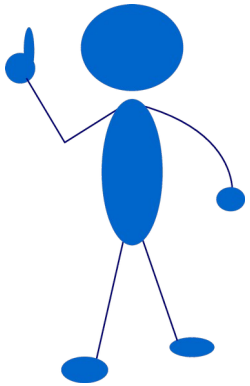
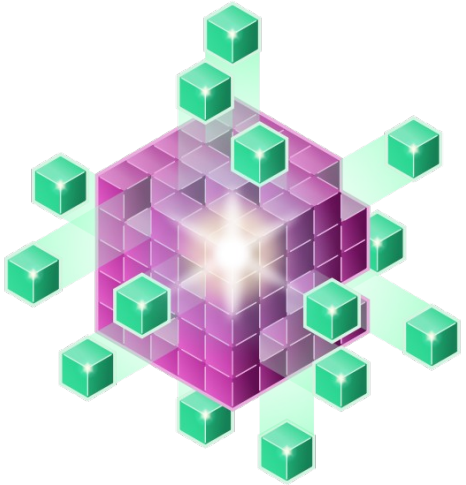
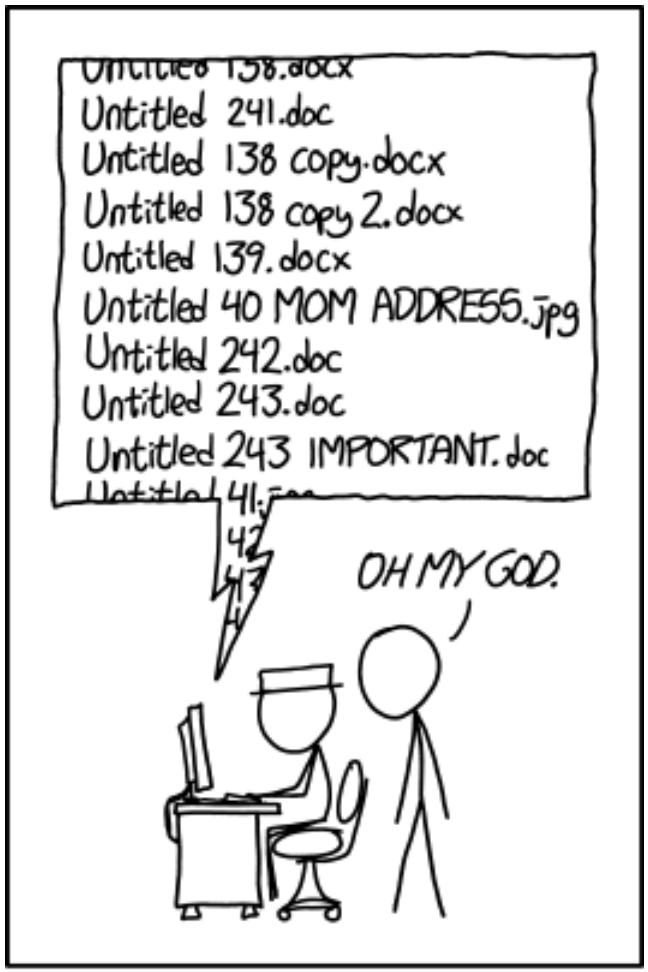
CELLULE DATA
GRENOBLE ALPES

Comment se faciliter ses données ?

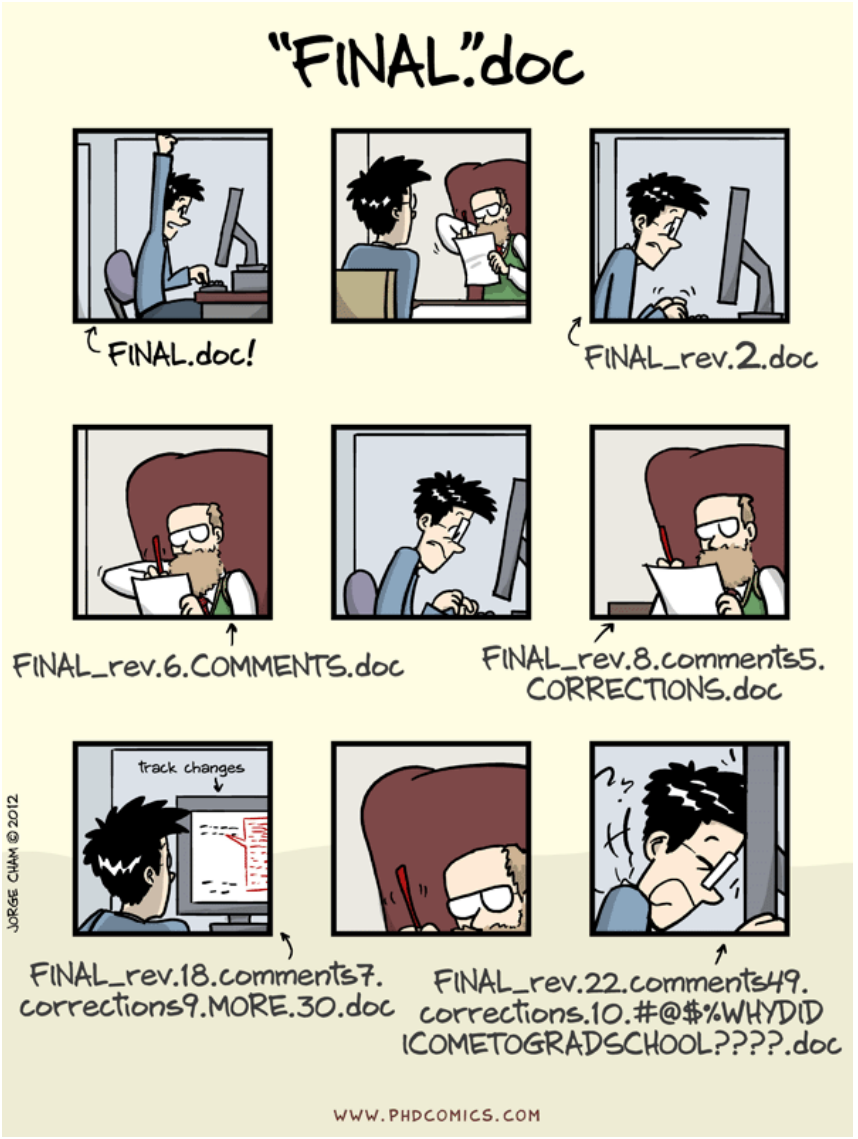


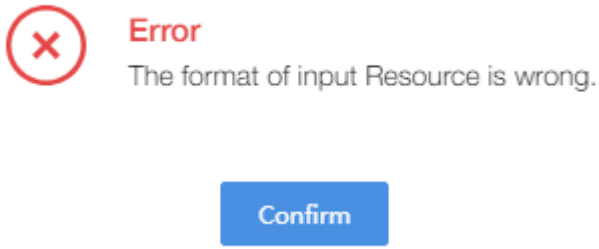


Mai 2023

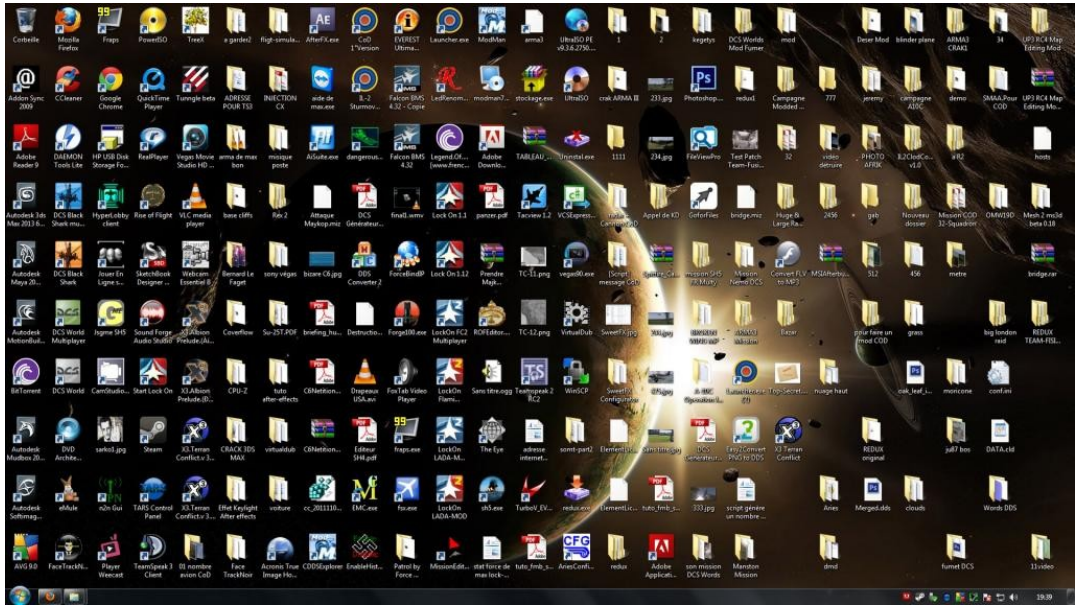


PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.




Error
The format of input Resource is wrong.
Confirm


HOME
LOG IN
WARNING
Are you sure you want to delete file?
CANCEL
OK



Quizz

- **Lequel de ces titres de fichier est le plus pérenne et adapté ?**
  - Ma thèse
  - ma_these
  - These_Martin_Dupond
  - These_Martin_Dupond_Histoire_UGA_2019
  - These_Martin_Dupond_Histoire_UGA_2019_sur_les_chevaliers_paysans_de_l'an_mil_au_lac_de_Paladru
- **Lequel de ces formats de fichier n’est pas ouvert ?**
  - gif
  - mp3
  - zip
  - txt
  - ppt

Pourquoi l’organisation de ses données est importante ?

- Avez vous déjà ouvert un fichier sans comprendre ce qu’il y avait dedans ?
- Avez vous déjà travaillé sur la mauvaise version d’un fichier ?
- Avez-vous déjà écrasé un fichier ?
- Avez vous déjà été dans l’impossibilité de retrouver un fichier ?
- Avez vous déjà été confronté au fait de ne plus pouvoir ouvrir un fichier ?

Pourquoi l’organisation de ses données est importante ?

- Avez vous déjà ouvert un fichier sans comprendre ce qu’il y avait dedans ?
- Avez vous déjà travaillé sur la mauvaise version d’un fichier ?
- Avez-vous déjà écrasé un fichier ?
- Avez vous déjà été dans l’impossibilité de retrouver un fichier ?
- Avez vous déjà été confronté au fait de ne plus pouvoir ouvrir un fichier ?

Les points importants : FAIR pour soi-même

- **L’organisation des données**
  - Etre en capacité de retrouver facilement et rapidement n’importe quel contenu : Findable / Accessible
- **Les formats de fichiers utilisés**
  - Etre en mesure d’ouvrir ses fichiers même si on change de système ou plusieurs années après : Accessible / Interoperable
- **Le nommage des fichiers**
  - Etre capable d’identifier rapidement le contenu d’un fichier : Interoperable / Reusable

Réfléchir à l’organisation de ses données

- Prendre du temps pour en **gagner** ensuite
- Etre **pragmatique et réaliste** : faire ce qu’il faut sans en faire trop !
- Pas de règle : la bonne organisation est celle qui **vous convient**. La perfection n’existe pas !
- Mais il faut **prendre en compte le contexte** : qui va utiliser les données ? Vous seul, l’ensemble de l’équipe, tous les partenaires du projet …
- N’oubliez pas de **décrire l’organisation** que vous aurez choisie, même si ce n’est que pour vous même !

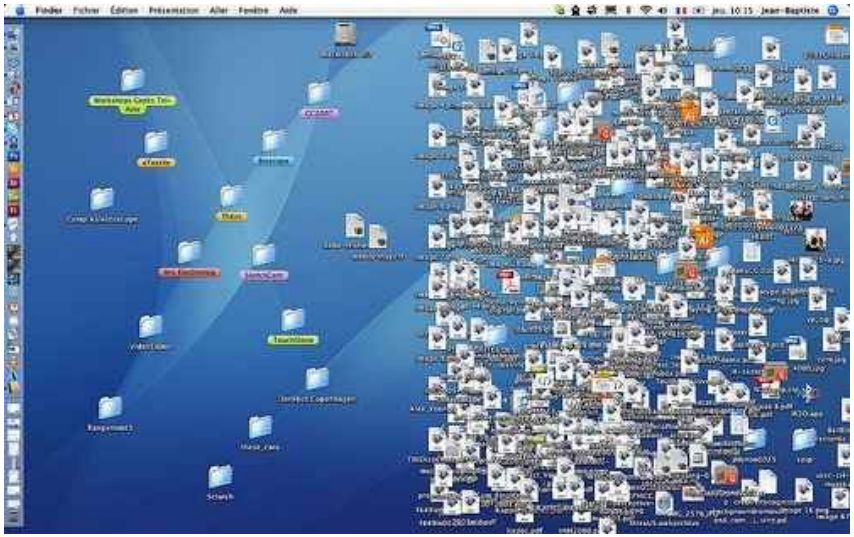
En résumé, les différentes étapes :

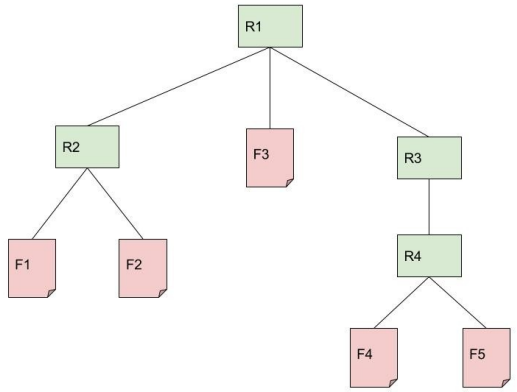
- Déterminer la **structure de l’organisation**
- Définir la **convention de nommage**
- Spécifier la **façon de gérer les versions**
- Identifier les **formats de fichiers** des différents types de données

Structure de l’organisation

Organisation hiérarchique

- Mettre en place une **arborescence** à base de répertoires et sous-répertoires en regroupant les données logiquement
- Ne pas hésiter à mettre un fichier **Readme** dans chaque répertoire pour expliquer le type de données qui s’y trouvent
- Eviter une **trop grande profondeur**





Nommer les répertoires et fichiers

- De façon à ce qu'on en **identifie directement** le contenu. Par exemple pour les données d'un projet de recherche :
  - Nom ou acronyme du projet (identifiant unique)
  - Nom ou initiales du chercheur
  - Date de la constitution des données
  - Type de données
  - Conditions de collecte
- Et toujours construit de la **même manière**
  - Les informations dans le même ordre, la plus importante en premier
  - Penser qu'on peut vouloir traiter automatiquement les fichiers (scripts, codes)
  - Documenter la construction des noms des répertoires et fichiers

Bonnes pratiques de nommage

- Ne pas répéter les informations déjà contenues dans les noms de répertoire
  - Des **noms les plus courts** possibles
- Utiliser la **langue la plus adaptée** à l'équipe de recherche
- Pas d'**espace ni de caractère spécial** (y compris éviter les accents, cédille …) : utiliser des majuscules pour séparer les mots ou des « _ » (underscore)
- Dates au format **AAAMMJJ** (facilite l'affichage par ordre chronologique)
- Chiffres à écrire avec le nombre de **caractères significatifs** :
  - Pour une séquence de 1à 10 : 01-10
  - Pour une séquence de 1 à 100 : 001-100

NomProjet/NomExperience/Data/20180524_Temperature_s29.csv

→ collecte de données de température issues du capteur 29 le 24 mai 2018 dans le cadre de l'expérience « NomExpérience » du projet « NomProjet »

Bon, pas bon ??

1

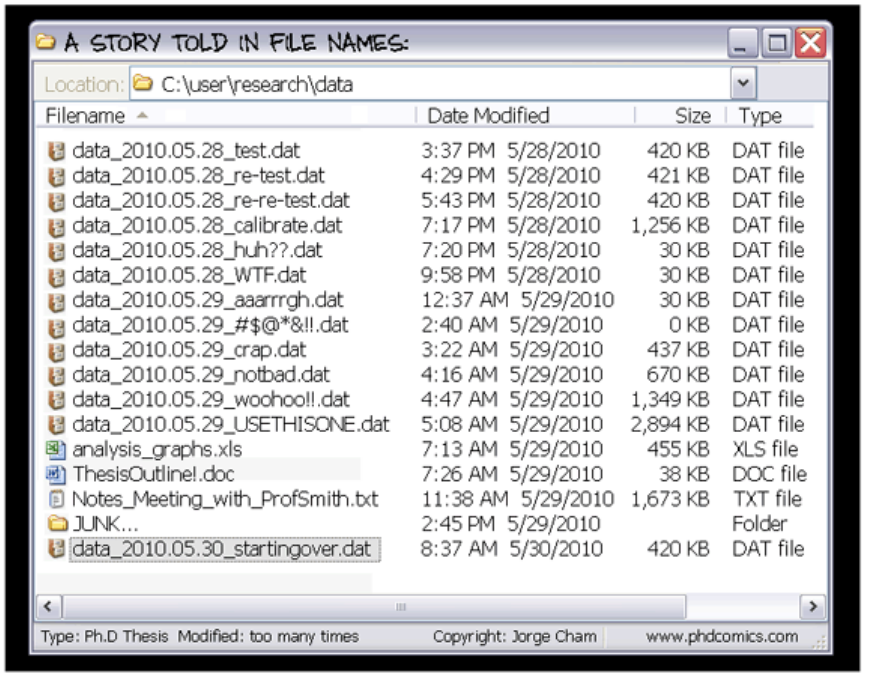
1) réunion 11 février 2021.doc
2) 20210211_CR.odt

2

1) script.py
2) clustering_analyses.py

3

1) 20160114_recording_001.wav
2) Enregistrement du 14/01.wav



4

1) Sans nom 2
2) Readme.txt

5

1) 20160318_dmp_old.odt
2) 20160318_dmp_v02.odt

Bon, pas bon ??

1

1) réunion 11 février 2021.doc
2) 20210211_CR.odt

2

1) script.py
2) clustering_analyses.py

3

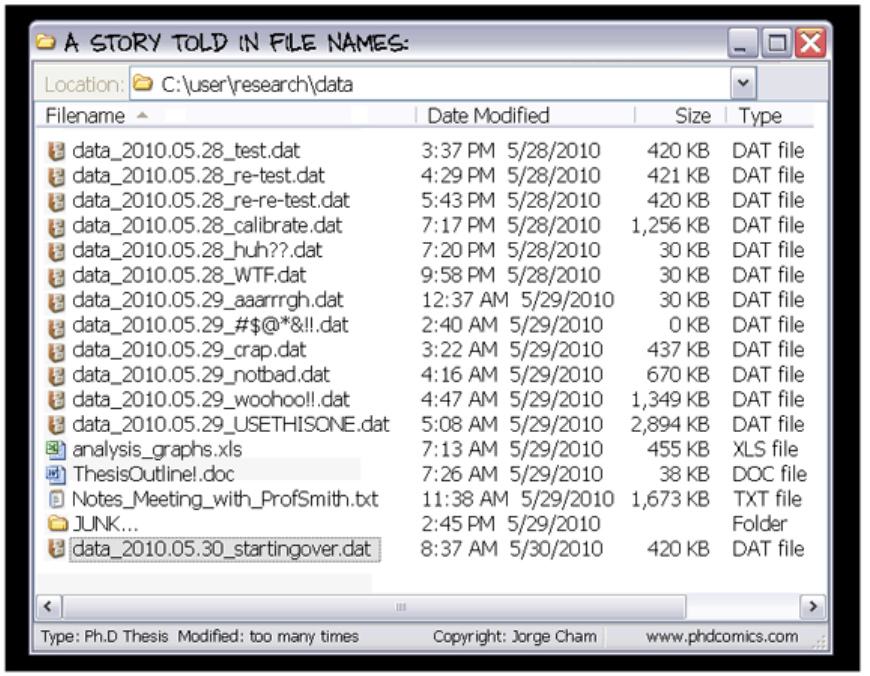
1) 20160114_recording_001.wav
2) Enregistrement du 14/01.wav

4

1) Sans nom 2
2) Readme.txt

5

1) 20160318_dmp_old.odt
2) 20160318_dmp_v02.odt



La gestion des versions

- Certains documents peuvent être amenés à évoluer, à avoir plusieurs versions avant une version finale, définitive.

- Deux solutions :

  - préciser la version dans le **nom du fichier**
  - Utiliser un **gestionnaire de version** (important quand plusieurs personnes sont susceptibles de modifier les fichiers)

- Pourquoi ?

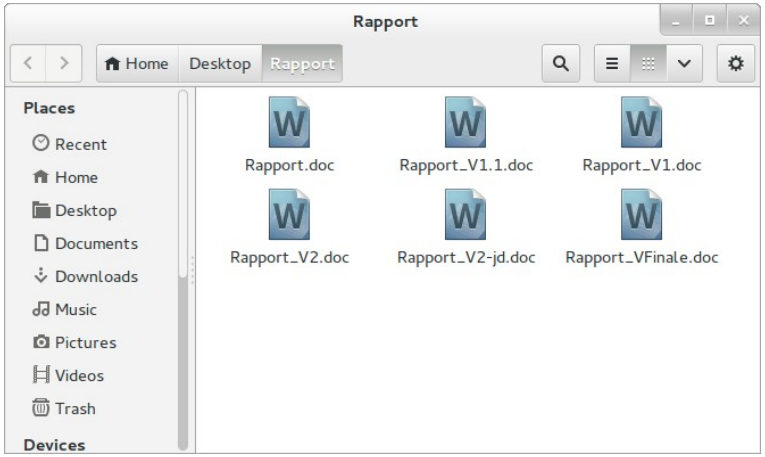
  - Pouvoir rétablir une **version antérieure** en cas d'erreur
  - Pouvoir accéder à l'**historique des modifications** et donc de la vie du fichier
  - Pouvoir identifier les **auteurs des modifications** (utilisation d'un gestionnaire de version)
  - Eviter de travailler sur une **version périmée** du fichier

Gestion des versions via le nommage des fichiers

- Etre **cohérent et consistant** avec les règles globales mises en place
- Pour faire simple, ne rédigez que la lettre **V suivi du numéro de la version**, en deux unités
  - Evolution majeure : fichier_v03
  - Modification mineure : fichier_v03-02
- Autre convention de nommage possible : VP pour Version Provisoire, VF pour Version Finale.
- Utilisation des **dates** également possible mais attention aux versions réalisées le même jour !
- Déposer les versions précédentes dans un dossier **Archives**
- **Documenter** les règles de nommage des versions

Pour tout document de type txt (code source, markdown, article en latex…) → utiliser un **gestionnaire de version** !!

Pourquoi utiliser un gestionnaire de version ?



- La version la plus à jour est-elle Rapport.doc ou Rapport_VFinale.doc ?
- Les versions n'apparaissent pas dans l'ordre (1.1, 1, 2)
- La version 2-jd vient elle avant ou après la version 2 ?

Source : https://perso.liris.cnrs.fr/pierre-antoine.champin/enseignement/intro-git/

- La gestion des versions est un travail fastidieux et méthodique.
- Les humains ne sont pas doués pour les travaux fastidieux et méthodiques.
- Laissons cela à l'ordinateur ...

✓ Fichiers texte : fichiers markdown, fichiers source, fichiers latex, ...

✗ Fichiers images, pdf, exécutables, ...

Gestion de version avec git

- Initialement dédiés à la **gestion de code source** pour les projets logiciels mais également :
  - Documentation
  - site web (wiki, git pages)
  - travail collaboratif : facilité d'échange, traçabilité, gestion des conflits

- **Principe** : un fichier est une référence à de l'information qui est physiquement stockée dans la mémoire d'un dispositif individuel ou d'un serveur collectif. Lorsque cette information change, le fichier est modifié : la référence pointe maintenant à de l'information qui, physiquement, est stockée de manière différente par rapport à la version précédente.
  Le système de gestion de versions a comme objectif de pouvoir retracer des **modifications physiques** dans la mémoire et associer ces changements au même fichier ou ensemble de fichiers, sans avoir à créer des nouvelles références (i.e. des fichiers avec des noms différents) à chaque passage.

- Ces changements sont stockés dans un **dépôt** (repository).

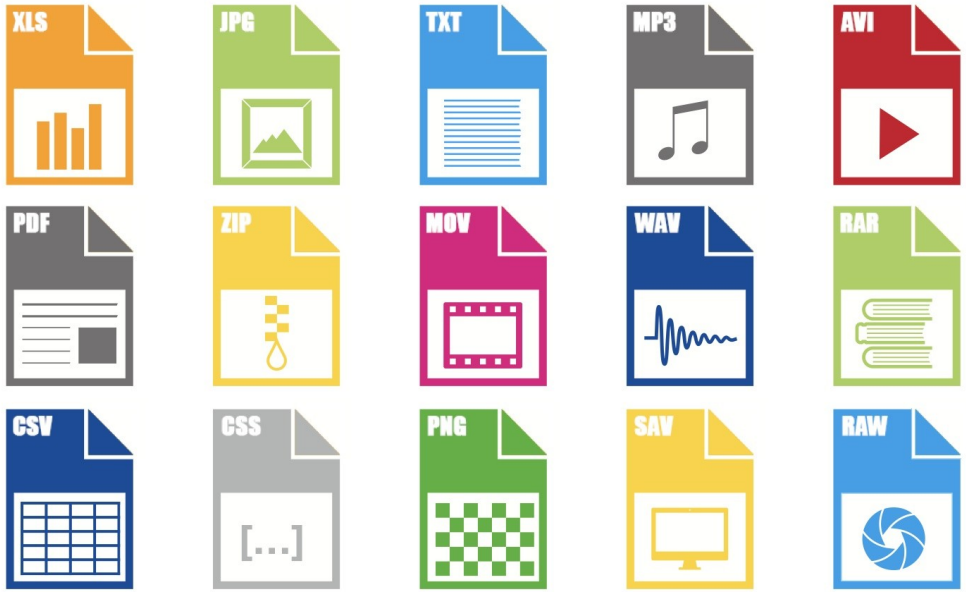
- Git fonctionne en créant **deux dépôts des changements** :
  - Le premier se trouve sur la même machine des fichiers de travail
  - Le deuxième se trouve dans une autre machine, souvent un serveur ou une plateforme cloud comme GitLab, qui s'occupe de centraliser les changements

- Pour aller plus loin :
  - https://www.adum.fr/script/formations.pl?mod=349076&site=UDG
  - https://pole-calcul-formation.gricad-pages.univ-grenoble-alpes.fr/ced/

Problématique du format des fichiers

- **Format** = manière dont les données sont structurées, organisées et encodées sur le support physique. Mode de représentation et de stockage des données.

- En général, défini par **l'extension** dans le nom du fichier



Pourquoi c’est important ?

- Etre capable d’ouvrir les fichiers sur un **maximum de types de système** avec un maximum de logiciels
  - Qui n’a pas été confronté au problème de l’ouverture d’un fichier envoyé par un collègue ?
- Etre capable de le faire **maintenant et dans 5 ans**
  - Qui n’a pas été confronté à une version de format de fichier qui n’est plus supportée par le logiciel ?
- Pouvoir **partager** ses données au sein de son équipe de recherche ou de façon plus large
- Pouvoir traiter et analyser les données grâce à **différents logiciels**, être capable de croiser des données de différentes sources, et donc différents fichiers, éventuellement avec des formats distincts
  - Assurer **l’interopérabilité** des données

Ouvert / fermé

- **Format ouvert** = mode de représentation (spécifications techniques) rendu public par son auteur et aucune entrave légale ne s'oppose à sa libre utilisation (droit d'auteur, brevet, copyright).

- **Standard ouvert** = format ouvert ou libre qui a été approuvé par une organisation internationale de standardisation.
  Plusieurs organisations de standardisation acceptent certaines formes de limitations à la diffusion de leurs standards : un standard ouvert peut par conséquent être basé sur un format ouvert mais non-libre.

- **Définition légale en France** (loi du 21 juin 2004 sur l'économie numérique) : « 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. »

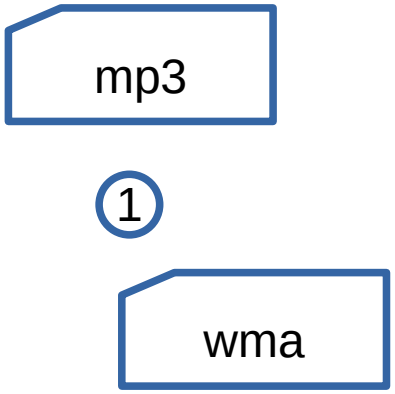
- **Format propriétaire ou fermé** = format de données dont les spécifications sont contrôlées par des intérêts privés. Souvent l'objet d'un brevet.
  Spécifications non connues, auquel cas il est difficile de développer des logiciels qui puissent lire ou écrire ce type de format, ou spécifications diffusées, mais des restrictions légales associées à son utilisation existent, et son évolution reste contrôlée par son propriétaire.
  Dépendant du logiciel développé par l'entreprise.

Les données ont de la valeur : ne pas se faire piéger par un format fermé qui contraindra la liberté de choix quant aux programmes, libres comme propriétaires, qu'on voudra utiliser.

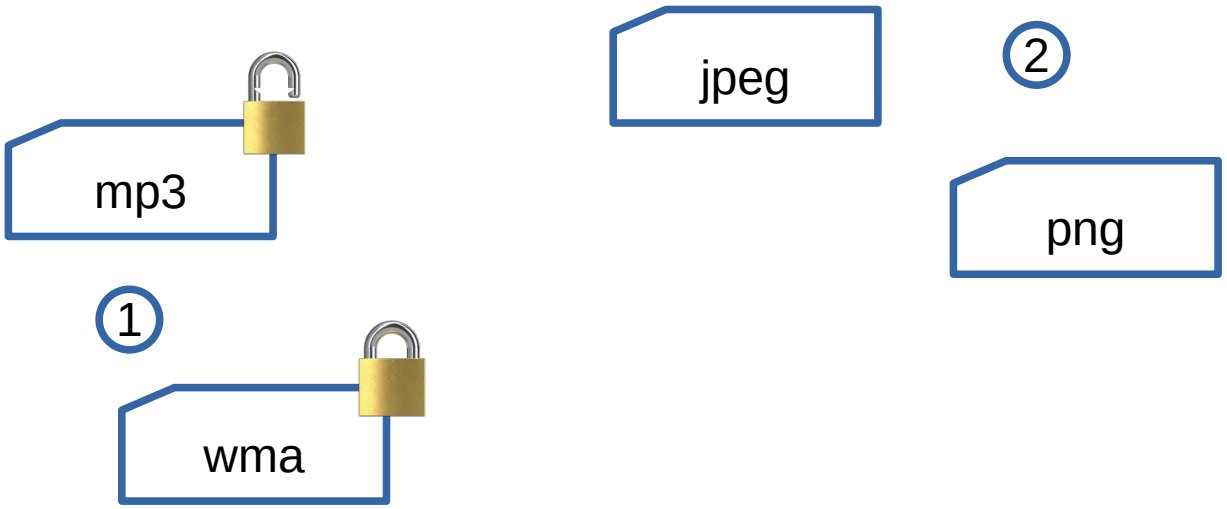
Comment savoir quel format utiliser ?

- Règle 1 : utiliser des **formats ouverts**
- Règle 1bis : utiliser des **formats ouverts**
- Règle 2 : utiliser les formats **les plus fréquents** dans votre communauté
- Beaucoup de **liste de formats ouverts** sur le web
  - Par exemple https://doranum.fr/stockage-archivage/quiz-format-ouvert-ou-ferme/
- Un site intéressant : https://facile.cines.fr/
  - outil qui permet de vérifier si un fichier est **valide et bien formé** par rapport au format déclaré, et donc de savoir s'il est éligible à l'archivage proposé par le CINES.
- Lien entre format de données et **préservation numérique** à la BnF : https://www.bnf.fr/fr/formats-de-donnees-pour-la-preservation-numerique

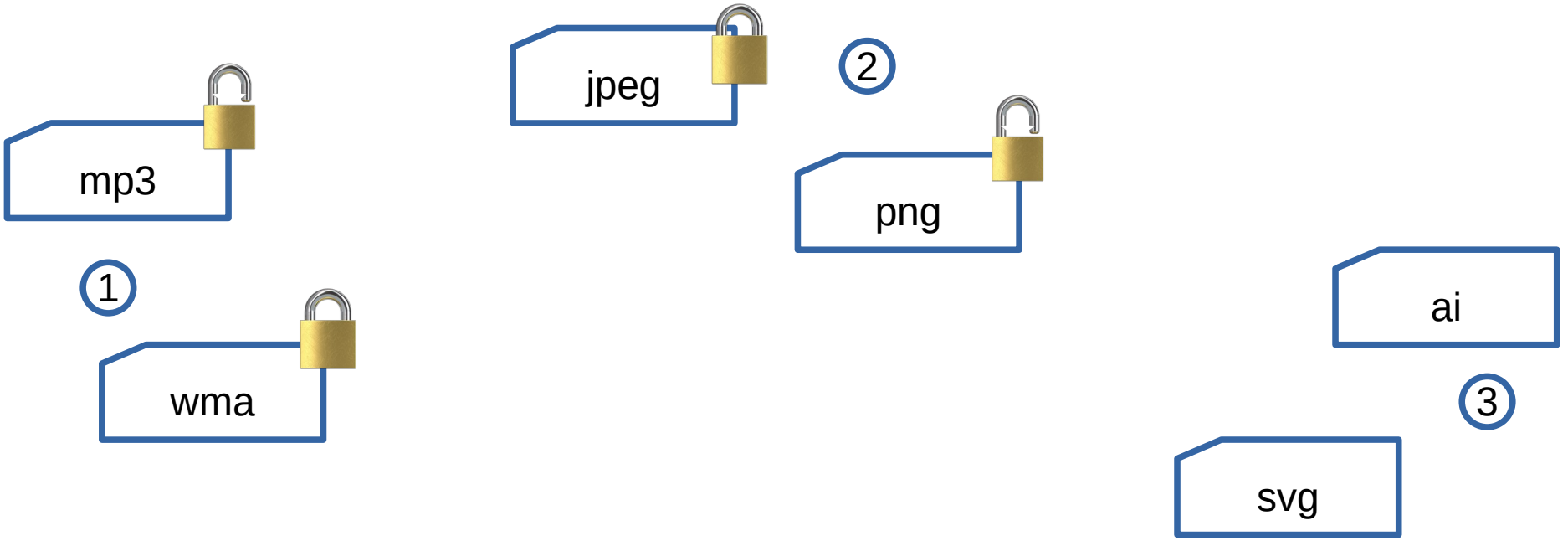
Ouvert ou fermé ??



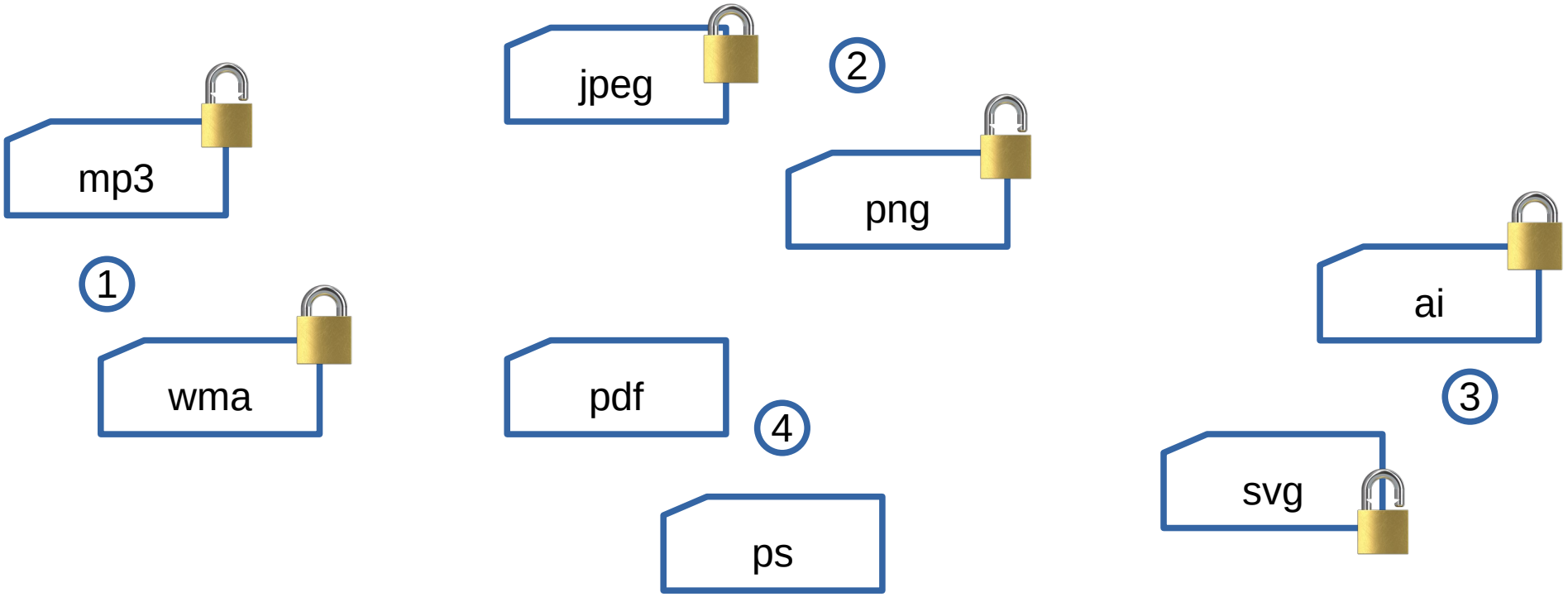
Ouvert ou fermé ??



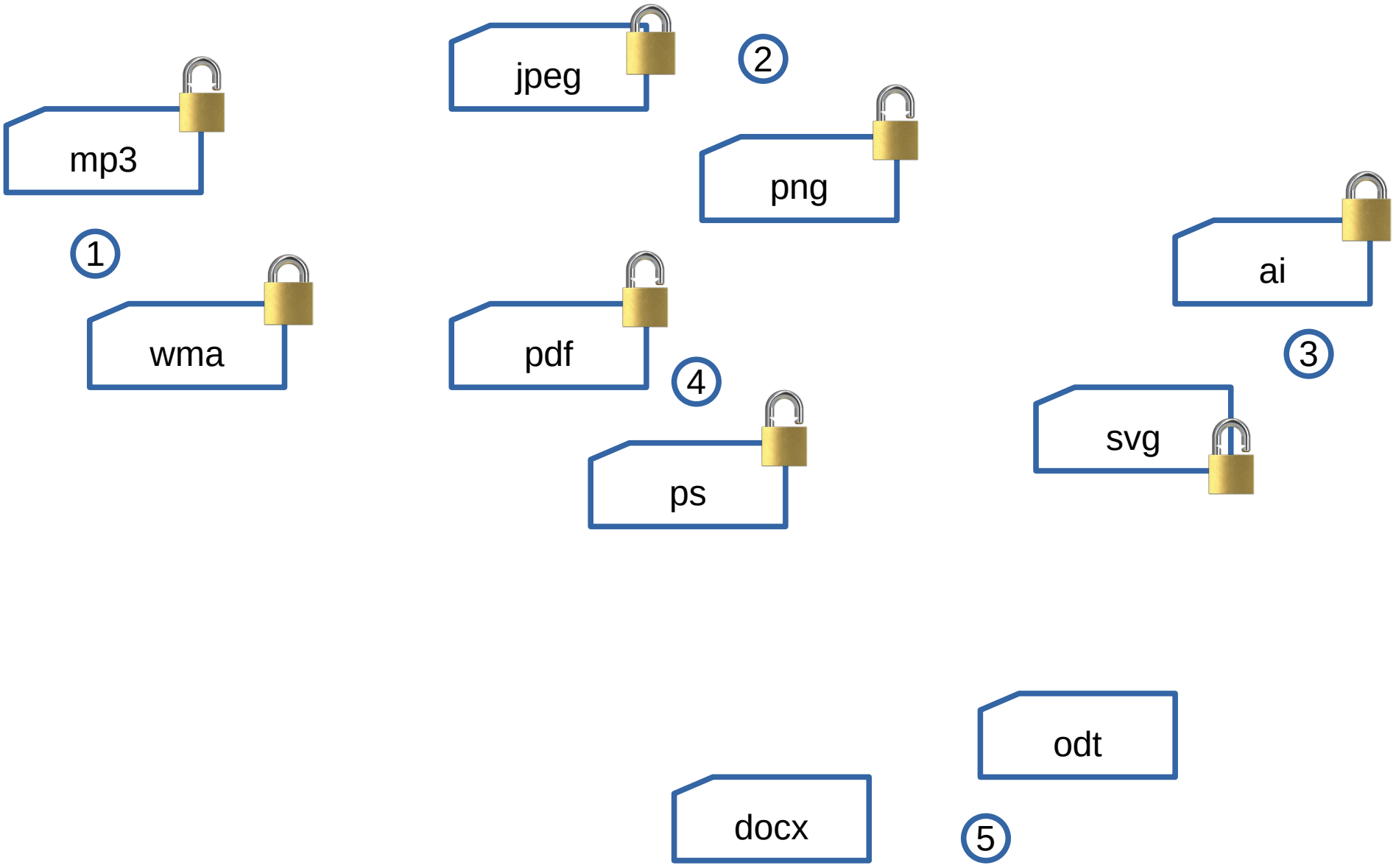
Ouvert ou fermé ??



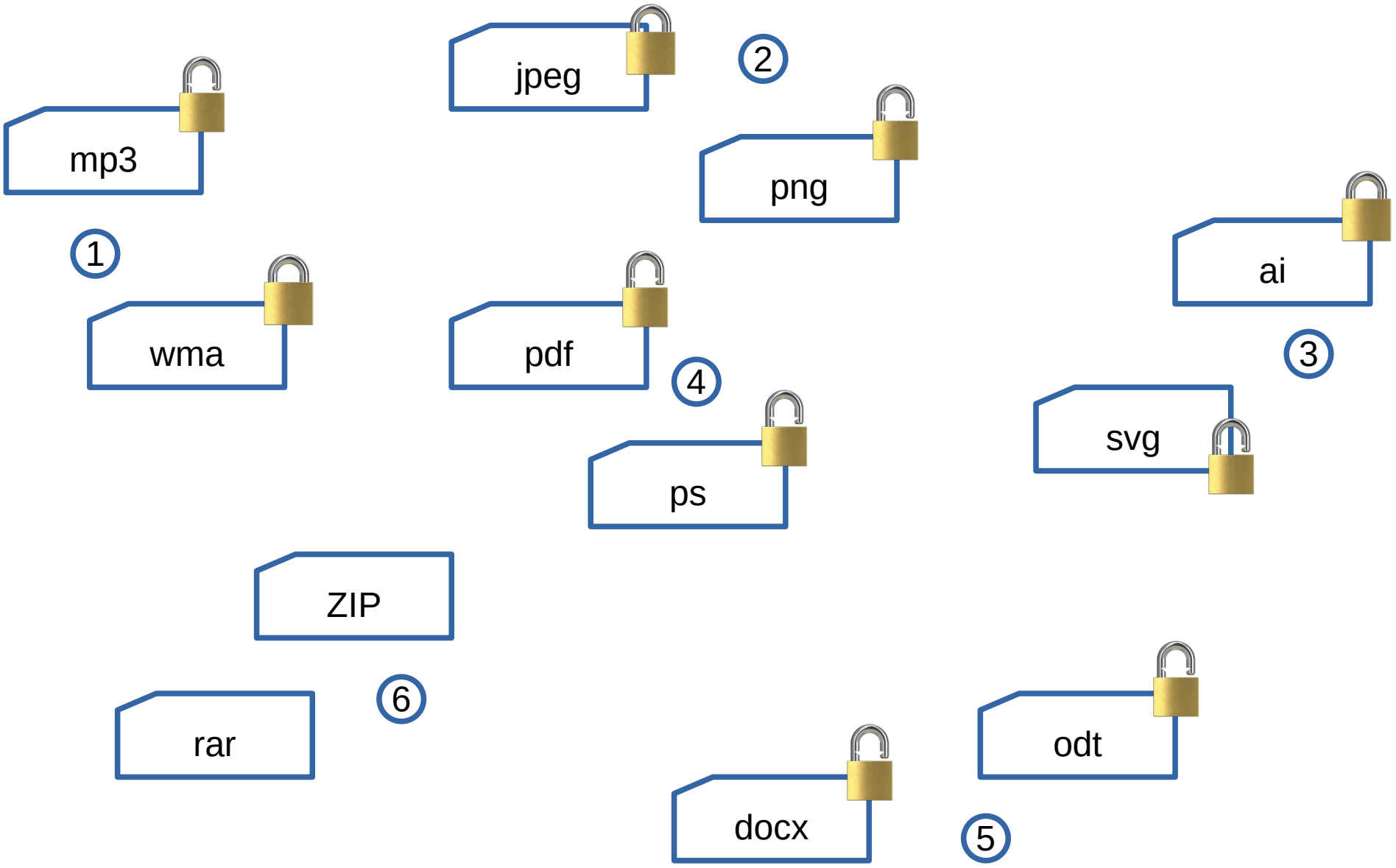
Ouvert ou fermé ??



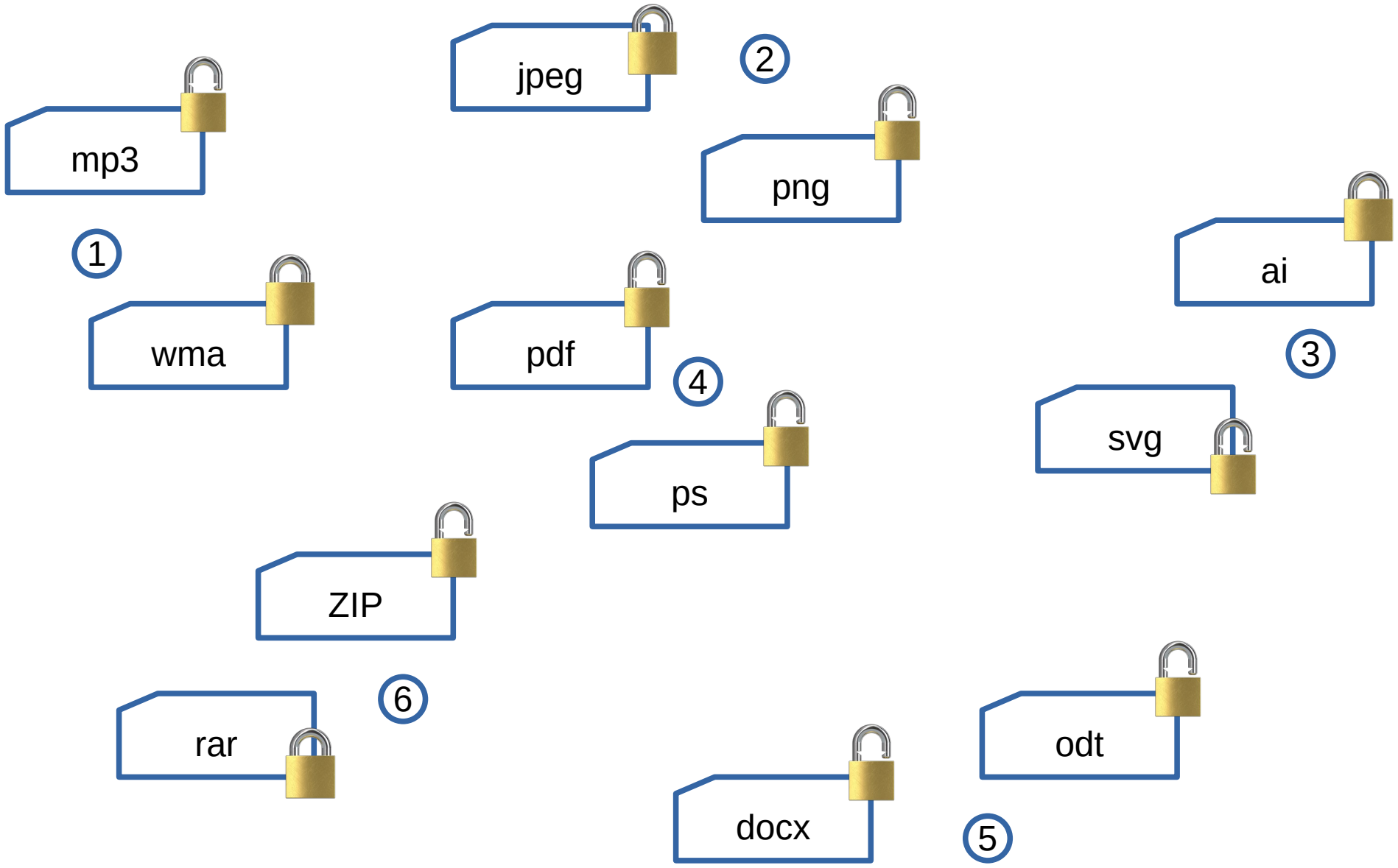
Ouvert ou fermé ??



Ouvert ou fermé ??



Ouvert ou fermé ??



Pour ceux que ça intéresse

- **Format jpeg** : Dans la pratique, seule la partie concernant le codage arithmétique est brevetée, et par conséquent protégée par IBM, son concepteur. Plusieurs tentatives pour le moment infructueuses pour récupérer la paternité du format par l'entreprise Forgent puis la société Global Patent Holdings, filiale d'Acacia Research Corporation (qui ne vit que des dommages et intérêts d'actions de ce type).

- **Format docx** : Office Open XML est une norme ISO/CEI 29500 créée par Microsoft, destinée à répondre à la demande d'interopérabilité dans les environnements de bureautique et à concurrencer la solution d'interopérabilité OpenDocument soutenue par tous les autres éditeurs de suites bureautiques, notamment Apache et The Document Foundation. La normalisation a provoqué une certaine polémique avec pas mal de mise en cause aussi bien techniques que politiques (voir la page de wikipedia : https://fr.wikipedia.org/wiki/Office_Open_XML)

- **Format pdf** : langage de description de page présenté par la société Adobe Systems en 1992 et qui est devenu une norme en 2008 sous l'appellation ISO 32000. Gérée par l'ISO (Organisation internationale de normalisation), la norme ISO 32000 est développée dans le but de protéger l'intégrité et la longévité du format PDF. L'évolution du format PDF ne sont donc plus du ressort des seules décisions de la société Adobe Systems.

- **Format mp3** : Cette technologie a fait l'objet de brevets et d'une licence commerciale. Jusqu'à 20 brevets, expirant entre 2007 et décembre 2017 aux États-Unis, date à laquelle le format est entré dans le domaine public (c'est-à-dire totalement libre de droits).

Quizz : à vous cette fois !

- **Lequel de ces titres de fichier est le plus pérenne et adapté ?**
  - Ma thèse
  - ma_these
  - These_Martin_Dupond
  - These_Martin_Dupond_Histoire_UGA_2019
  - These_Martin_Dupond_Histoire_UGA_2019_sur_les_chevaliers_paysans_de_l'an_mil_au_lac_de_Paladru
- **Lequel de ces formats de fichier n’est pas ouvert ?**
  - gif
  - mp3
  - zip
  - txt
  - ppt

Réponses !

- **Lequel de ces titres de fichier est le plus pérenne et adapté ?**
  - Ma thèse
  - ma_these
  - These_Martin_Dupond
  - **These_Martin_Dupond_Histoire_UGA_2019**
  - These_Martin_Dupond_Histoire_UGA_2019_sur_les_chevaliers_paysans_de_l'an_mil_au_lac_de_Paladru
- **Lequel de ces formats de fichier n'est pas ouvert ?**
  - gif
  - mp3
  - zip
  - txt
  - **ppt**

En pratique ! Range ta chambre

- Petit travail en groupe de **2 à 4 personnes**
- Chacun présente au reste du groupe la façon dont il **organise les données de sa thèse** (sans tricher …), oralement ou directement sur son ordinateur
- Le groupe pointe les **bonnes pratiques, les freins** et propose des **pistes d'amélioration**
- Restitution de chaque groupe à tout le monde sous le format :
  - **2 bonnes pratiques** identifiées et déjà mises en œuvre
  - **2 problématiques** à améliorer
  - **2 pistes d'amélioration**

A vous de jouer !