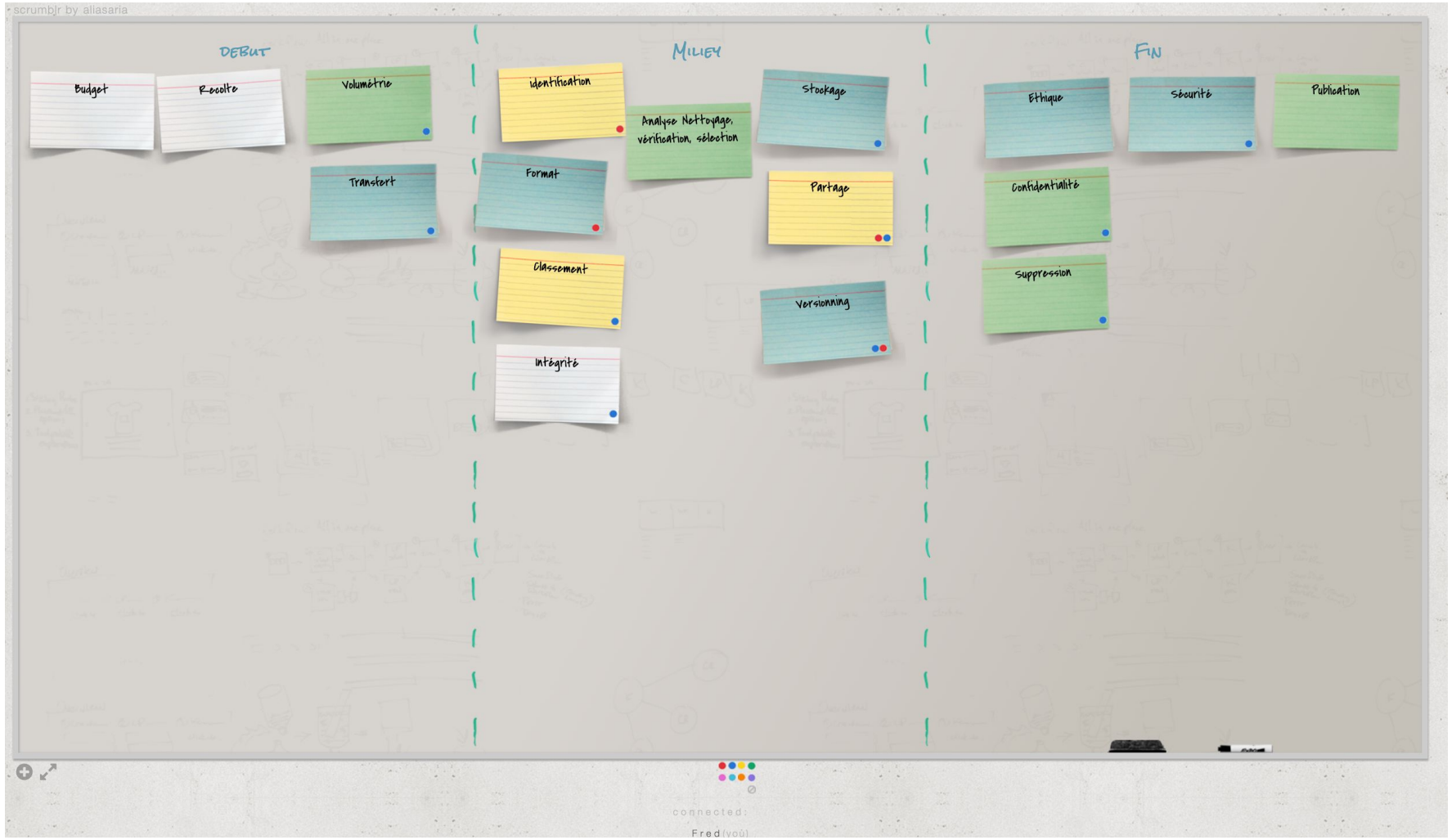
La vie des données

La vie des données le long du projet

Positionnez sur ce tableau les différents évènements de la vie des données le long d'un projet de recherche (5 minutes)



La vie des données le long du projet



Un cas d’usage

Un cas d’usage

Tout au long de cette session nous allons vous présenter des bonnes pratiques et des outils en nous mettant en situation au travers d'un cas d'usage.
Pour ce faire, imaginons que nous sommes une équipe de recherche et que nous souhaitons démarrer un nouveau projet de recherche.
Ce projet, nécessitera de mener de nombreuses expérimentations et acquisition de données diverses.
Nous espérons également qu'il nous permettra de proposer quelques bons papiers.
Nous nous efforcerons également tout au long du projet de garder en tête les principes FAIR que nous souhaiterons notamment mettre en oeuvre au travers de la publication de données

Attention, cette formation contient du placement de produits :-)

Un cas d’usage

Contexte : les chercheurs/ingénieurs effectuent leur travail sur leur PC, mais avec obligation de sauvegarder sur un serveur.

Situation : ce matin, je m’aperçois que mon PC est inaccessible : visiblement le disque dur est mort.

À part acquérir un nouveau poste de travail, comment vais-je récupérer mon environment logiciel & données ?

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 :

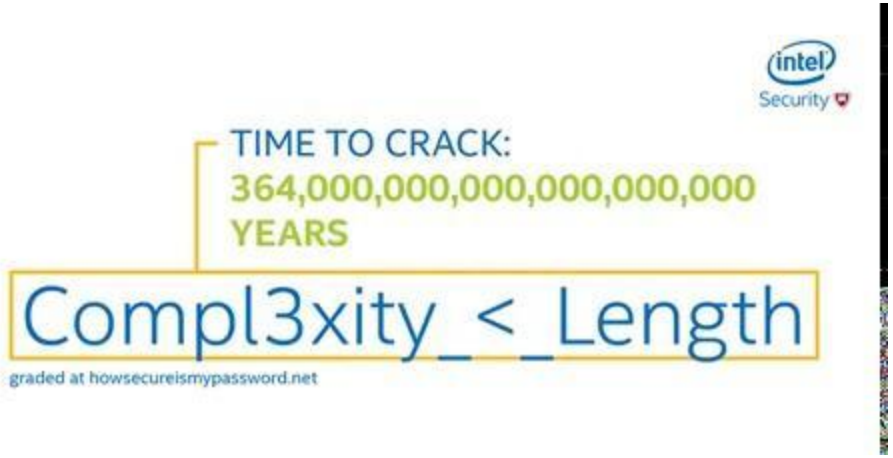
- Y'a-t-il des sauvegardes (stratégie 3-2-1) ?
  - **3** copies sur au moins **2** systèmes différents dont au moins **1** est distant = **0** inquiétude
    Par exemple : stockage en RAID (copie locale) + sauvegarde sur un disque externe qui reste au labo
- Votre environnement est-il mis à jour régulièrement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

Vos solutions de stockage :

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

Un environnement de travail sûr

Vos mots de passes (au pluriel)



Un environnement de travail sûr

Vos mots de passes (au pluriel)

- Utilisez-vous des mots de passe robustes ?

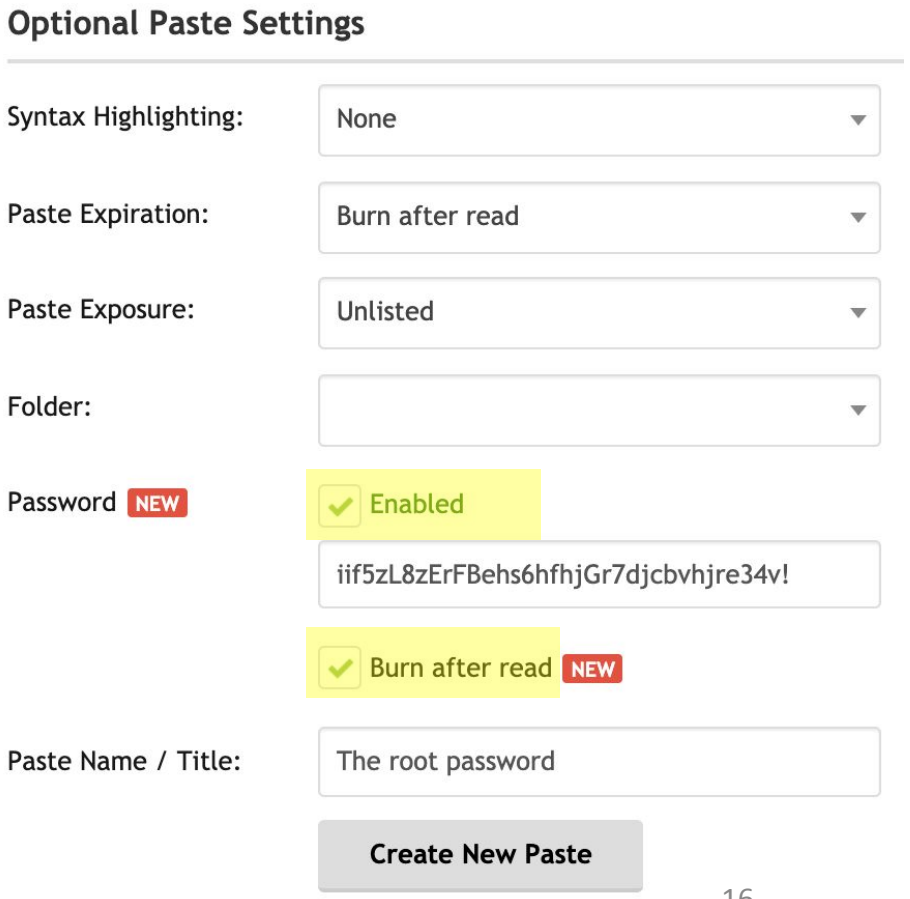
Type de mot de passe	Taille de clé équivalente	Force	Commentaire
Mot de passe de **8 caractères** dans un alphabet de **70 symboles**	49	Très faible	Taille usuelle
Mot de passe de **10 caractères** dans un alphabet de **90 symboles**	65	Faible	
Mot de passe de **12 caractères** dans un alphabet de **90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

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

Un environnement de travail sûr

Vos mots de passes (au pluriel)

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



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	**Gratuit**	**Accessible**
et donc pérenne	*mais n'hésitez pas à payer la souscription Premium pour soutenir le projet*	*Application Mac, Windows, Linux, Web, iPhone et Androïd*

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

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)

Notre espace de stockage

Un cas d’usage

Situation : après 7 mois d’attente, Sam Lee me transmet des données critiques par le biais d’une clé USB.

Q : quelle est ma démarche ?

Stockage des données

Fonction fondamentale : **la conservation des données**

Stockage :

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

Critères de sélection pour choisir un support de stockage :

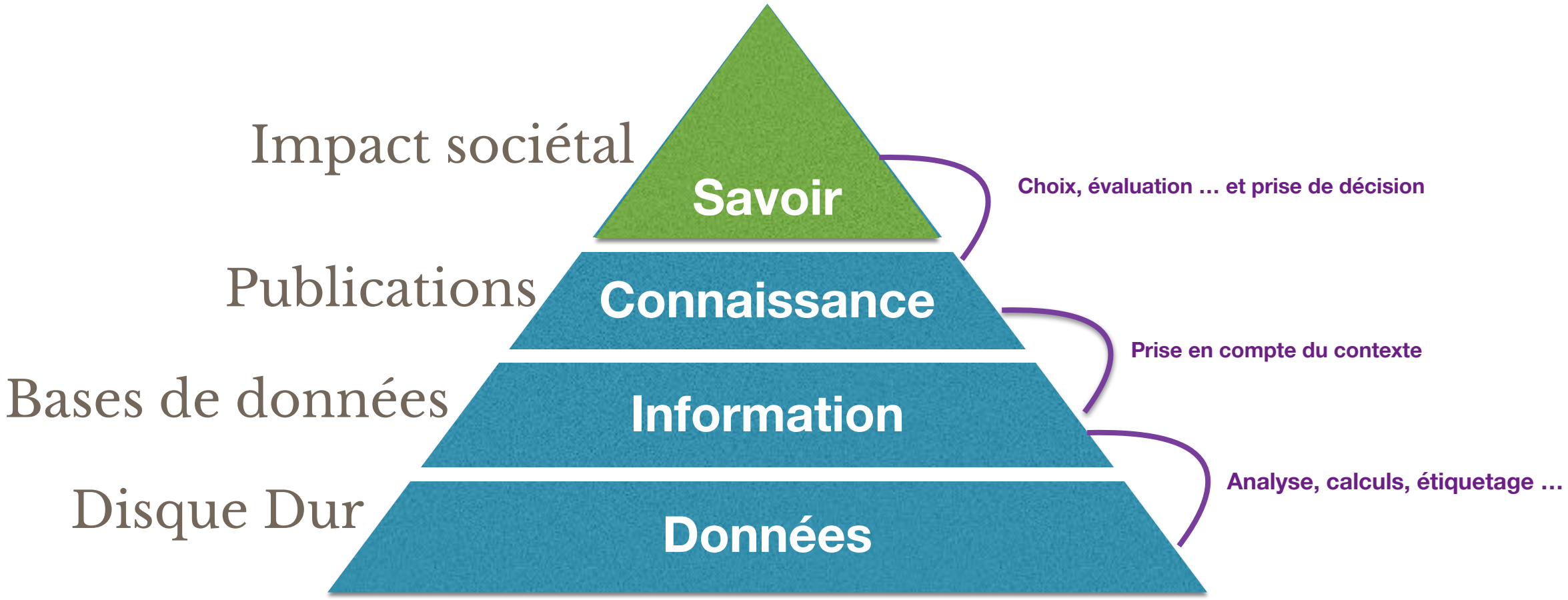
- la fréquence d'utilisation des données,
- les besoins en capacité de stockage (taille),
- la sécurité des données,
- la vitesse d'accès à la donnée
- la fiabilité et le coût du support

Stockage des données

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

Data Pyramid



Stocker et sécuriser : quels compromis ?

Comparatif de systèmes de stockage des données

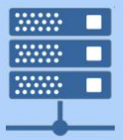
Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 Ordinateur professionnel	★★☆☆ Sujet au piratage informatique, aux détériorations et pannes	★☆☆☆ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud…)	★★★★ Pas de coût supplémentaire ou coût peu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

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

Comprenez l’infrastructure que vous utilisez

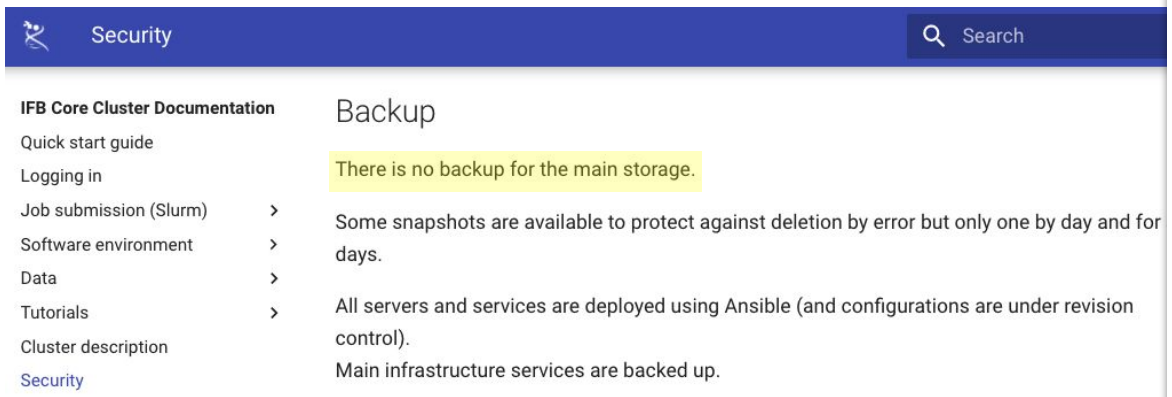
Performance vs Sécurité

- 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 voire 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

Comprenez l'infrastructure que vous utilisez

- 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 :

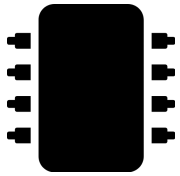
- consulter, corriger et améliorer les informations contenues sur le site pour toute question
- consulter les *notes de maintenance* sur le site web et sur les messages d'accueil des machines
- ne pas utiliser la machine comme espace de stockage ou de sauvegarde
- ne pas utiliser la machine comme passerelle 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

Le NNCR IFB

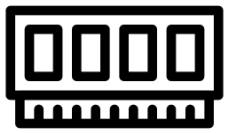
National Network of Computing Resources

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

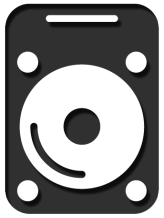
Le cluster national IFB



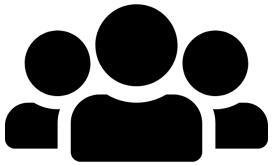
4300 coeurs



20 To RAM



2 Po



Une communauté d'entraide

Plus de 400 outils

SSH
Jupyter
RStudio
Galaxy

Le NNCR IFB

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	1 084	7 000	350
BiRD	Nantes	560	4 000	500

Le cluster national de l’IFB

Allocation des espaces de stockage par projet :

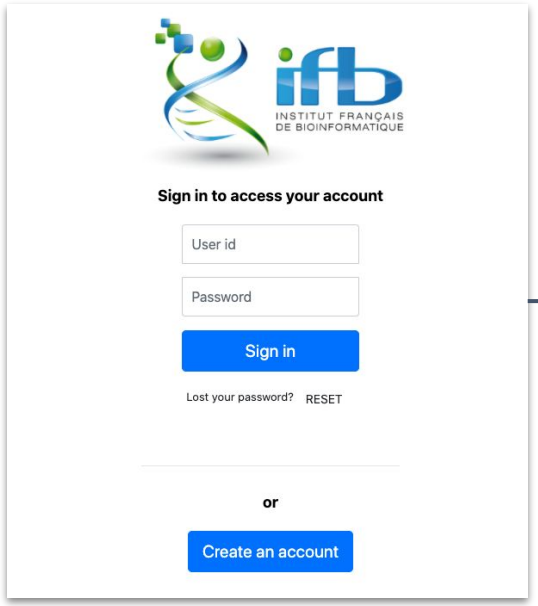
- 250 Go par projet extensible sur demande argumentée
- Un projet peut être accessible à plusieurs utilisateurs
- Un utilisateur peut demander plusieurs espaces projet
- Pas de sauvegarde

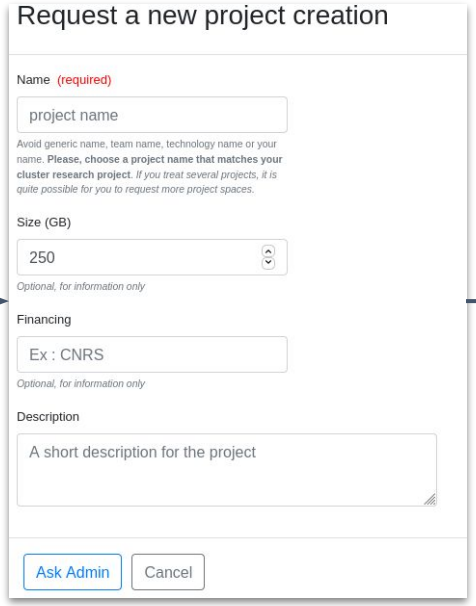
Bientôt disponible :

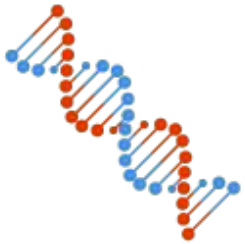
- Mise à disposition d’un espace scratch avec un quota plus important pour des besoins ponctuel (suppression automatique des fichiers les plus anciens)
- Sauvegarde des espaces projets

Demander un espace projet IFB cluster

https://my.cluster.france-bioinformatique.fr







Créer un compte

Vérifier son adresse email

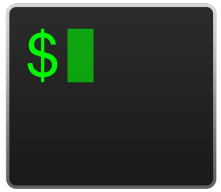
Attendre la validation de son compte

Demander un espace projet

Commencer à travailler

Accès à l'espace projet

FileZilla



Terminal

MobaXterm

SSH/SFTP
core.cluster.france-bioinformatique.fr

JupyterHub
.cluster.france-bioinformatique.fr

RStudio
.cluster.france-bioinformatique.fr

Transfert de vos données de recherche

Comment transmettre vos données ?

Pas bien → Bien

Messagerie instantanée

Email

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

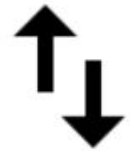
Envoi d'un disque

Dropbox, Drive, etc

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

Cloud privé

Service d'un consortium



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

Transfert de vos données de recherche

Démonstration :

- A l'aide de son terminal : la copie via SSH avec scp
- A l'aide d'un client sFTP : FileZilla
- A l'aide de son navigateur : JupyterHub

La vitesse de transfert dépend de :

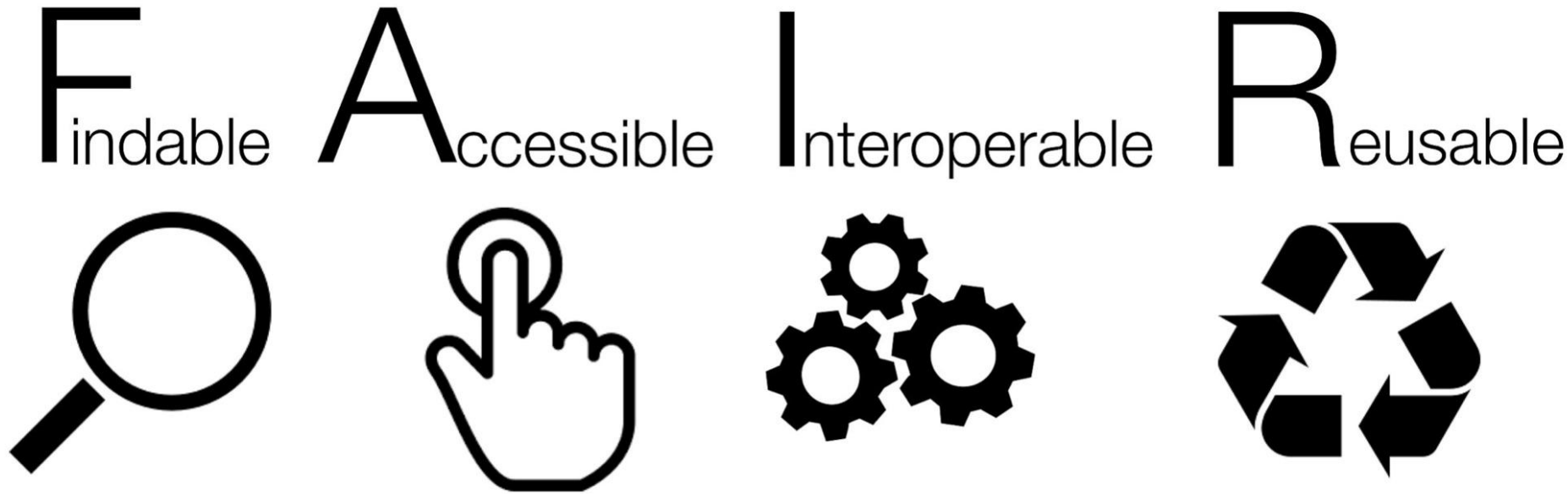
- L'outil utilisé
- L'infrastructure source et destination
- Le réseau
- La granularité des données

N'oubliez pas le chiffrement !!!!

Un cas d’usage

Situation : C’est parti pour un projet de biologie intégrative sur 3 ans, au programme acquisitions de nombreux types de données (imagerie, séquençage, phénotypage) et analyses intensives.

Q : Expliquez votre approche de nommage et d’organisation des fichiers (le nom des fichiers doit obligatoirement comprendre au moins la date)



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

Le nommage des fichiers

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*]

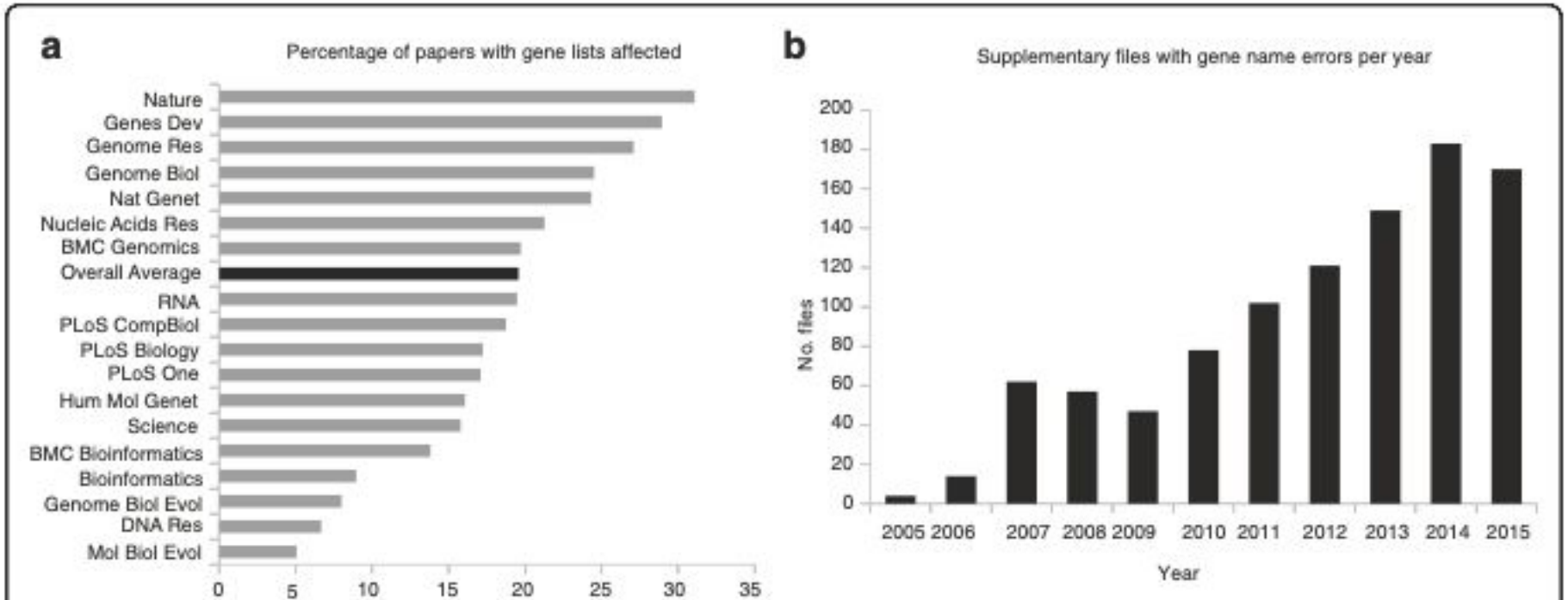


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

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

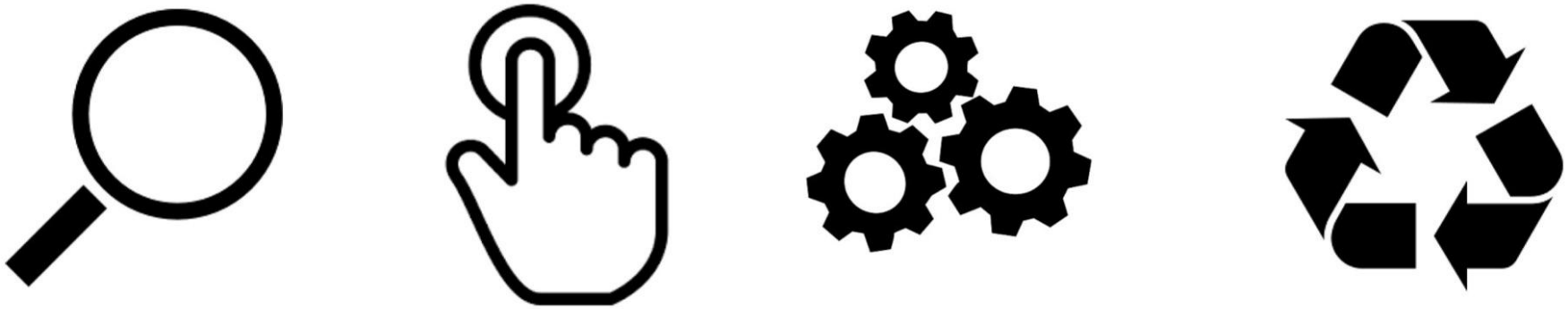
REPUBLIQUE ET CANTON DE GENEVE
Collège spécialisé des systèmes d'information

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)

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

Format de fichier

Un cas d’usage

Situation : vous devez traiter un fichier avec un format ‘propriétaire’, c’est à dire qui nécessite un logiciel non gratuit pour lire le fichier. Votre institution n’a aucune licence pour ce logiciel, et ne projette pas d’en acquérir.

Q : quelles sont les solutions possibles ?

Deux grandes catégories de formats : **textuels** et **binaires**.

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

Ex. de format textuel : HTML
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) :

Caractères ordinaires + caractères ayant une valeur spéciales : **< > /**, etc.
Mots ayant des valeurs spéciales en HTML (« balises ») si encadrés par < > ou </>: <body>, </body>, etc...

Ex. de format textuel : RTF (texte structuré) Contenu lisible dans un éditeur texte

```
{\rtf1\adeflang1025\ansi\ansicpg1252\uc1\adeff0\deff0\stshfdbch37\stshfl
och37\stshfhich37\stshfbi0\deflang1036\deflangfe1036\themelang1036\theme
langfe0\themelangcs0{\fonttbl{\f0\fbidi \froman\fcharset0\fprq2{\*\panos
e 02020603050405020304}Times New Roman;}{\f34\fbidi \froman\fcharset0\fp
rq2{\*\panose 02040503050406030204}Cambria Math;}

\mlMargin0\mrMargin0\mdefJc1\mwrapIndent1440\mintLim0\mnaryLim1}{\info{\
author Mathieu Saby}{\operator Mathieu Saby}{\creatim\yr2018\mo6\dy10\hr
13\min44}{\revtim\yr2018\mo6\dy10\hr13\min44}{\version2}{\edmins1}{\nofp
ages1}{\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\f37 Bonjour }{\rtlch\fcs1 \af0 \ltrch\fcs0 \cf6\insrsid16
651434\charrsid16651434

\hich\af37\dbch\af37\loch\f37 tout le monde}{\rtlch\fcs1 \af0 \ltrch\fcs
0 \insrsid16651434
```

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

Formats « binaires »

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

Ex. de format binaire : PNG (image)
Contenu illisible dans un éditeur texte (à part «?PNG » au début)

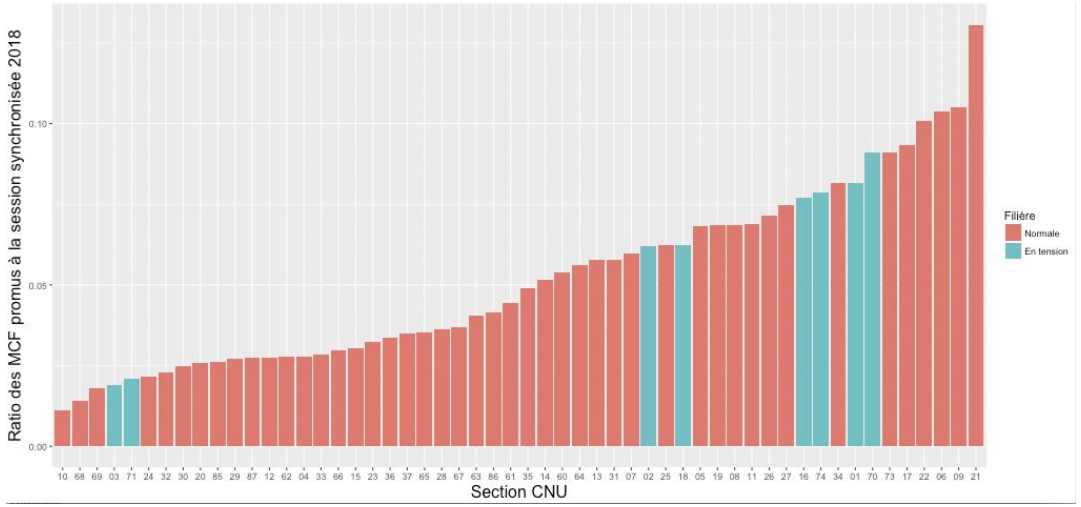
```
?PNG

?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
```

Uniquement lisible et interprétable avec une visionneuse d'images.

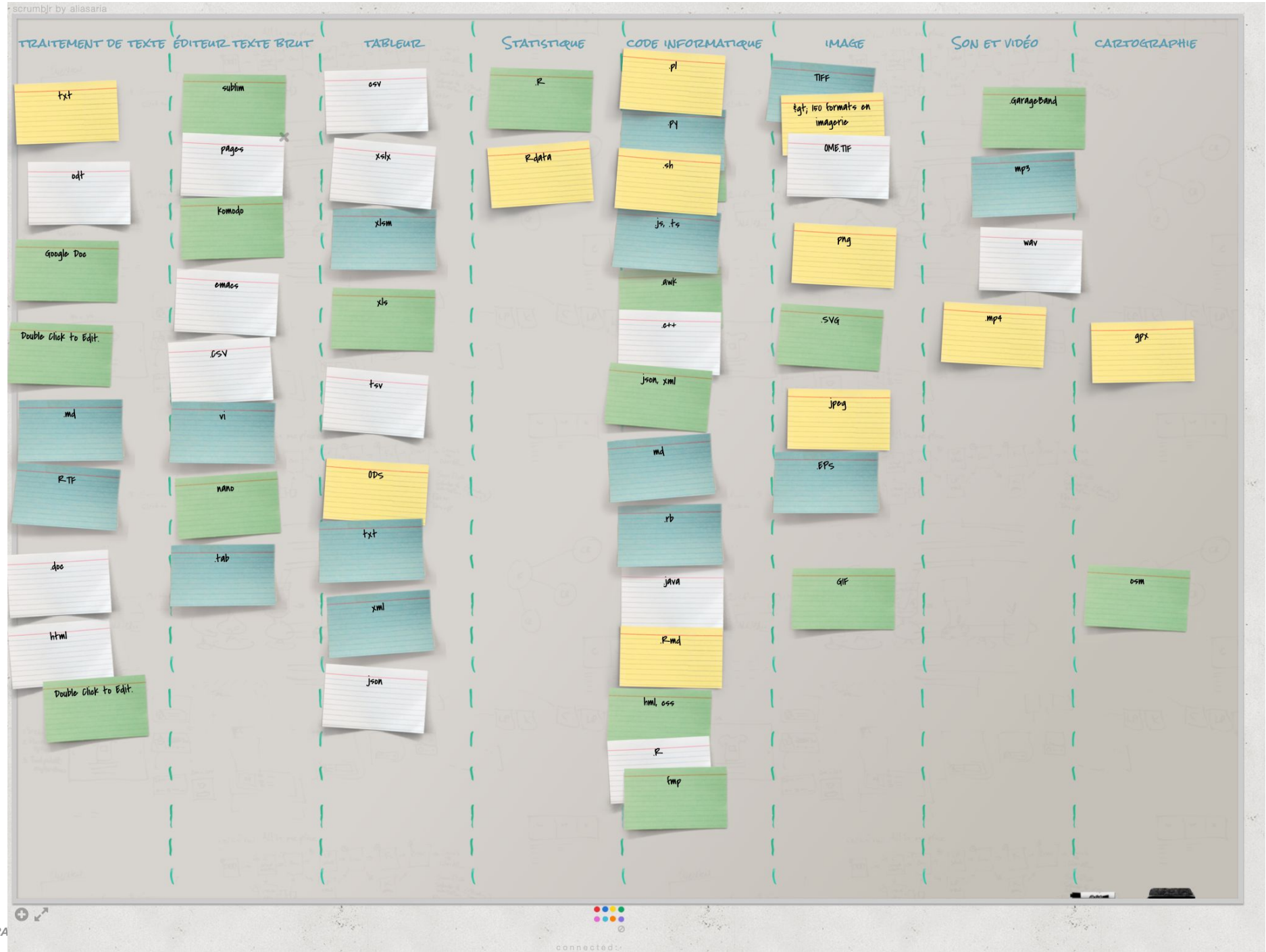


Quelles conséquences pour vous ?
Et pour ceux qui arriveront plus tard ?

Formats et logiciel ?

Allez à : http://scrumblr.ca/fair_data_m2_format

et listez les formats que vous connaissez dans les colonnes idoines


scrumblr by aliasaria
TRAITEMENT DE TEXTE
ÉDITEUR TEXTE BRUT
TABLEUR
STATISTIQUE
CODE INFORMATIQUE
IMAGE
SON ET VIDÉO
CARTOGRAPHIE
txt
odt
Google Doc
Double Click to Edit.
md
R-TF
.doc
html
Double Click to Edit.
sublim
pages
Komodo
emacs
.DSV
vi
nano
.tab
csv
x-s-x
xlsm
xls
tsv
ODS
txt
xml
json
.R
R-data
.pl
.PY
.sh
.js, .ts
.awk
.c++
json, xml
md
.rb
.java
.R-md
html, css
.R
fmp
TIFF
> 150 formats en imagerie
OME.TIF
png
.SVG
.jpeg
.EPS
GIF
GarageBand
mp3
wav
.mp4
gpx
osm

Les logiciels nécessaires pour traiter les formats cités sur scrumblr :

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

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

Sont-ils lies 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 ?

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

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	GeoTIFF (.tif, .tiff)	TIFF World File
Raster	ASCII GRID (.asc, .txt)	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.

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

Content type	High	Medium	Low
Text	• Plain text (encoding: USASCII, UTF-8, UTF-16 with BOM) • XML (includes XSD/XSL/XHTML, etc.; with included or accessible schema) • PDF/A-1 (ISO 19005-1) (*.pdf)	• Cascading Style Sheets (*.css) • DTD (*.dtd) • Plain text (ISO 8859-1 encoding • PDF (*.pdf) (embedded fonts) • Rich Text Format 1.x (*.rtf) • HTML (include a DOCTYPE declaration) • SGML (*.sgml) • Open Office (*.sxw/*.odt) • OOXML (ISO/IEC DIS 29500) (*.docx) • Microsoft Word 2007 or newer (*.docx)	• PDF (*.pdf) (encrypted) • Microsoft Word 2003 or older (*.doc) • WordPerfect (*.wpd) • DVI (*.dvi) • All other text formats not listed
Raster image	• TIFF (uncompressed) • JPEG2000 (lossless) (*.jp2) • PNG (*.png)	• BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy) (*.jp2) • TIFF (compressed) • GIF (*.gif) • Digital Negative DNG (*.dng)	• MrSID (*.sid) • TIFF (in Planar format) • FlashPix (*.fpx) • PhotoShop (*.psd) • RAW • JPEG 2000 Part 2 (*.jpf, *.jpx) • All other raster image formats not listed
Vector graphics	• SVG (no Java script binding) (*.svg)	• Computer Graphic Metafile (CGM, WebCGM) (*.cgm)	• Encapsulated Postscript (EPS) • Macromedia Flash (*.swf) • All other vector image formats not listed
Audio	• AIFF (96kHz 16bit PCM) (*.aif, *.aiff) • WAV (96kHz 24bit PCM) (*.wav)	• SUN Audio (uncompressed) (*.au) • Standard MIDI (*.mid, *.midi) • Ogg Vorbis (*.ogg) • Free Lossless Audio Codec (*.flac) • Advance Audio Coding (*.mp4, *.m4a, *.aac) • MP3 (MPEG-1/2, Layer 3) (*.mp3)	• AIFC (compressed) (*.aifc) • NeXT SND (*.snd) • RealNetworks 'Real Audio' (*.ra, *.rm, *.ram) • Windows Media Audio (*.wma) • Protected AAC (*.m4p) • WAV (compressed) (*.wav) • All other audio formats not listed
Video	• Motion JPEG 2000 (ISO/IEC 15444-4)??*.mj2) • AVI (uncompressed/native, motion JPEG) (*.avi) • QuickTime Movie (uncompressed/native, motion JPEG) (*.mov)	• Ogg Theora (*.ogg) • MPEG-1, MPEG-2 (*.mpg, *.mpeg, wrapped in AVI, MOV) • MPEG-4 (H.263, H.264) (*.mp4, wrapped in AVI, MOV)	• AVI (others) (*.avi) • QuickTime Movie (others) (*.mov) • RealNetworks 'Real Video' (*.rv) • Windows Media Video (*.wmv) • All other video formats not listed

Formats standardisés

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

Ex :

PDF/A1 est une version standardisée (ISO 19005) du format PDF. Les autres versions de PDF ne sont pas standardisées
Les formats Libre office (ODS, ODT…) sont standardisés (ISO/IEC 26300)
Le format XML est standardisé par une « recommandation » du W3C (équivaut à une norme)
Le format CSV est décrit dans la RFC 4180 de l'IETF, mais n'est pas réellement standardisé (la RFC est un document indicatif), plusieurs versions existent

Les formats bureautique Microsoft (XLSX, DOCX…) sont standardisés (ISO/IEC 29500). Mais les logiciels semblent parfois s'écarter du standard

http://dotwhat.net

dot What?
The leading File Extension resource

HOME ADD EXTENSION EDIT EXTENSION ARTICLES FAQ CONTACT US

J'aime 285 Tweet

Search...

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z OTHER

WELCOME TO DOTWHAT? ... THE LEADING FILE EXTENSION RESOURCE

Thanks to years of research and help from our loyal visitors, we now have one of the world's largest and most detailed databases of file extension information, covering multiple operating systems from Microsoft's Windows, Apple's OS X and all variations of Unix to those used on the latest mobile devices and phones.

EVERYTHING YOU NEED TO KNOW! IF NOT, JUST ASK!

We try to provide as much information on each file extension as possible and we encoura visitors to contact us if they have any additional information on an extension or if they think a new file extension should be added to the database. Alternatively, each entry can be edited and visitors have the option of adding a comment, question or tip!

Sections

Software Developers

Software Products

Common File Extensions

Categories

3D/CAD Files

Audio Files

Backup Files

Compressed Files

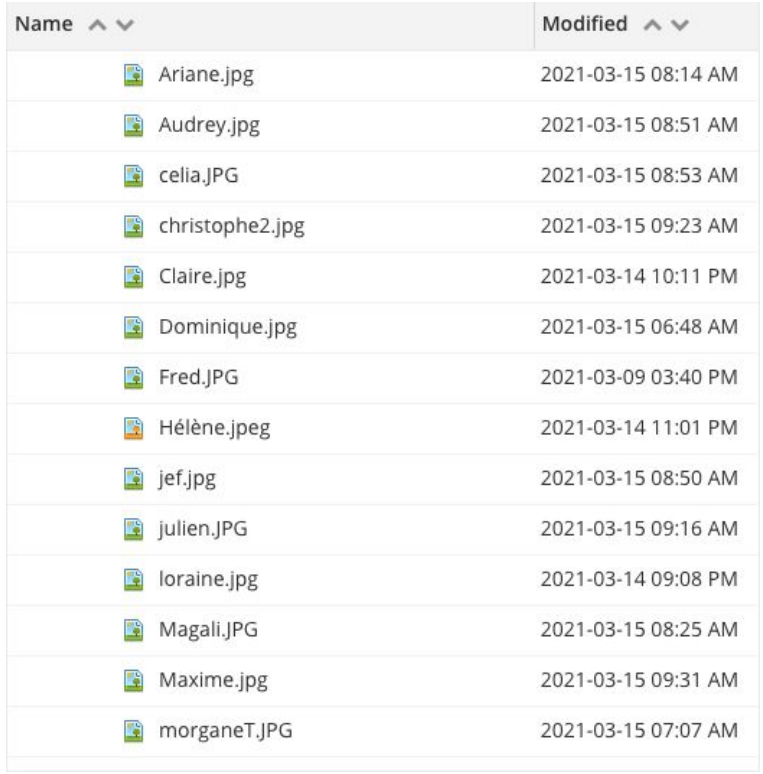
Configuration Files

Data Files

Organisation des données

Organisation des données

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



Name	Modified
Ariane.jpg	2021-03-15 08:14 AM
Audrey.jpg	2021-03-15 08:51 AM
celia.JPG	2021-03-15 08:53 AM
christophe2.jpg	2021-03-15 09:23 AM
Claire.jpg	2021-03-14 10:11 PM
Dominique.jpg	2021-03-15 06:48 AM
Fred.JPG	2021-03-09 03:40 PM
Hélène.jpeg	2021-03-14 11:01 PM
jef.jpg	2021-03-15 08:50 AM
julien.JPG	2021-03-15 09:16 AM
loraine.jpg	2021-03-14 09:08 PM
Magali.JPG	2021-03-15 08:25 AM
Maxime.jpg	2021-03-15 09:31 AM
morganeT.JPG	2021-03-15 07:07 AM

Organisation des données

Organisation des dossiers

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

Pour chaque dossier, ajoutez un fichier README:

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

Organisation des données

Exemple :

Un dossier par projet

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

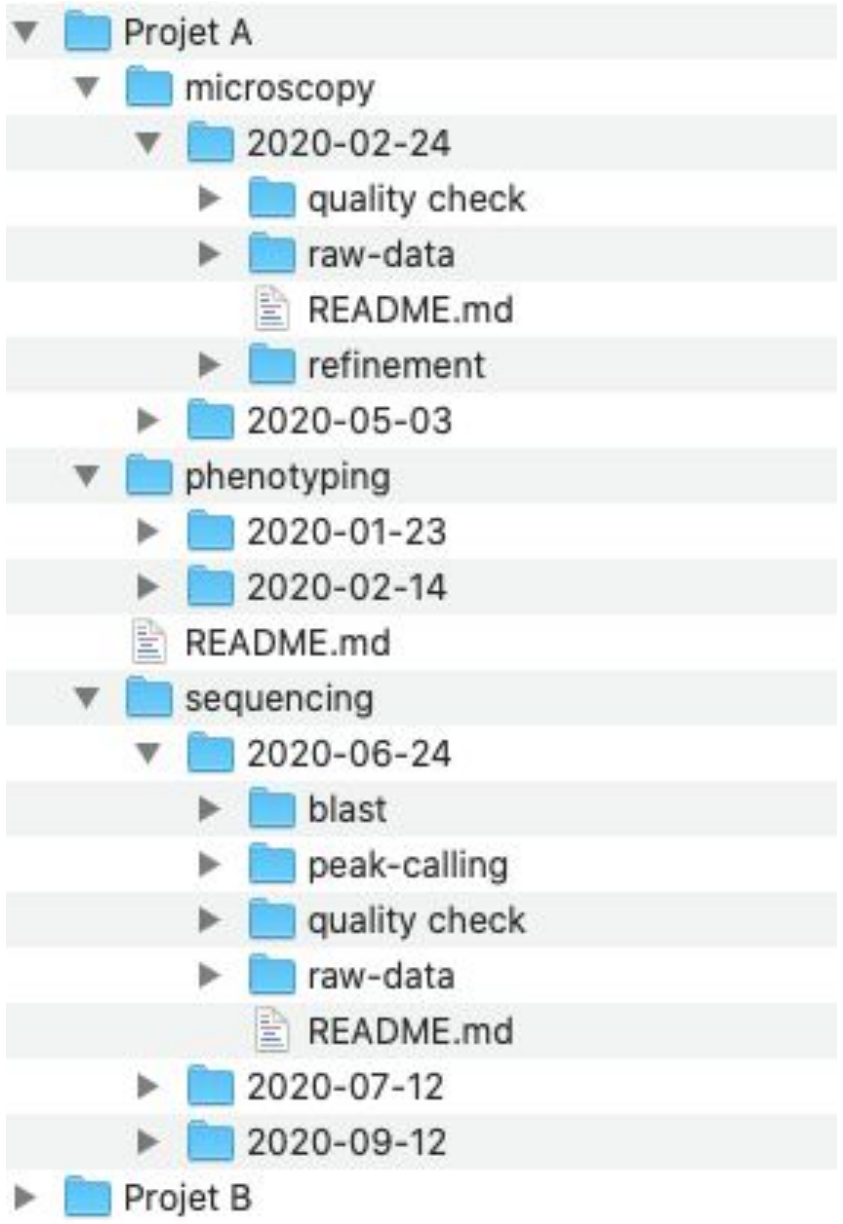
Un sous-dossier par date (2020-02-24, 2020-05-03)

Un sous-dossier pour les données brutes

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

Un sous-dossier par publication

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



Un cas d’usage

Situation : une partie de vos données est considérée comme données sensibles.

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, collègue un peu c**…)
- un transfert de données défectueux
- un plantage d’un disque dur
- ...

Intégrité des données

Identifier et contrôler la corruption des données

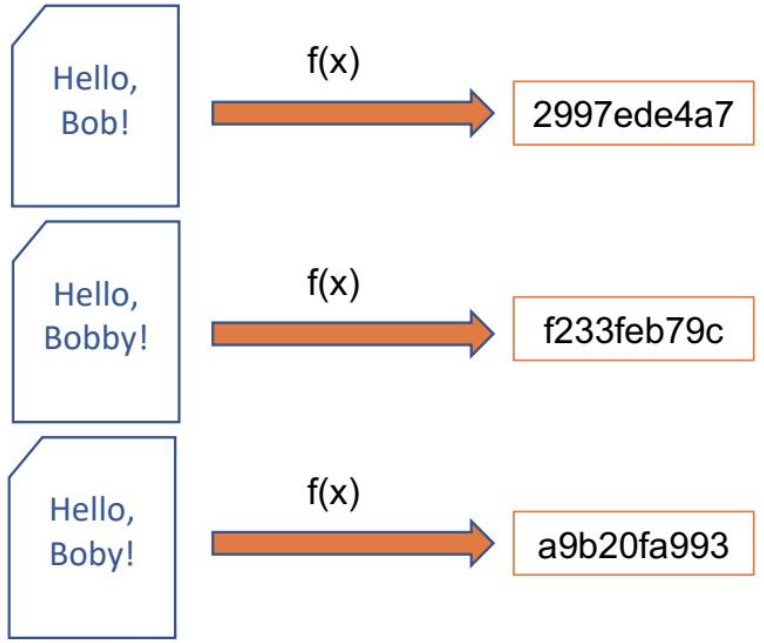
Solution 1 : générer des sommes de contrôles

Comment ?

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

Quand ?

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



Intégrité des données

Identifier et contrôler la corruption des données

Solution 2 : utilisez le contrôle d'accès

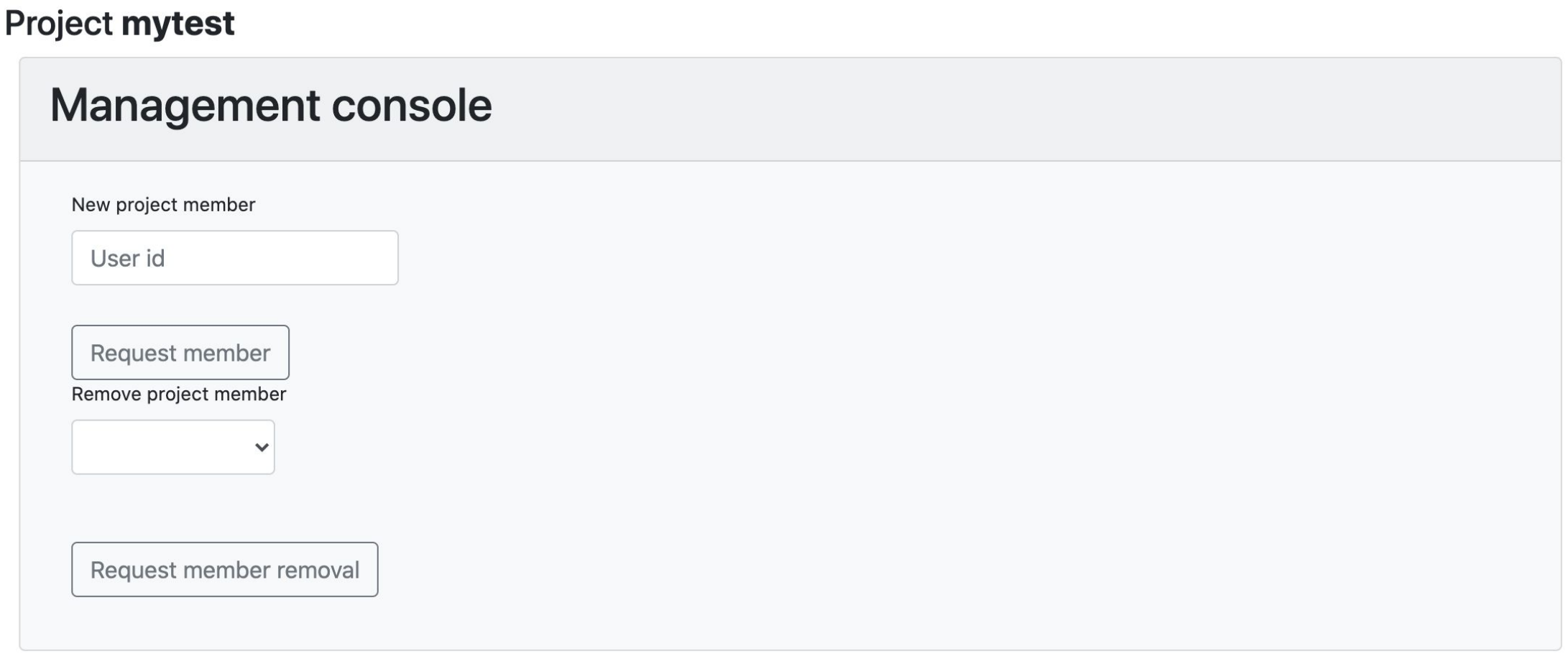
N'accordez que les permissions d'accès nécessaire :

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

Mettez les données brutes en lecture seule

L'accès aux données sensibles doit être documenté

Gérer l’accès à son projet



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

Un cas d’usage

Situation : vous êtes régulièrement obligés (par votre institution etc…) de procéder au nettoyage des données que vous sauvegardez sur le serveur. Cette obligation s’accompagne de la nécessité de justifier la raison pourquoi les dossiers listés doivent être conservés.
A cet effet, les dossiers sont tagués/catégorisés : par exemple “pour publication”, etc…

A REVOIR

Quelles catégories seraient pertinentes pour justifier la conservation des données ?

La suppression des données

Suppression des données

Est-ce que ces données peuvent être supprimés ?

Le stockage des données a un coût financier et écologique.

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

Un petit exercice :

Quels jeux de données puis-je supprimer ?
(on va pas être d'accord)

https://www.wooclap.com/JNHKXR

Conservation des données

“Je veux garder mes données pour l’éternité”

Ne manquez pas le module 4...

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

Les infrastructures de stockage sont vos amies

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

Essayons de nous améliorer

Où se situe mon fichier ?

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

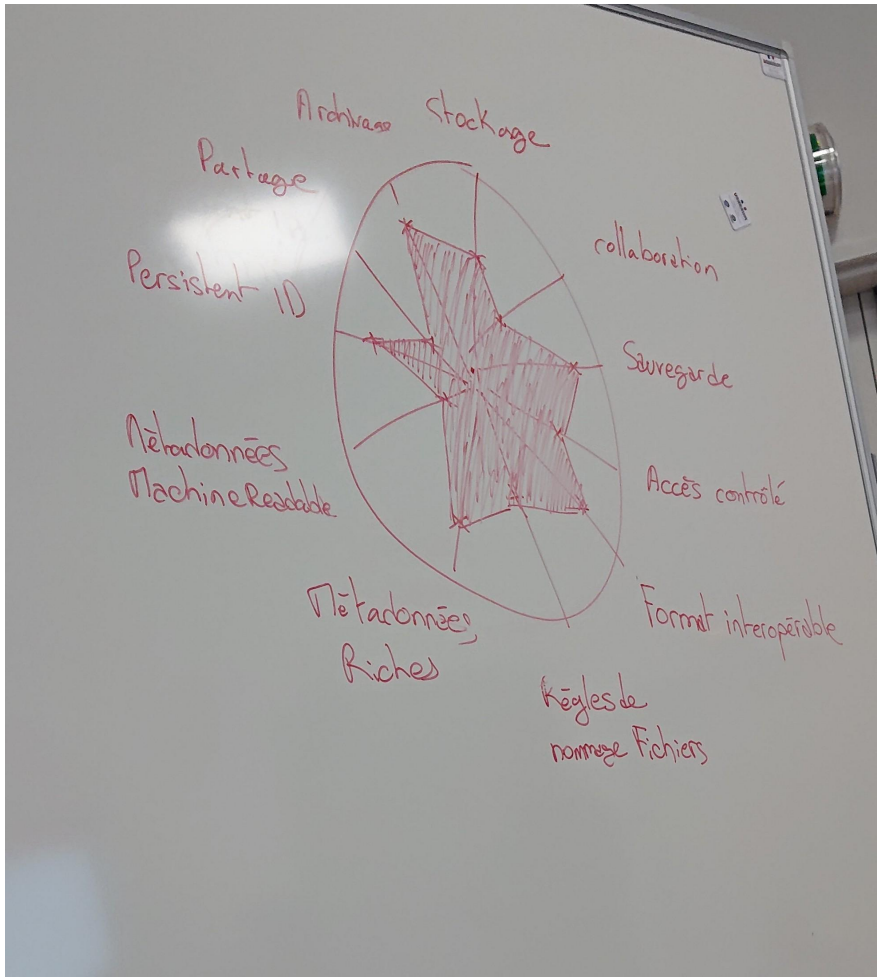
Propriétaire | Norme

Formalisation

Pas de norme | Norme propriétaire | Norme Iso

Exercice

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



Gestion Electronique de Documents

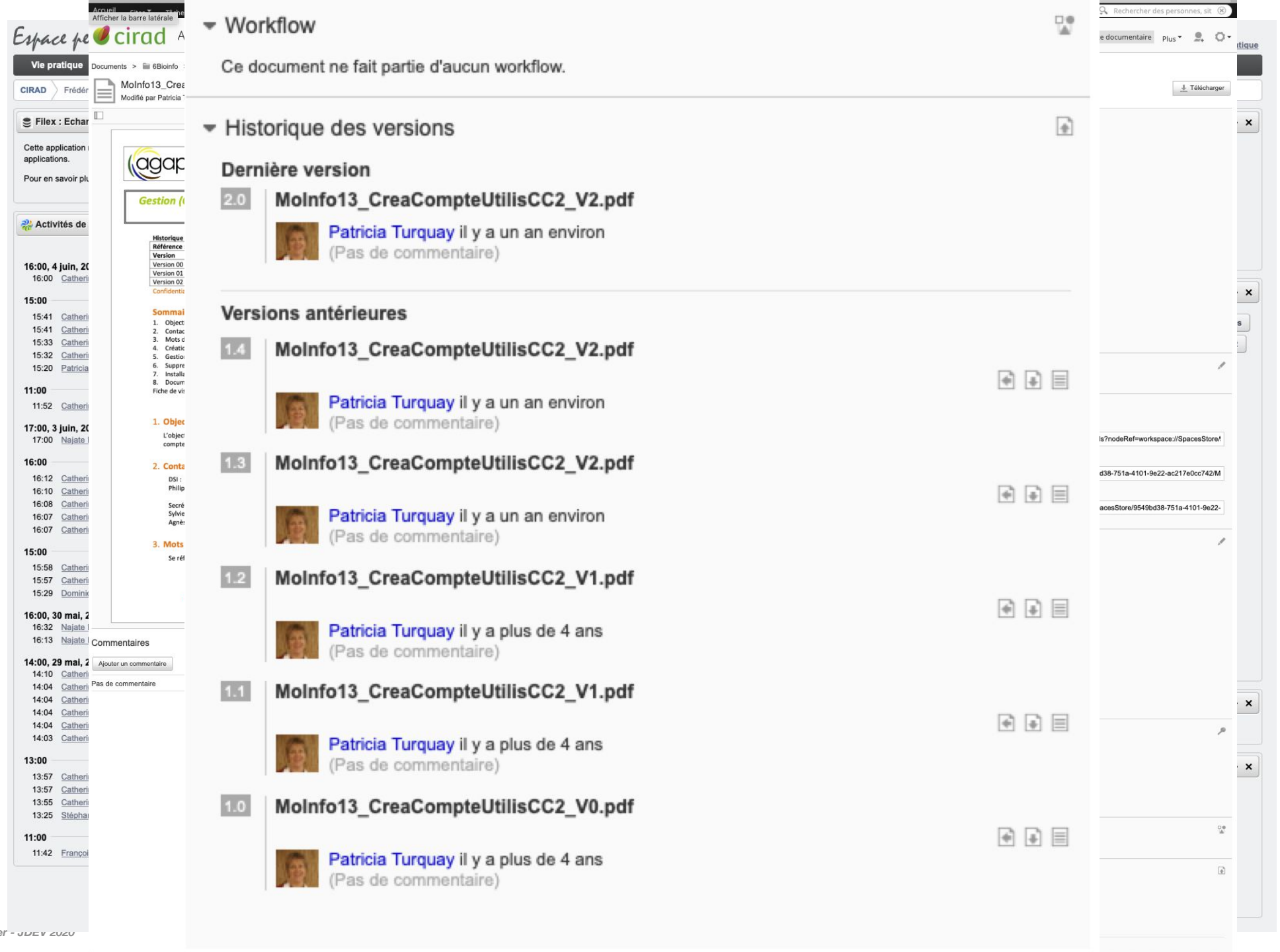
Espace de :
Préservation
et de
Partage
du
Savoir
du
Groupe

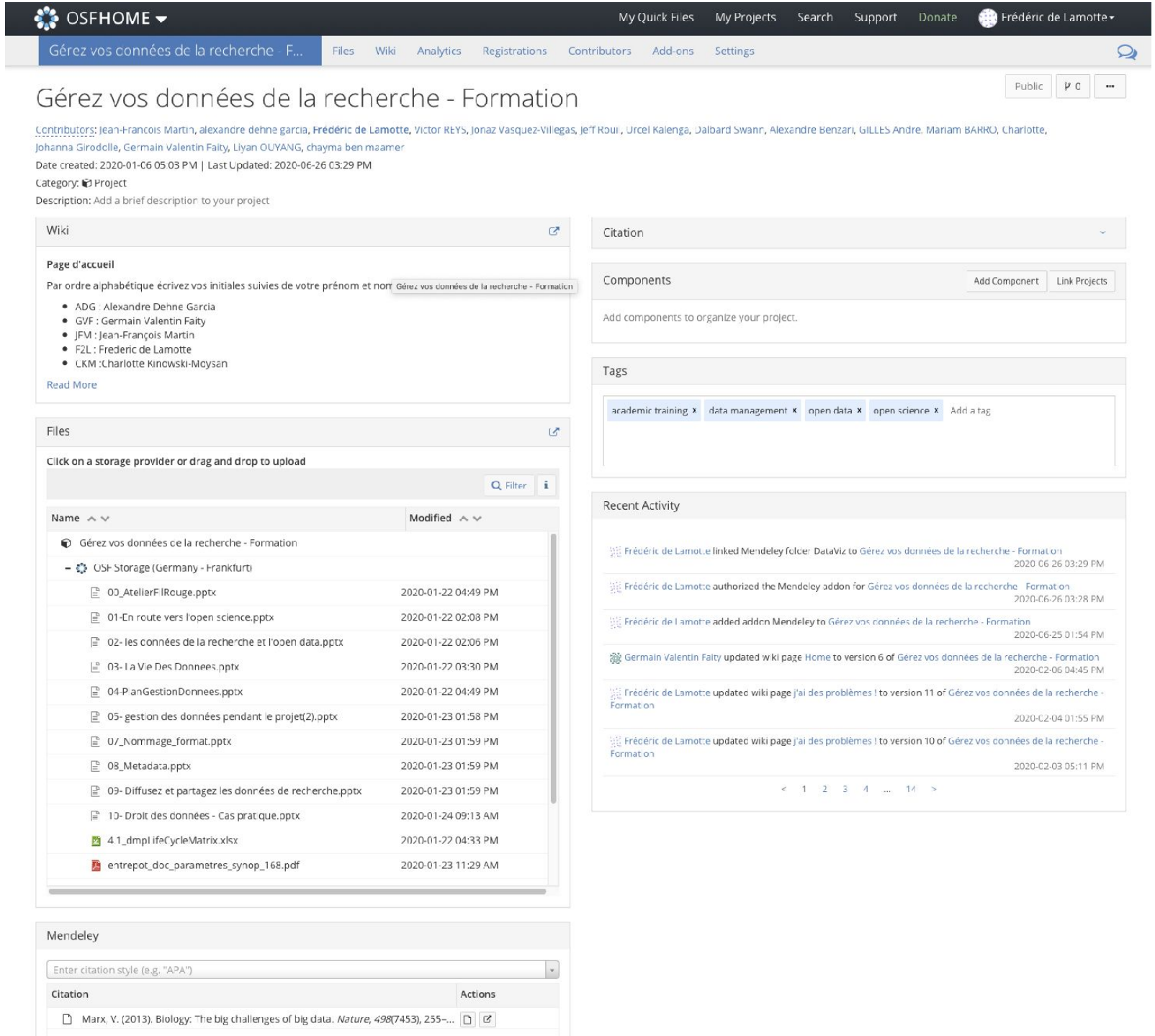
Tout le monde à sa GED

Gratuite ou payante

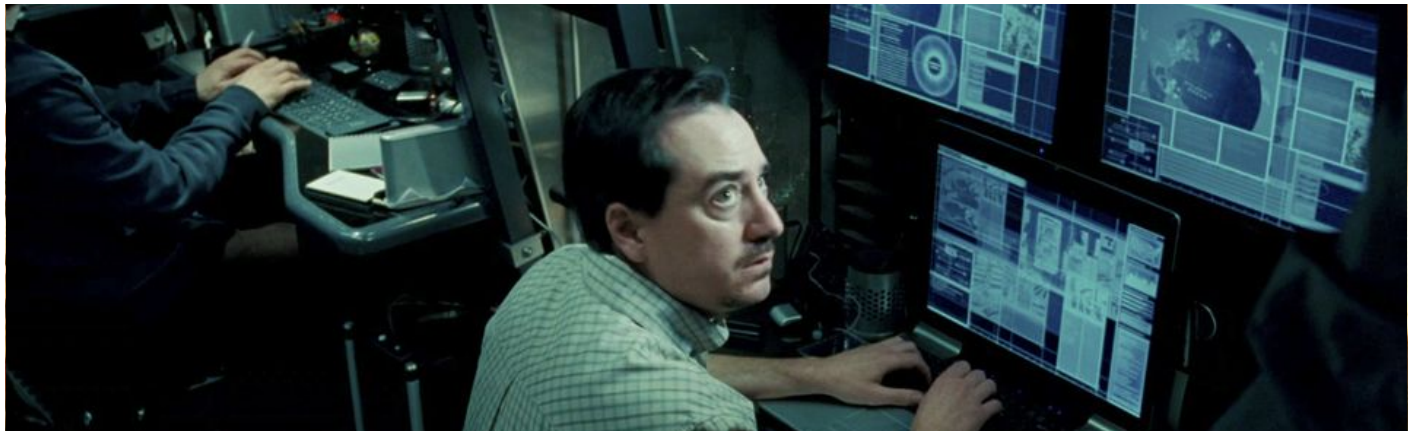
Bonne ou mauvaise

Choisie ou imposée


Workflow
Ce document ne fait partie d'aucun workflow.
Historique des versions
Dernière version
2.0
MoInfo13_CreaCompteUtilisCC2_V2.pdf
Patricia Turquay il y a un an environ
(Pas de commentaire)
Versions antérieures
1.4
MoInfo13_CreaCompteUtilisCC2_V2.pdf
Patricia Turquay il y a un an environ
(Pas de commentaire)
1.3
MoInfo13_CreaCompteUtilisCC2_V2.pdf
Patricia Turquay il y a un an environ
(Pas de commentaire)
1.2
MoInfo13_CreaCompteUtilisCC2_V1.pdf
Patricia Turquay il y a plus de 4 ans
(Pas de commentaire)
1.1
MoInfo13_CreaCompteUtilisCC2_V1.pdf
Patricia Turquay il y a plus de 4 ans
(Pas de commentaire)
1.0
MoInfo13_CreaCompteUtilisCC2_V0.pdf
Patricia Turquay il y a plus de 4 ans
(Pas de commentaire)

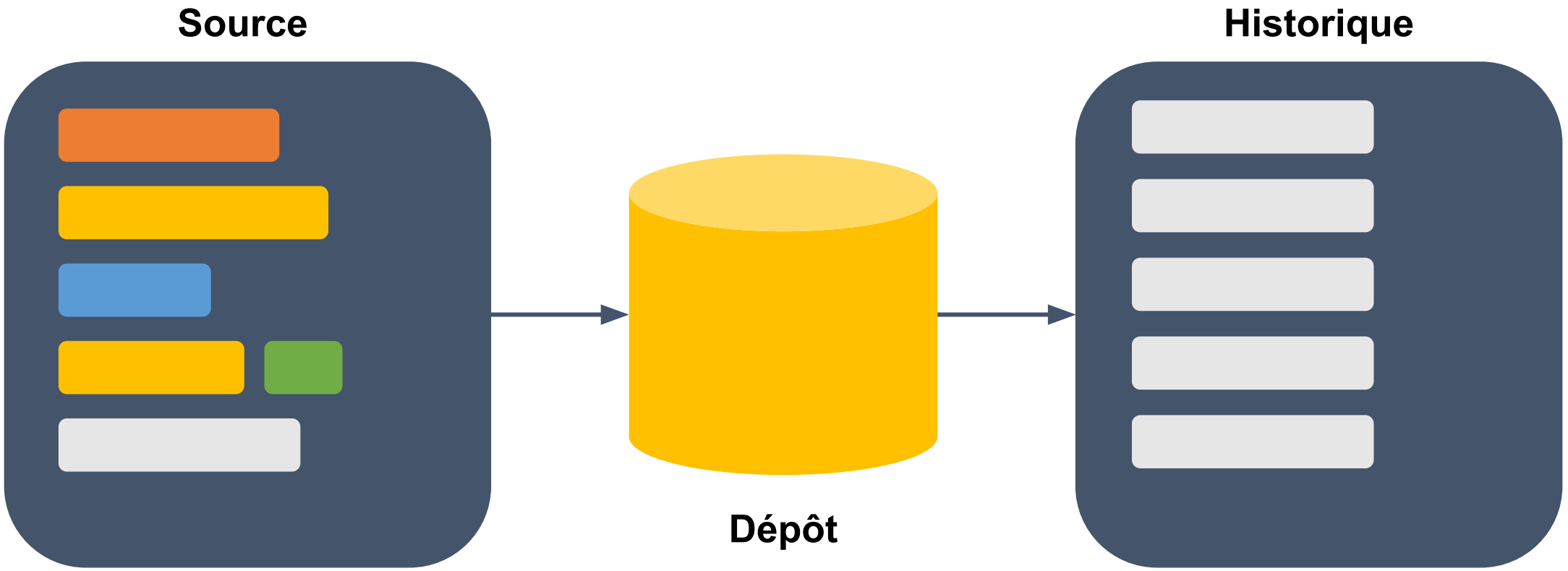

OSFHOME
My Quick Files
My Projects
Search
Support
Donate
Frédéric de Lamotte
Gérez vos données de la recherche F...
Files
Wiki
Analytics
Registrations
Contributors
Add-ons
Settings
Gérez vos données de la recherche - Formation
Public
Contributors: Jean-Francois Martin, alexandre dehne garcia, Frédéric de Lamotte, Victor REYS, Jonaz Vasquez-Villegas, Jeff Rouf, Urcel Kalenga, Dalbard Swann, Alexandre Benzari, GILLES Andre, Mariam BARRO, Charlotte, Johanna Girodolle, Germain Valentin Faity, 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
Gérez vos données de la recherche - Formation
ADG : Alexandre Dehne Garcia
GVF : Germain Valentin Faity
JFM : Jean-François Martin
FDL : Frederic de Lamotte
LKM :Charlotte Kinowski-Meysan
Read More
Citation
Components
Add Component
Link Projects
Add components to organize your project.
Tags
academic training
data management
open data
open science
Add a tag
Files
Click on a storage provider or drag and drop to upload
Filter
Name
Modified
Gérez vos données de la recherche - Formation
OSF Storage (Germany - Frankfurt)
00_AtelierFIRouge.pptx 2020-01-22 04:49 PM
01-En 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-PlanGestionDonnees.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_Metadata.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
Recent Activity
Frédéric de Lamotte linked Mendeley folder DataViz to Gérez vos données de la recherche - Formation
2020-06-26 03:29 PM
Frédéric de Lamotte authorized the Mendeley addon for Gérez vos données de la recherche - Formation
2020-06-26 03:28 PM
Frédéric de Lamotte added addon Mendeley to Gérez vos données de la recherche - Formation
2020-06-25 01:54 PM
Germain Valentin Faity updated wiki page Home to version 6 of Gérez vos données de la recherche - Formation
2020-02-06 04:45 PM
Frédéric de Lamotte updated wiki page j'ai des problèmes ! to version 11 of Gérez vos données de la recherche - Formation
2020-02-04 01:55 PM
Frédéric de Lamotte updated wiki page j'ai des problèmes ! to version 10 of Gérez vos données de la recherche - Formation
2020-02-03 05:11 PM
< 1 2 3 4 ... 14 >
Mendeley
Enter citation style (e.g. "APA")
Citation
Actions
Marx, V. (2013). Biology: The big challenges of big data. Nature, 498(7453), 255–...

Gestion des codes sources



C’est quoi Git ?

Git est un système de gestion de versions (version control system)



C’est quoi Git ?

v1.0.0

v1.0.1

v1.0.2

C’est quoi Git ?

Alice

v1.0.0

v1.0.1

v1.0.2

Bob

v1.0.0

v1.0.1

v1.0.2

C’est quoi Git ?

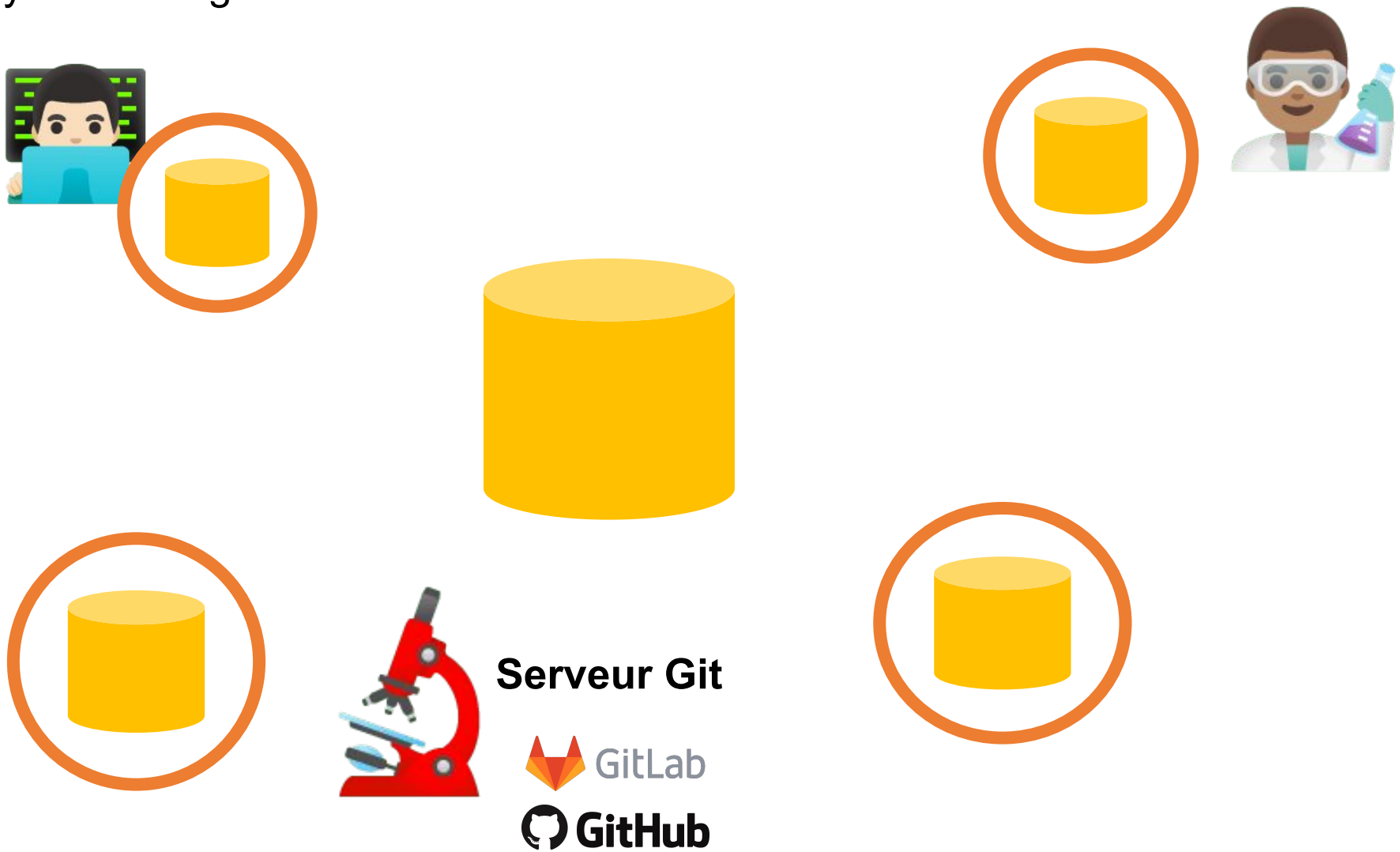
Système de gestion des versions

Suivre les changements

Travailler ensemble

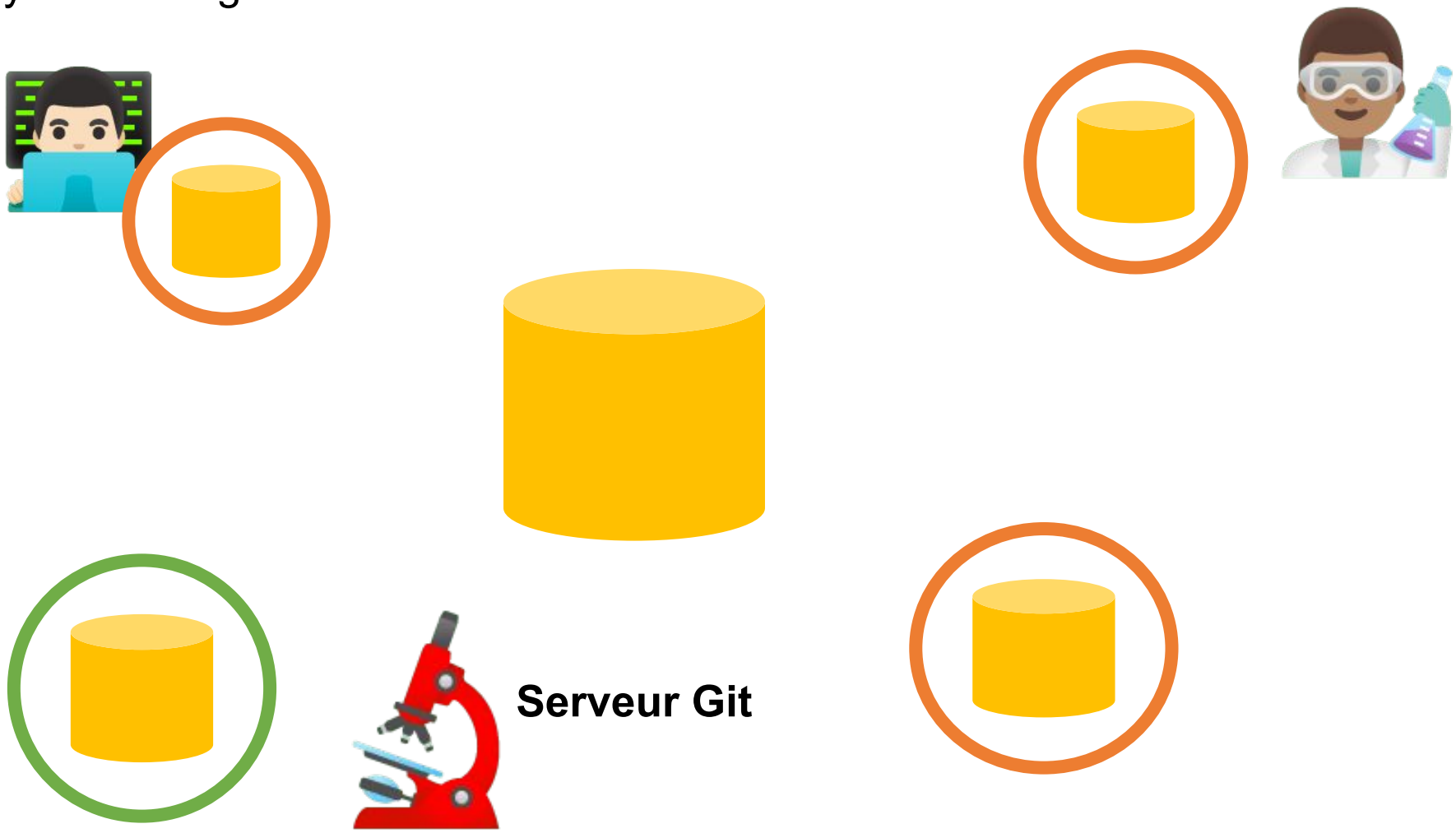
C’est quoi Git ?

Git est un système de gestion de versions DISTRIBUÉ



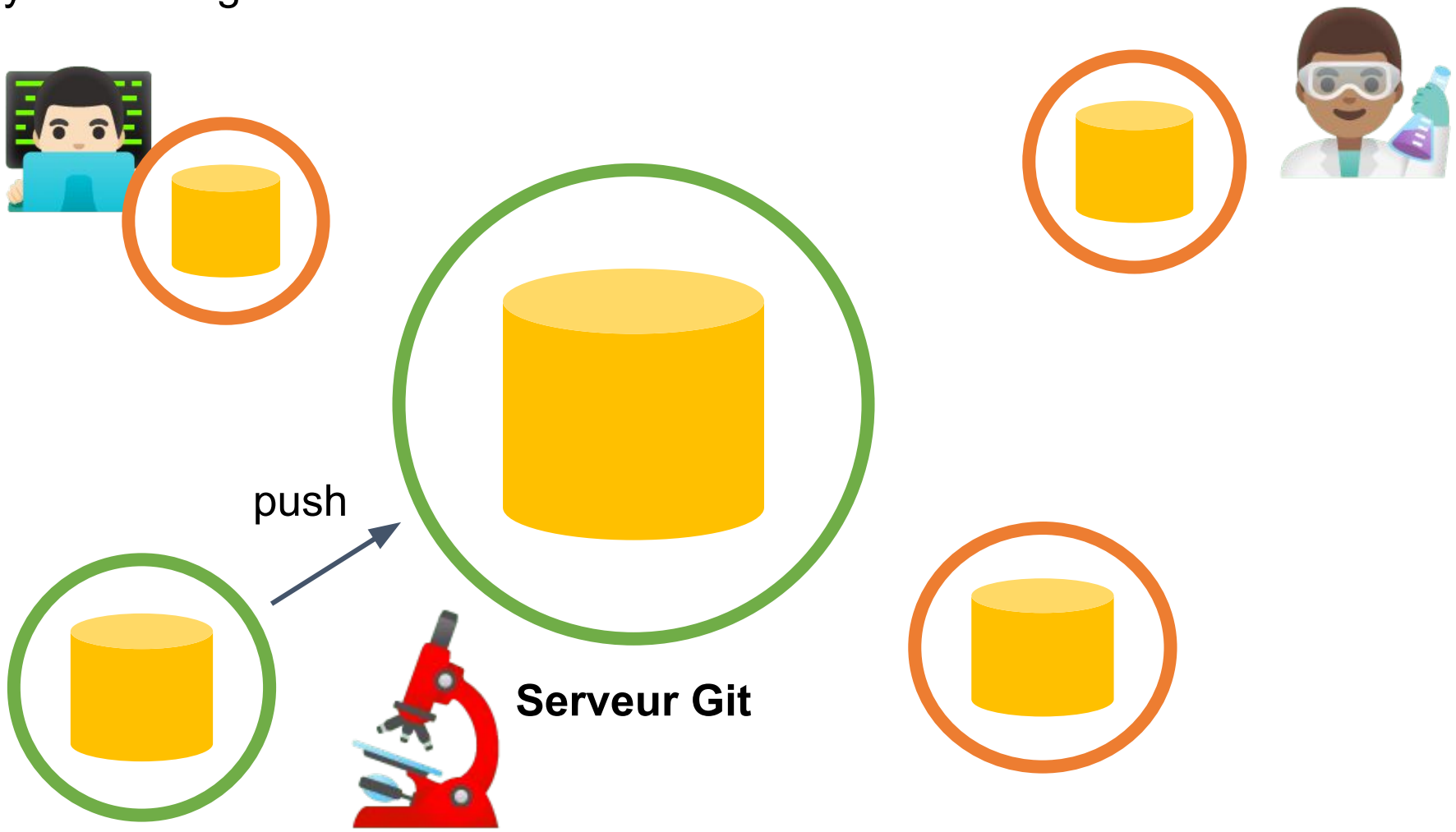
C’est quoi Git ?

Git est un système de gestion de versions DISTRIBUÉ



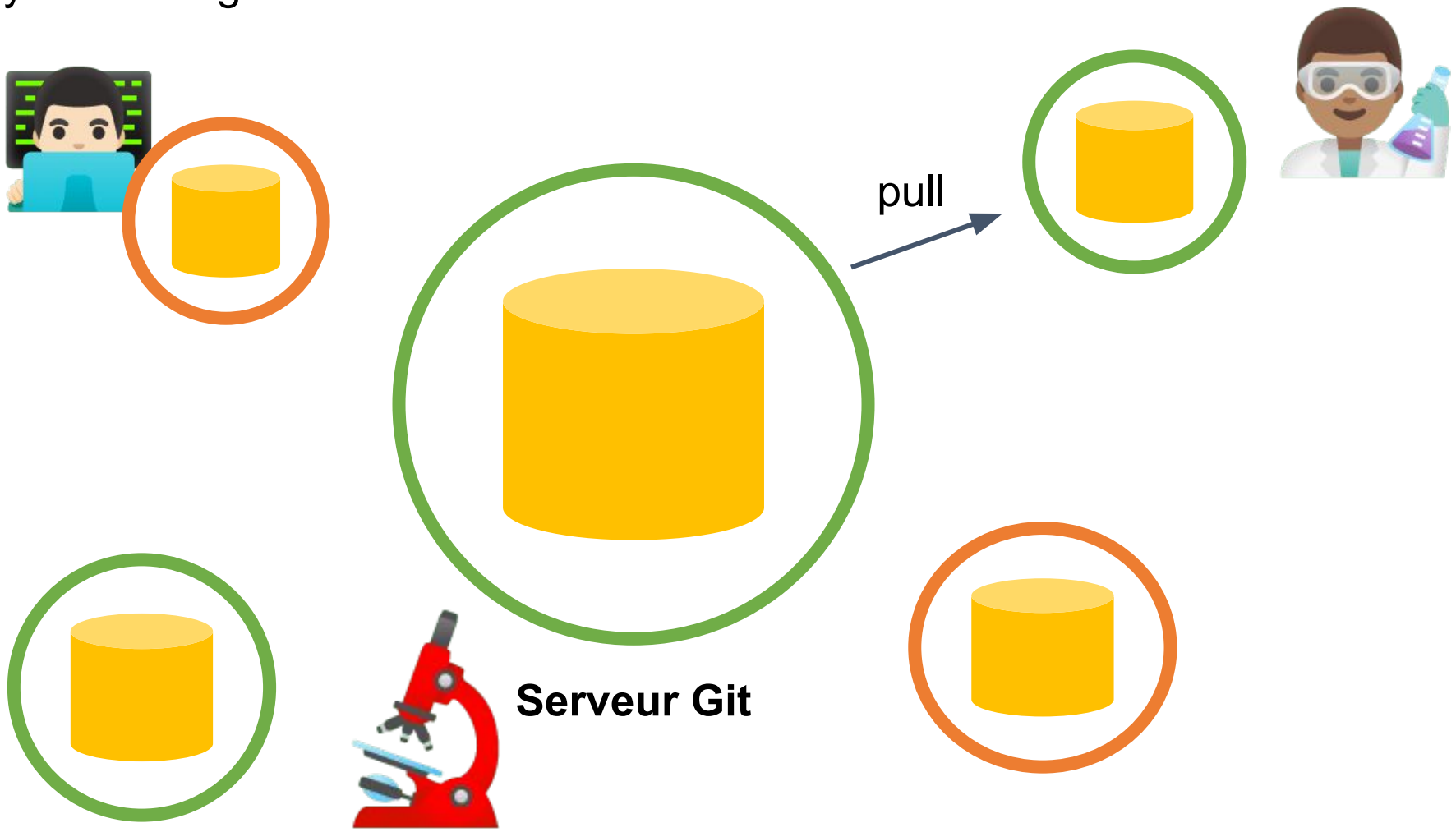
C’est quoi Git ?

Git est un système de gestion de versions DISTRIBUÉ



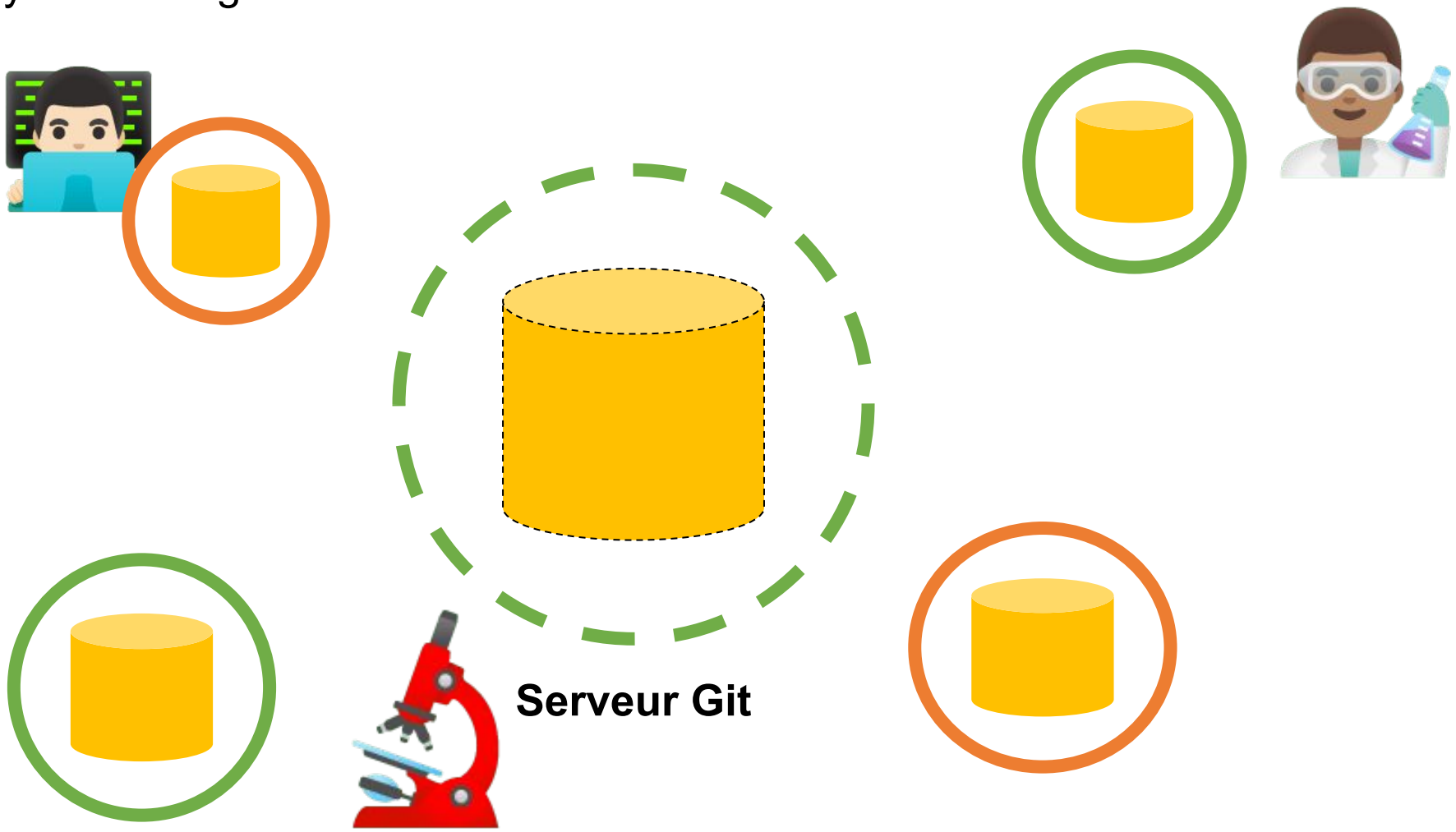
C’est quoi Git ?

Git est un système de gestion de versions DISTRIBUÉ



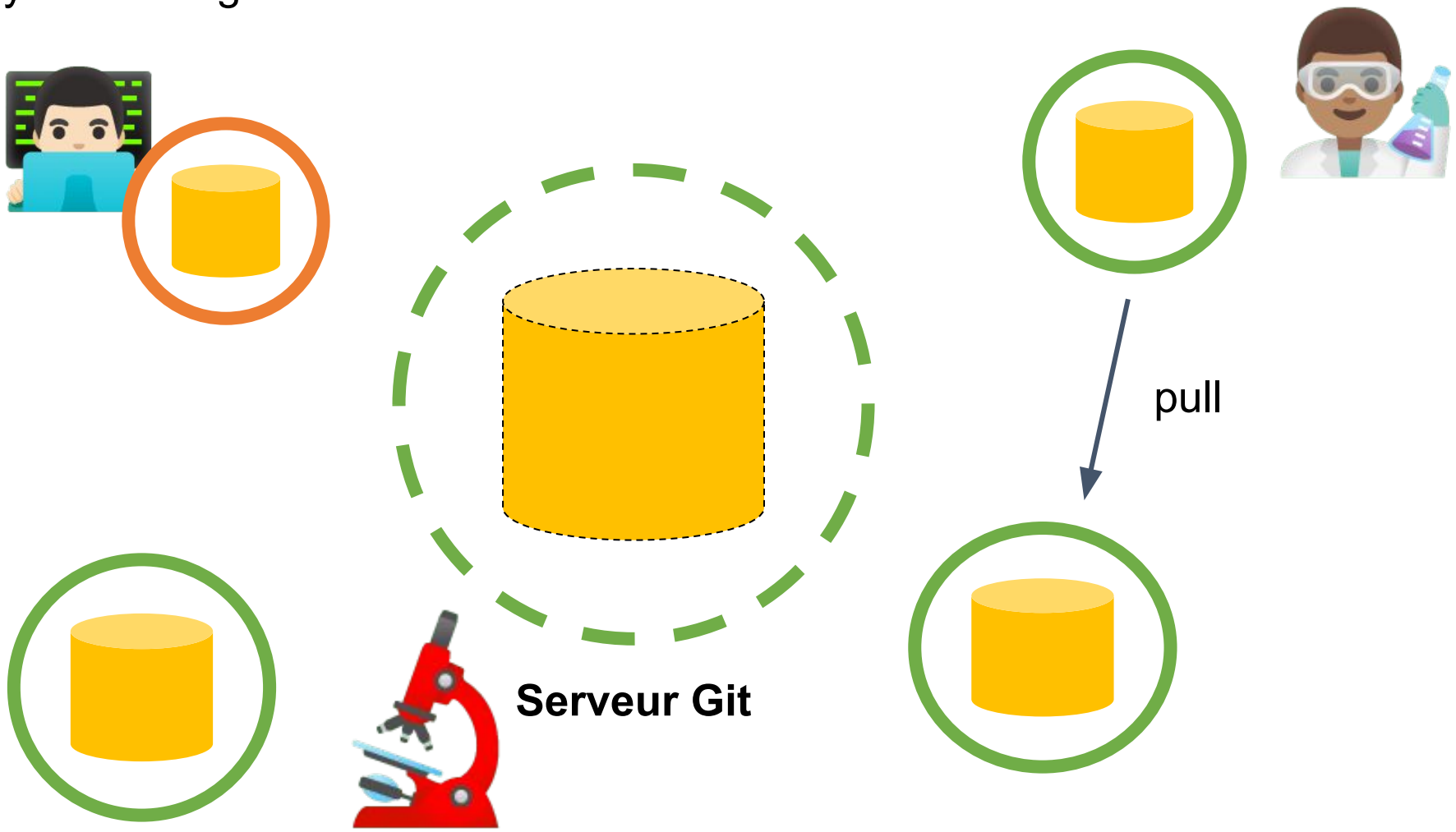
C’est quoi Git ?

Git est un système de gestion de versions DISTRIBUÉ

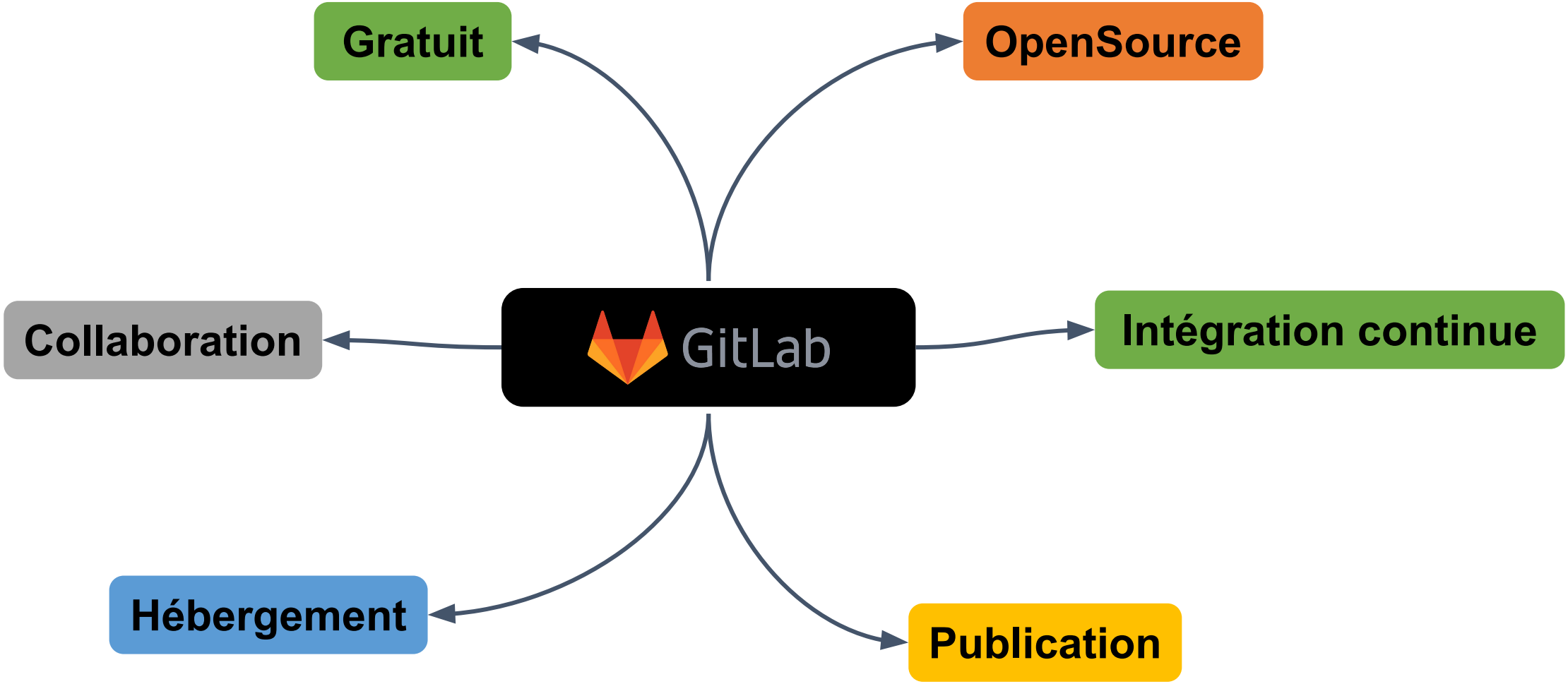


C’est quoi Git ?

Git est un système de gestion de versions DISTRIBUÉ



C’est quoi Git ?



Merci !
mais au fait…
vous les avez tous ?

La vie aquatique

Contagion

Alien, le retour

Indiana Jones

Bruce Tout Puissant

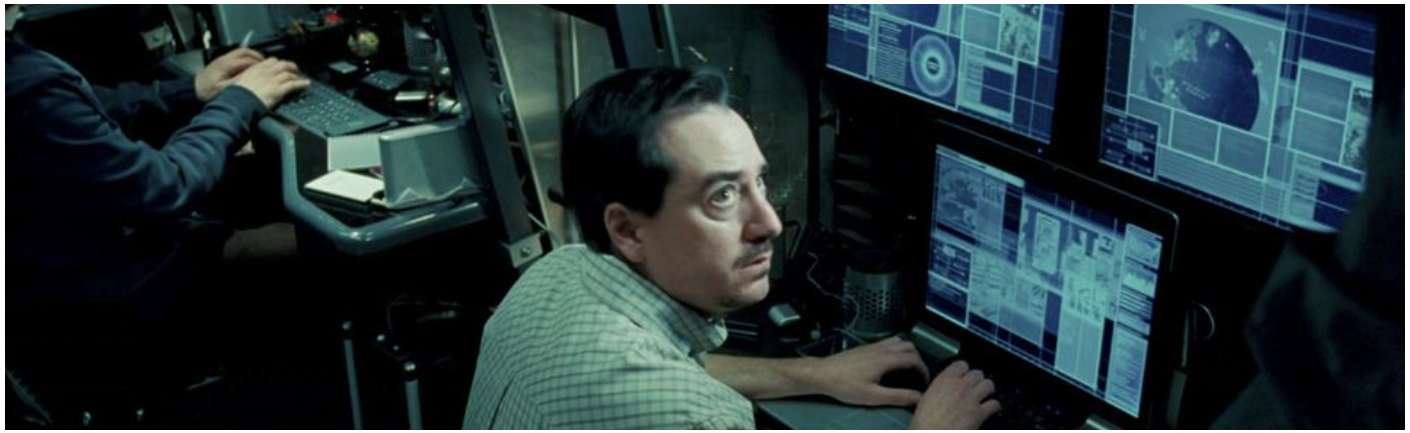
Pulp Fiction

La Momie

Matrix, reloaded

Wall-E

Lucy



Die Hard, retour en enfer