ON FAIT LE POINT sur les **données de la recherche** avec **Sorella !**

Université de Guyane


SORELLA !
JEAN-PAUL !


TIENS,
TOUT EST LÀ


DONNÉESLABOMODIFS8
DONNÉESLABO13
EXTRACTION LABO 15
MISE À JOUR EXTRACTION 8
DONNÉES RELECTURE JEANNE
EXTRACTIONS RELECTURE TOUFIK B5


JEAN-PAUL ! QU'EST-CE QUE C'EST QUE TOUS CES TRUCS À LA FIN ? QU'EST-CE QUE TU AS FAIT ?


AH ! MOI, ON M'A DIT : POUR GÉRER LES DONNÉES, FAUT DU TABLEUR EXCEL ! AH LÀ, Y A DU TABLEUR EXCEL ! 60 HEURES DE BOULOT, AVEC TROIS STAGIAIRES EN SUPPORT !
MAIS TU PENSES RÉELLEMENT QU'UNE PERSONNE NORMALEMENT CONSTITUÉE PEUT UTILISER ÇA ???


AH BEN ! J'ESPÈRE BIEN QU'ILS PEUVENT LIRE ÇA PUISQU'ILS LE DEMANDENT ! OU ALORS, C'EST JUSTE UNE LUBIE QUI LES PREND, D'EMBÊTER LE MONDE ? COMME SI ON N'AVAIT QUE ÇA À FAIRE DE PONDRE DES TRUCS INUTILES DONT TOUT LE MONDE SE FOUT !


GROMLMLMLLL...
OK, JE VOIS. SI ON NOUS DEMANDE DE GÉRER LES DONNÉES DE RECHERCHE DE NOTRE PROJET, C'EST QUE CE N'EST PAS UN TRUC INUTILE, TU SAIS ! ON VA REPRENDRE CELA ENSEMBLE DEPUIS LE DÉBUT...

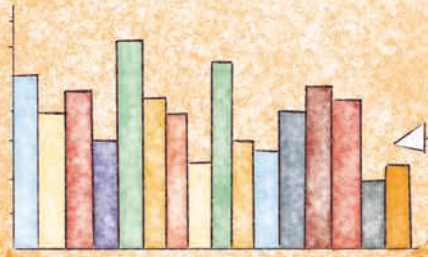
1. Qu'est-ce que des données de la recherche ?

EN FAIT, IL EXISTE UN GRAND NOMBRE DE TYPES DE DONNÉES, QUI NE SONT D'AILLEURS PAS LES MÊMES SELON LES THÈMES TRAITÉS ET LES DISCIPLINES COUVERTES. SELON LE CONTEXTE, ELLES PEUVENT ÊTRE :

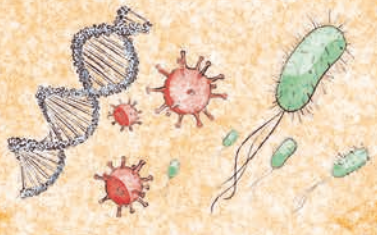
Des données expérimentales
(poids biomasse, séquences peptides, tests en psychologie, etc.)

Des données de simulation numérique
(modèles climatiques, modèles économiques, etc.)

Des données dérivées ou compilées
(bases de données compilées, fouilles de texte, statistiques de population, etc.)

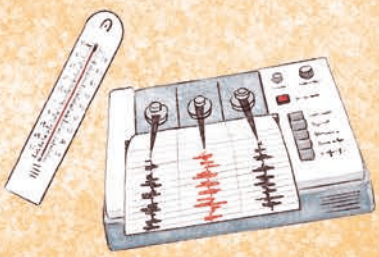


Des données de référence
(corpus textuels, séquences gènes TP53, structures chimiques, etc.)



AH ! TU VOIS ! C'EST EXACTEMENT CE QUE J'AI FAIT AU FINAL : DU TABLEUR EXCEL !

Des données d'observation
(relevés météo, mesures sismiques, images, enquêtes sociales, fouilles archéologiques, etc.)



Des données d'archives
(plaques de verre, fonds de photographies, textes de lois, graines de plantes, etc.)

En fait, selon la définition donnée par l'OCDE en 2007, les données de la recherche concernent « tous les enregistrements factuels (chiffres, textes, images et sons) utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires à la validation des résultats de recherche ».

Mais pour que ces données puissent servir comme preuves pour la communauté scientifique, encore faut-il les traiter, les organiser et les diffuser correctement !

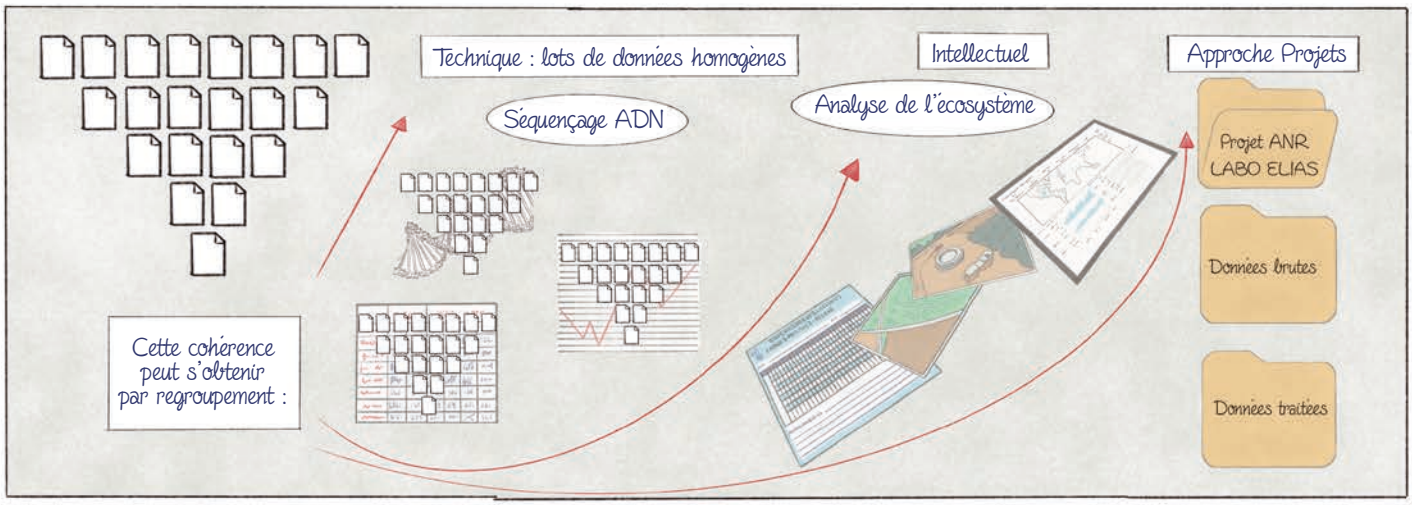
On distingue trois stades de la vie des données.

Stade 1 : les données sont brutes. Elles ont été recueillies, mais pas encore organisées, mises en forme ou analysées

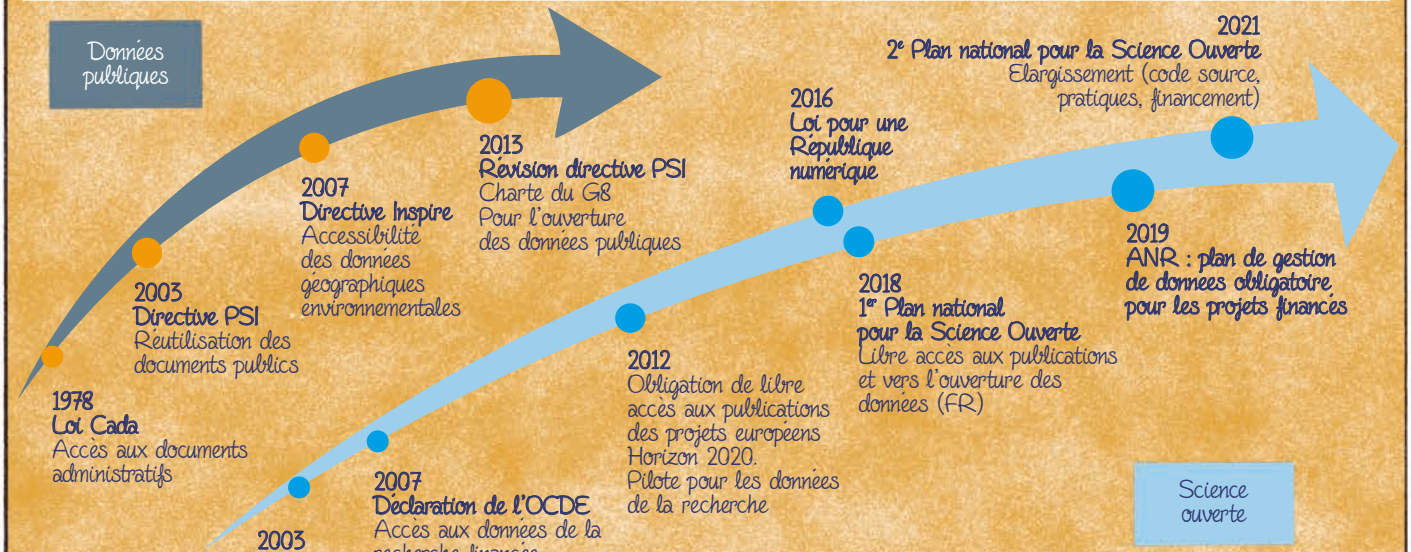
Stade 2 : les données ont été traitées. Elles ont subi par exemple une calibration, un étalonnage ou une correction des données brutes. Elles ont également été dérivées, et présentent un résumé ou une représentation/vue spécifique (agrégation, compilation, calcul, réorganisation, etc.)

Les données ont été traitées et analysées. C'est généralement le dernier stade avant la publication des résultats de la recherche via des articles scientifiques.

Chacun de ces stades peut aussi être le matériau de recherche d'un chercheur, selon ses besoins.

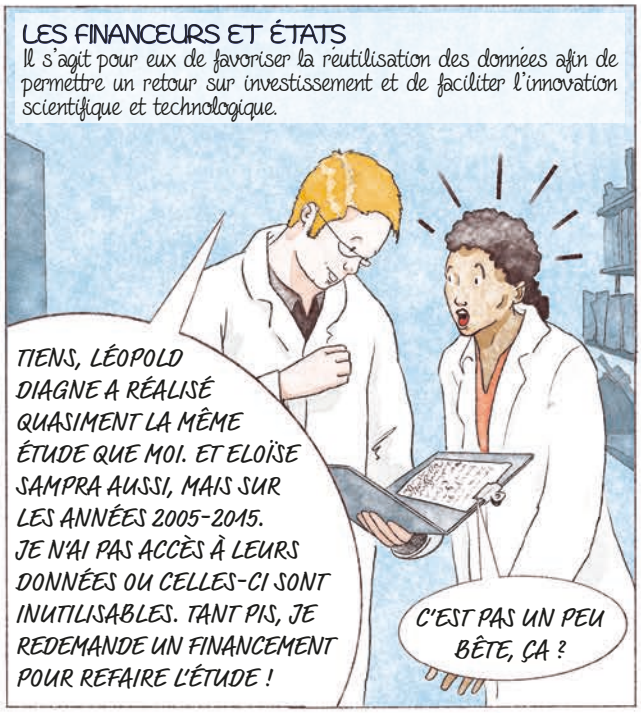


Le mouvement de la Science ouverte – c'est-à-dire l'ouverture des données et des productions scientifiques – est lié de près à celui de données publiques – les données issues de la recherche publique étant par défaut définies comme des données publiques. Ceci est d'ailleurs inscrit dans le Code de la Recherche depuis la Loi pour une République numérique en 2016



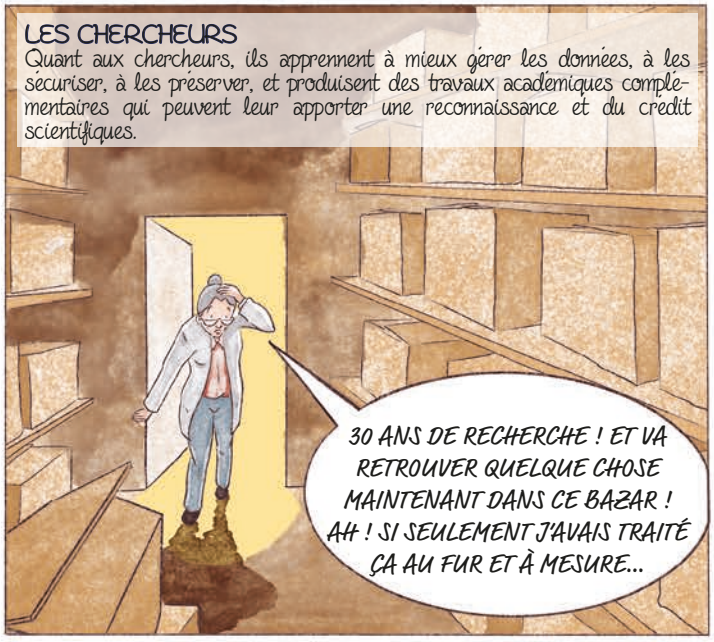
Inscrit dans une dynamique européenne, le 2e axe du 2e Plan national pour la Science ouverte établi pour 2021-2024 préconise aux chercheurs français de « structurer, partager et ouvrir les données de la recherche ». De nombreux financeurs imposent de respecter cette condition pour l'attribution de crédits, ou a minima de déposer un plan de gestion des données (PGD).


PARCE QUE LA SCIENCE OUVERTE EST AU CŒUR DES STRATÉGIES DE CINQ ACTEURS CLÉS QUI Y TROUVENT LEURS INTÉRÊTS.
Les financeurs et États
Les éditeurs scientifiques
Les organismes de recherche et les universités
Les chercheurs
La société civile


LES FINANCEURS ET ÉTATS
Il s'agit pour eux de favoriser la réutilisation des données afin de permettre un retour sur investissement et de faciliter l'innovation scientifique et technologique.
TIENS, LÉOPOLD DIAGNE A RÉALISÉ QUASIMENT LA MÊME ÉTUDE QUE MOI. ET ELOÏSE SAMPRA AUSSI, MAIS SUR LES ANNÉES 2005-2015. JE N'AI PAS ACCÈS À LEURS DONNÉES OU CELLES-CI SONT INUTILISABLES. TANT PIS, JE REDEMANDE UN FINANCEMENT POUR REFAIRE L'ÉTUDE !
C'EST PAS UN PEU BÊTE, ÇA ?

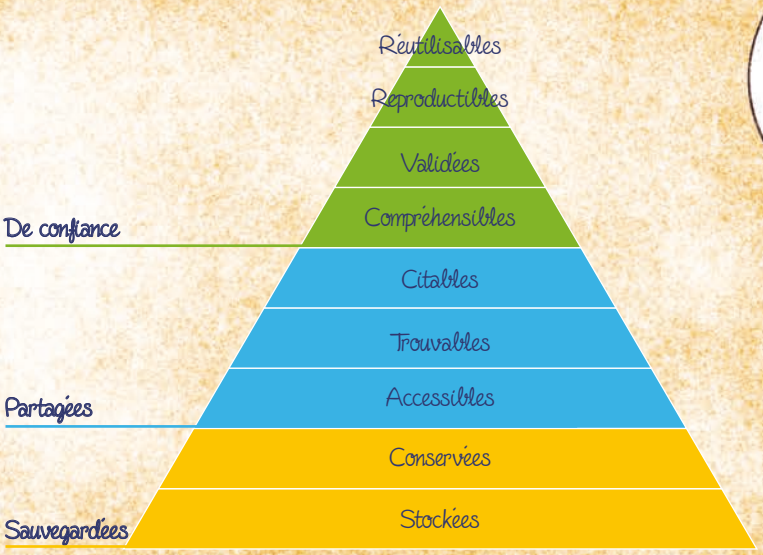

LES ÉDITEURS SCIENTIFIQUES
Outre le devoir de répondre à la demande des financeurs, les éditeurs scientifiques voient dans le partage des données un moyen scientifique d'aboutir à une meilleure validation des travaux en s'assurant notamment de leur reproductibilité et de renforcer la confiance dans les résultats présentés.
C'EST BIZARRE. RAOUL NEWTON A PUBLIÉ UNE ÉTUDE CERTIFIANT QUE LE PLOMB POUVAIT VOLER SOUS CERTAINES CONDITIONS ATMOSPHÉRIQUES TERRESTRES. MAIS PERSONNE N'A ENCORE RÉUSSI À REPRODUIRE L'EXPÉRIENCE AVEC LES DONNÉES QU'IL A PARTAGÉES...


LES ORGANISMES DE RECHERCHE ET LES UNIVERSITÉS
C'est pour eux l'occasion de promouvoir une bonne éthique et une reproductibilité des recherches au sein des laboratoires, en améliorant la gestion des ressources et en suscitant des collaborations entre institutions.
VOICI JEANNINE DUGLAS, SPÉCIALISTE DE LA REPRODUCTION DES DROSOPHILES EN NOUVELLE-GUINÉE, QUI NOUS PRÉSENTE SON NOUVEAU PROJET
AH ! NOUS, ON TRAVAILLE SUR L'ADN DES DROSOPHILES : TU CROIS QU'ON POURRAIT LUI DEMANDER LES DONNÉES ?


LES CHERCHEURS
Quant aux chercheurs, ils apprennent à mieux gérer les données, à les sécuriser, à les préserver, et produisent des travaux académiques complémentaires qui peuvent leur apporter une reconnaissance et du crédit scientifiques.
30 ANS DE RECHERCHE ! ET VA RETROUVER QUELQUE CHOSE MAINTENANT DANS CE BAZAR ! AH ! SI SEULEMENT J'AVAIS TRAITÉ ÇA AU FUR ET À MESURE...


LA SOCIÉTÉ CIVILE
Et enfin, la société civile peut ainsi profiter d'une source d'information fiable, favoriser l'innovation et la participation des citoyens à la recherche.
SUPER ! J'AI TOUTES LES DONNÉES DE GÉOLOCALISATION DES TOILETTES PUBLIQUES DE PARIS : GO POUR UNE APPLICATION DE RÉSERVATION DE PLACES !

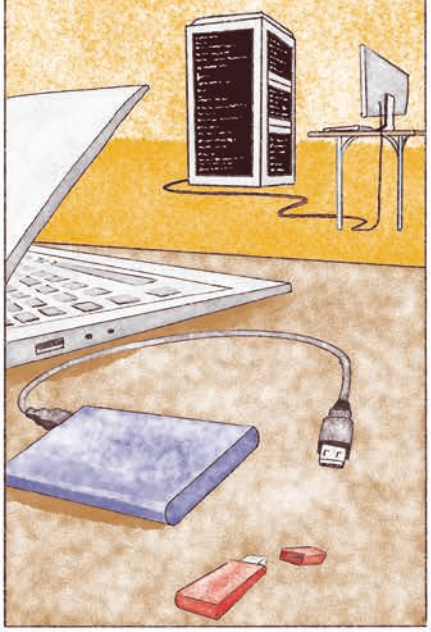
AVANT DE POUVOIR PARTAGER ET DIFFUSER DES DONNÉES – ET AFIN DE LE FAIRE DANS LES MEILLEURES CONDITIONS POSSIBLES – IL Y A DES ÉTAPES DE BONNE GESTION À RESPECTER, RENDANT AU FUR ET À MESURE LES DONNÉES PLUS FIABLES ET DE MEILLEURE QUALITÉ.



D'après Anita de Waard, Helena Cousijn et IJsbrand Jan Aalbersberg, CC BY NC
10 aspects of highly effective research data :
www.elsevier.com/connect/10-aspects-of-highly-effective-research-data

Etape 1 : Données sauvegardées

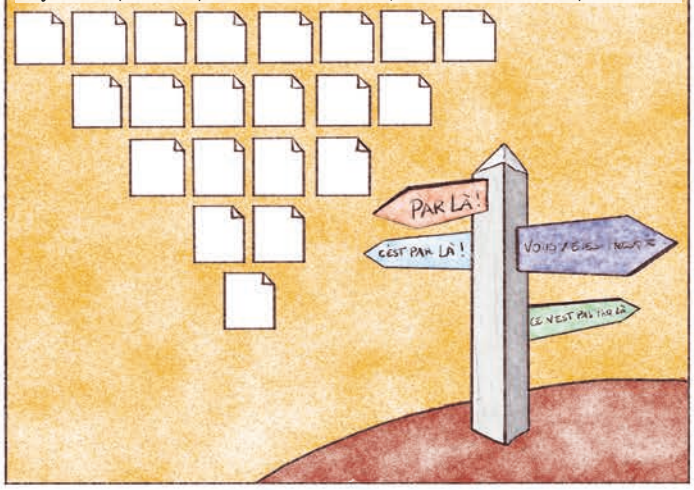
Les données doivent d'abord être **stockées**, idéalement selon la règle du 3-2-1 : 3 copies identiques, stockées sur 2 supports différents + 1 copie hors site



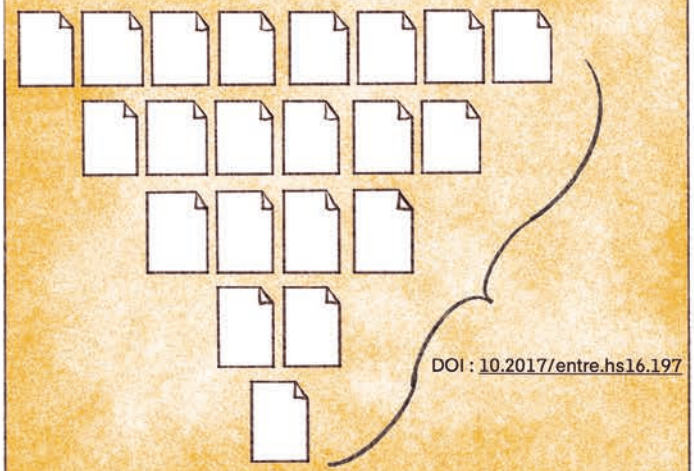
Mais cela ne suffit pas. Afin de prévenir l'obsolescence des technologies informatiques, les données doivent être **conservées**, c'est-à-dire sélectionnées sur différents critères et selon différentes durées de vie (conservation durant 6 mois, 3 ans, 10 ans, etc.)

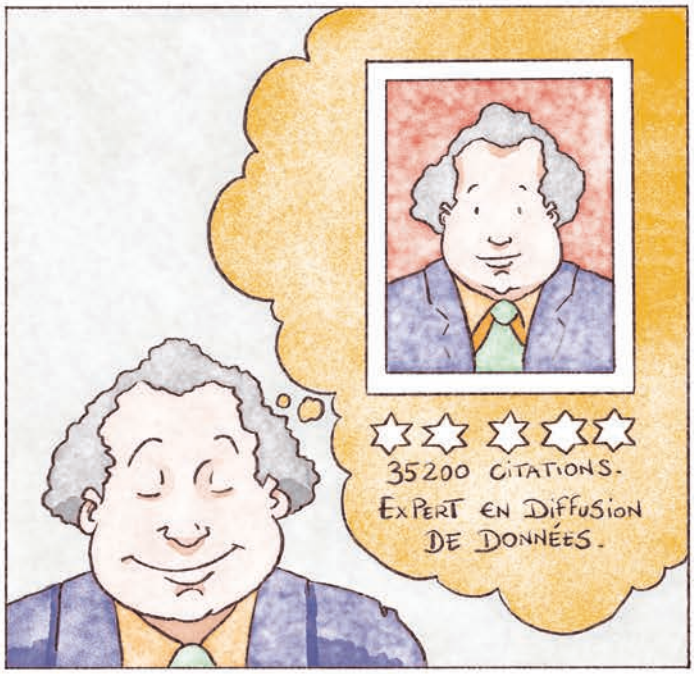
Même lorsque les données sont stockées et préservées, elles ne sont pour autant pas toujours disponibles pour les chercheurs et les machines qui voudraient les interroger. Il faut donc *a minima* les rendre **accessibles** en ligne.

Mais même en ligne, les données de la recherche ne sont pas toujours aisément **trouvables** par d'autres chercheurs. Il convient donc de les signaler – par exemple en améliorant la qualité de leur description.



Et afin de suivre la réutilisation de ces données et de s'assurer que les chercheurs qui les ont produites bénéficient du crédit scientifique qu'ils méritent, il est conseillé de les rendre **citables**, par exemple en leur attribuant un DOI (Digital Object Identifier) ou en les reliant à un *data-paper*.





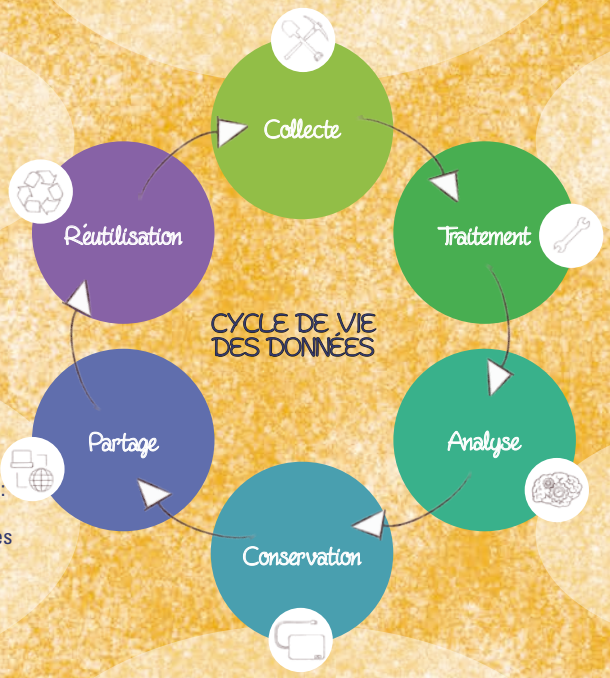
Des données qui ont été collectées pour un usage interne ne sont pas forcément **compréhensibles** par un tiers. Il convient de documenter leur collecte : quelles unités de mesure ont été utilisées ? Quel est le contexte ? Quelles abréviations et paramètres ont été employés ? Il est également nécessaire de les décrire le mieux possible – notamment grâce à des métadonnées complètes et précises.

Afin d'apporter une validation scientifique à ces données, il peut être utile de les soumettre à une **validation** par ses pairs. Il existe en effet également des systèmes de *revieuwing* pour les données de recherche (comme les *data papers*).

La **reproductibilité** des recherches permet d'accroître la crédibilité des résultats.

Compréhensibles, fiables et reproductibles, les données de recherche seront plus à même d'être **réutilisables** et citées par d'autres chercheurs. Il convient alors de leur appliquer une licence utilisateur qui encadrera strictement cette réutilisation par les autres chercheurs.

- Rédaction d'un Plan de gestion de données (PGD)
- Localisation des données existantes
- Obtention si besoin des autorisations/consentements (Respect du Règlement Général sur la Protection des Données (RGPD) pour les données personnelles, sensibles)
- Collecte / Acquisition des données
- Création des métadonnées de description



- Vérification, nettoyage et validation des données
- Description des données
- Gestion et stockage des données

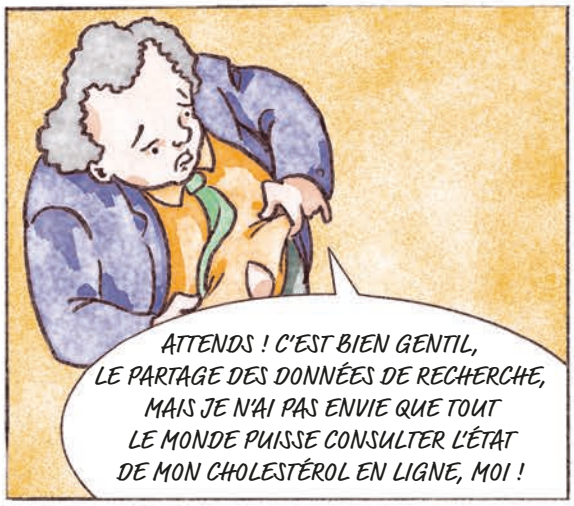
- Interprétation des données
- Production des résultats scientifiques
- Protection éventuelle par un brevet
- Publication des articles scientifiques

- Migration des données vers le meilleur format et le support adapté
- Augmentation des métadonnées et de la documentation
- Sauvegarde et stockage des données
- Dépôt des données dans un entrepôt
- Sélection et archivage des données à long terme

- Définition des droits d'accès des données comprenant : le niveau d'ouverture (de l'ouverture interne à l'ouverture publique)
- Les conditions d'utilisation des données : choix d'une licence de réutilisation
- Définition des droits d'accès des données comprenant : le niveau d'ouverture (de l'ouverture interne à l'ouverture publique, **éventuellement ouverture différée**)

- Nouvelle recherche
- Réexamen des résultats/des données
- Enseignement et apprentissage
- Respect de la licence de réutilisation

D'après l'institut Pasteur :
www.pasteur.fr/fr/ceris/bibliotheque/gerer-ses-donnees-recherche



QUESTIONS JURIDIQUES : DONNÉES OUVERTES OU NON OUVERTES ?

Données diffusables

- Communication libre si (cf. loi pour une République numérique, oct. 2016) :
  - données issues d'une activité de recherche financée au moins pour moitié par des fonds publics
  - non protégées par un droit spécifique
  - rendues publiques par le chercheur ou l'établissement (établissement décide quelles données seront ouvertes, le lieu et les conditions de dépôt)
- Communication obligatoire pour certaines données géographiques et environnementales (cf. convention Inspire et convention d'Arrhus)

Données diffusables sous conditions

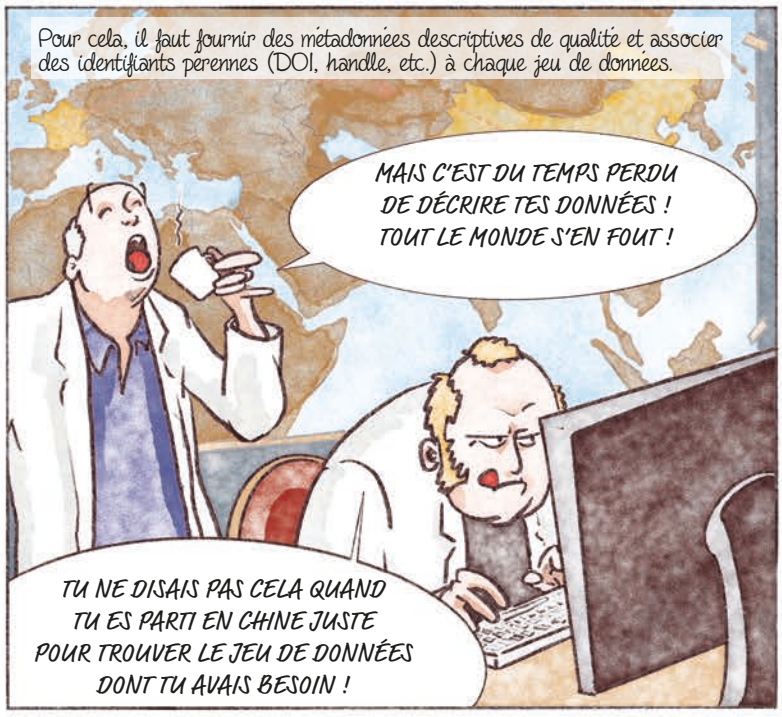
- Données présentant des risques pour la protection du potentiel scientifique et technique de la nation (cf. laboratoire dit « unité protégée »)
- Zones à régime restrictif (ZRR) : accès physique et numérique soumis à autorisation
- Données protégées par le droit d'auteur et autres droits de propriété intellectuelle
- Données personnelles (cf. Règlement général sur la protection des données (RGPD))

Données interdites à la diffusion

- Données présentant des risques pour la protection du secret de la défense nationale
- Données présentant des risques pour la sécurité de l'Etat, la sécurité publique, la sécurité de l'établissement
- Secret professionnel ou confidentialité (secret médical, secret de l'instruction, secret bancaire et fiscal)

Afin d'atteindre cet objectif, le partage des données doit respecter un ensemble de bonnes pratiques permettant leur découverte et leur utilisation par des humains – mais aussi par des machines. Cet engagement nécessaire est résumé par l'acronyme : **FAIR** ».

D'après Dominique L'Hostis, Du plan de gestion au data paper, juin 2019
https://gricad-media.univ-grenoble-alpes.fr/video/plan-gestion-donnees-au-data-paper



ACCESSIBLES : Les données et les métadonnées doivent être stockées durablement, avec des accès et/ou des téléchargements facilités, en spécifiant les conditions d'accès et d'utilisation.

Cela nécessite de rendre les données accessibles par leurs identifiants via un protocole de communication standardisé (ex : HTTPS, API REST). Il est conseillé d'utiliser au maximum des protocoles ouverts, libres, pouvant être implémentés de manière universelle.

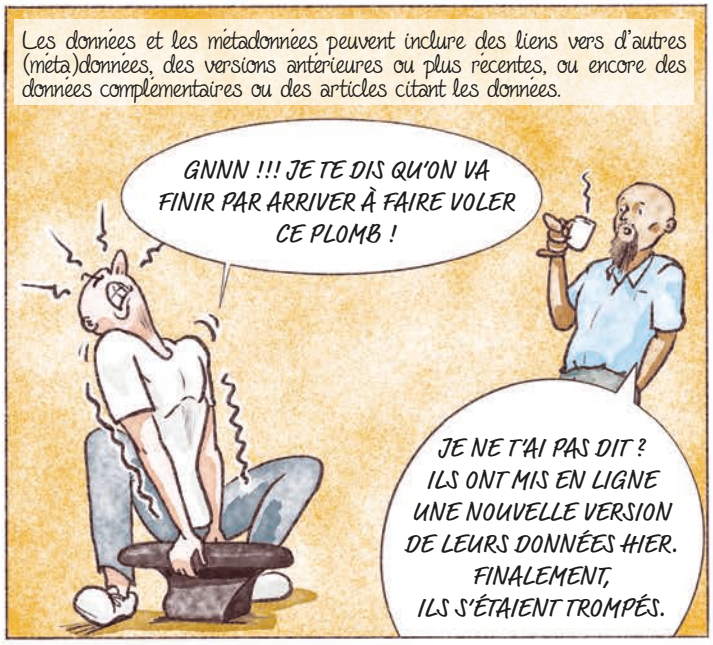
Cela suppose également de prévoir des protocoles permettant l'authentification et l'autorisation si besoin, par exemple pour limiter ou restreindre la consultation de données sensibles, stratégiques ou confidentielles à un certain type d'utilisateurs identifiés.

Enfin, il convient de rendre les métadonnées accessibles même quand les données ne le sont plus, ce qui suppose de mettre en place des protocoles d'archivage pérenne.

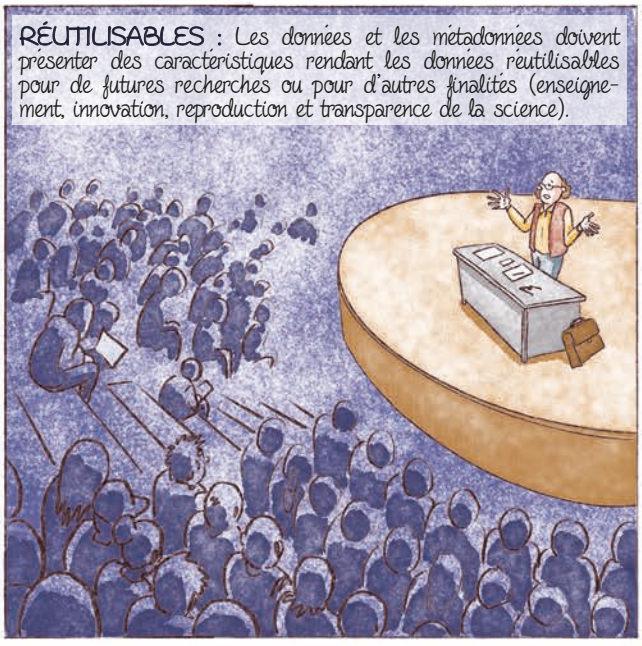
INTEROPÉRABLE : Les données et les métadonnées doivent être téléchargeables, utilisables, intelligibles et combinables avec d'autres données, par des humains ET par des machines

Il est souhaitable que les données et les métadonnées utilisent un langage formel, accessible et partagé largement applicable à la représentation des connaissances, comme les technologies du Web sémantique. Il est conseillé de s'appuyer sur des ontologies et des vocabulaires contrôlés standards.

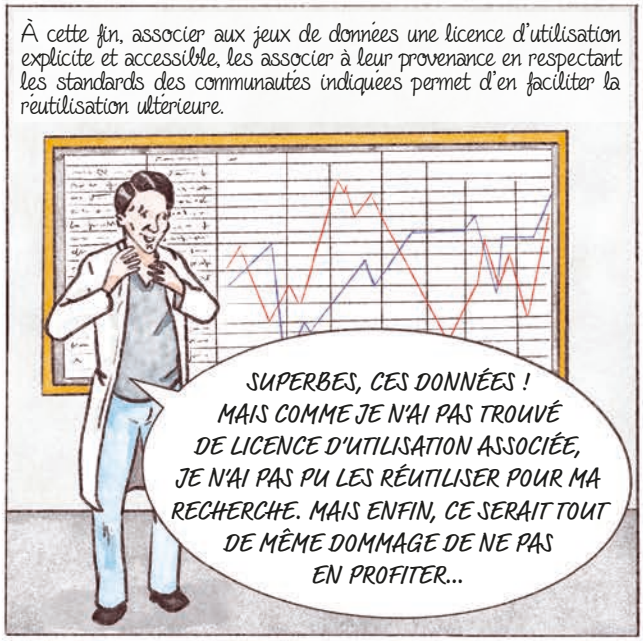
Les données et les métadonnées peuvent inclure des liens vers d'autres (méta)données, des versions antérieures ou plus récentes, ou encore des données complémentaires ou des articles citant les données.

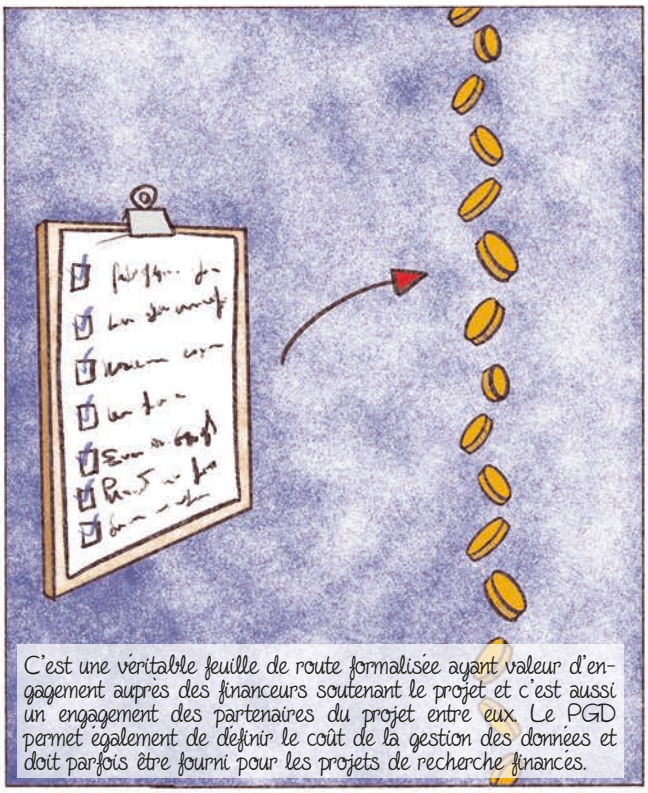


RÉUTILISABLES : Les données et les métadonnées doivent présenter des caractéristiques rendant les données réutilisables pour de futures recherches ou pour d'autres finalités (enseignement, innovation, reproduction et transparence de la science).



À cette fin, associer aux jeux de données une licence d'utilisation explicite et accessible, les associer à leur provenance en respectant les standards des communautés indiquées permet d'en faciliter la réutilisation ultérieure.





C'est une véritable feuille de route formalisée ayant valeur d'engagement auprès des financeurs soutenant le projet et c'est aussi un engagement des partenaires du projet entre eux. Le PGD permet également de définir le coût de la gestion des données et doit parfois être fourni pour les projets de recherche financés.

Le Plan de gestion de données (PGD) se construit par le biais d'une collaboration entre plusieurs acteurs professionnels aux compétences diverses et complémentaires.

L·E·A CHERCHEUR·SE

Le chercheur est responsable de la collecte, de la description et du découpage des données en différents jeux cohérents.

L'INGÉNIEUR·E PROJET

Quand un ingénieur-projet est nommé sur le projet, il devient responsable de la coordination des actions menées autour des données.

L'INFORMATICIEN·NE

L'informaticien est l'interlocuteur à privilégier pour les questions de stockage et de sécurisation des données, pour les aspects liés aux infrastructures et sur les coûts.

L·E·A BIBLIOTHÉCAIRE

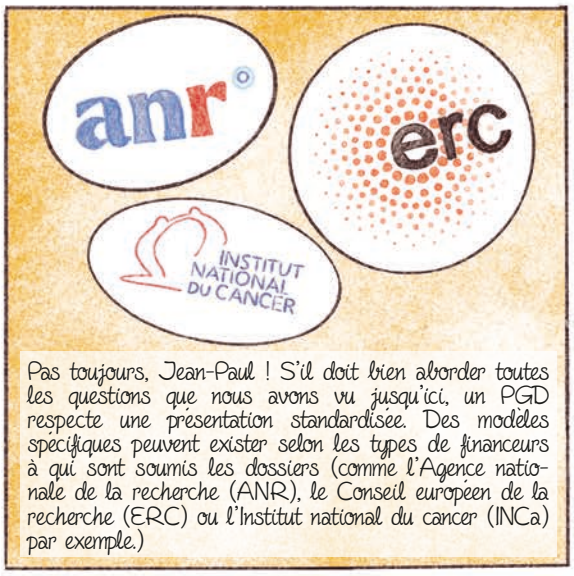
Le spécialiste de l'Information Scientifique et Technique – qu'il soit bibliothécaire, documentaliste ou archiviste – aide le chercheur à sélectionner les données, à définir les durées de conservation et les solutions techniques appropriées au partage des données.

L·E·A JURISTE

Le juriste et le DPD apportent leurs conseils sur les procédures d'anonymisation des données, sur les licences à utiliser, ou encore sur les permissions d'accès qui seront accordées aux différents groupes d'utilisateurs (accès ouvert, restreint ou sous embargo). Il conseille aussi sur la protection des données en vue de leur valorisation économique sous forme de brevets ou autres.

LES AUTRES PROFESSIONNEL·LE·S

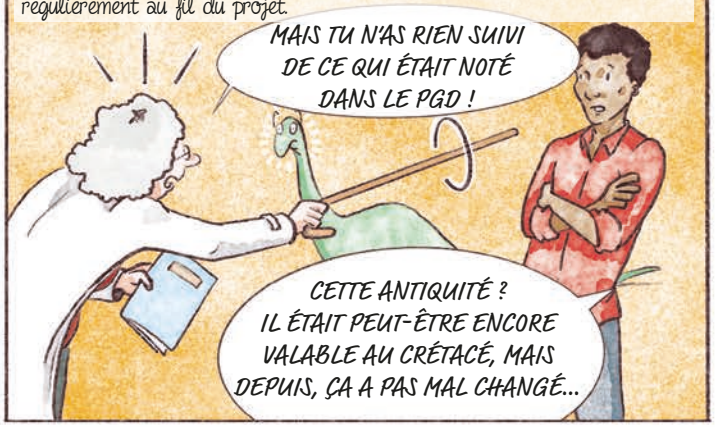
Cette liste de professionnels pouvant intervenir dans l'élaboration d'un plan de gestion de données (PGD) n'est bien-sûr pas exhaustive. Beaucoup d'autres compétences peuvent être convoquées.



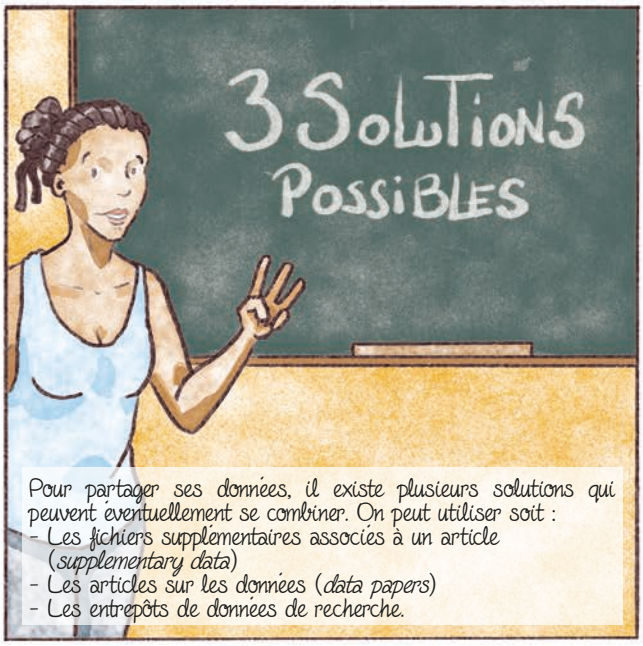
Pas toujours, Jean-Paul ! S'il doit bien aborder toutes les questions que nous avons vu jusqu'ici, un PGD respecte une présentation standardisée. Des modèles spécifiques peuvent exister selon les types de financeurs à qui sont soumis les dossiers (comme l'Agence nationale de la recherche (ANR), le Conseil européen de la recherche (ERC) ou l'Institut national du cancer (INCa) par exemple.)

Pour aider à sa rédaction, un outil est disponible en ligne : « DMP OPIDoR ». Ce dernier propose des modèles-types de PGD à remplir selon le financeur choisi. Il permet la collaboration à plusieurs, et offre aux équipes la possibilité de partager leur PGD une fois finalisé afin de mutualiser les expériences entre laboratoires.

Un Plan de gestion de données reste toutefois un objet vivant, qui se construit avant la collecte des premières données et qui est mis à jour régulièrement au fil du projet.



3. Partager les données de recherche : quels outils ?



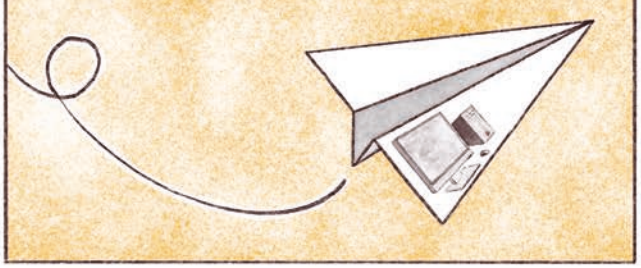
Pour partager ses données, il existe plusieurs solutions qui peuvent éventuellement se combiner. On peut utiliser soit :
- Les fichiers supplémentaires associés à un article (*supplementary data*)
- Les articles sur les données (*data papers*)
- Les entrepôts de données de recherche.

Une deuxième solution est de publier un article spécifique pour décrire un jeu de données : il s'agit d'un article de données ou *data paper*. Il pourra prendre place au sommaire d'une revue classique (si la ligne éditoriale de celle-ci le permet) ou dans une revue entièrement dédiée à ce type de publications : les *data-journals*.

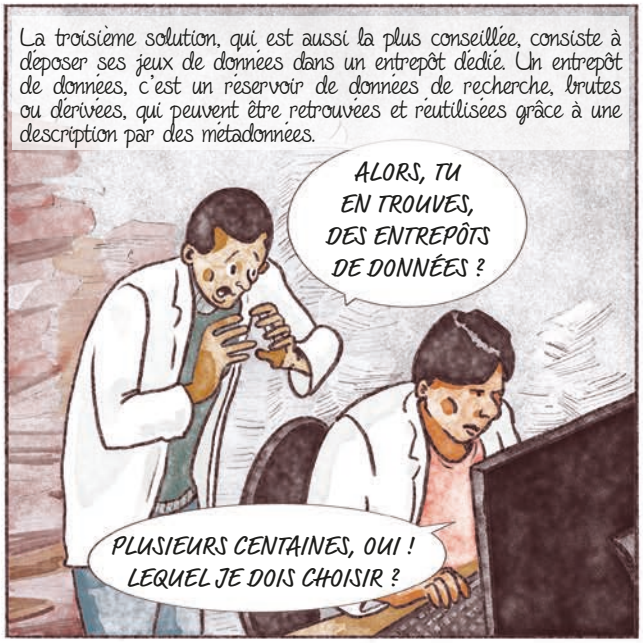
Contrairement à un article scientifique, le *data-paper* ne comprend ni hypothèse, ni conclusion, ni interprétations issues de l'analyse des données mais se contente d'analyses techniques et statistiques.

Classiquement, le *data-paper* comporte deux parties : un ensemble de fichiers (*data files*) accessibles directement sous forme de fichiers annexes ou via un entrepôt, et une partie descriptive qui donne le lien des jeux de données. Déposer les jeux de données dans un entrepôt est la solution à privilégier notamment car on peut leur attribuer des identifiants pérennes comme le DOI.

La partie descriptive d'un *data-paper* est très souvent composée de neuf parties :
- La page de titre
- L'introduction
- Une description suffisante des matériels et méthodes
- Une description suffisante des données
- Une partie d'informations et de discussions justifiant la rigueur des données
- Des conseils si besoin pour la réutilisation des données
- Les remerciements et la mention d'éventuels conflits d'intérêt
- La liste des références bibliographiques
- Les figures, tableaux et annexe relatifs à la méthodologie, à la qualité des données ou proposant une synthèse de celles-ci.

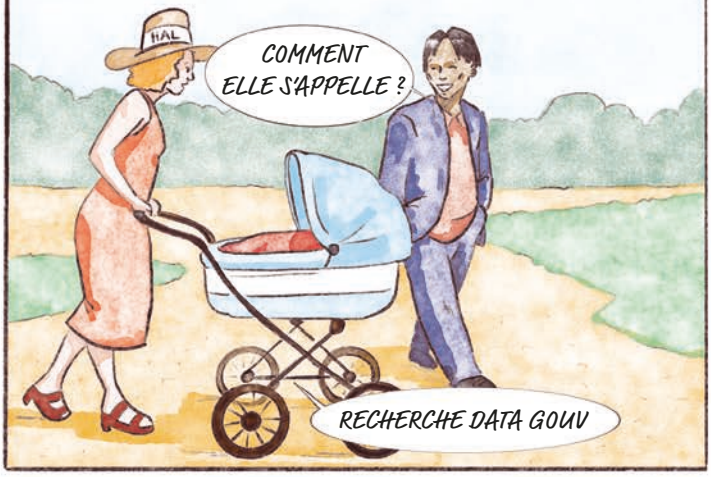


La troisième solution, qui est aussi la plus conseillée, consiste à déposer ses jeux de données dans un entrepôt dédié. Un entrepôt de données, c'est un réservoir de données de recherche, brutes ou dérivées, qui peuvent être retrouvées et réutilisées grâce à une description par des métadonnées.

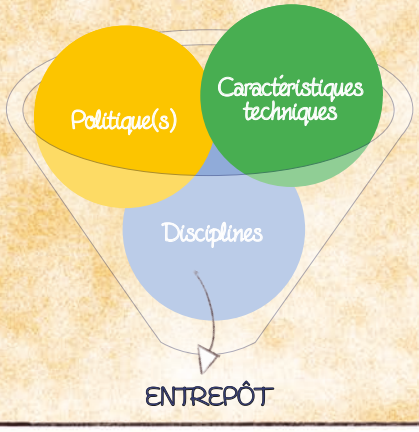


Eh oui ! Des entrepôts de données, dans le monde, il en existe plusieurs milliers. Ils se divisent en trois grandes catégories : les entrepôts disciplinaires, généralistes et institutionnels.

En France, il existe depuis fin 2021 un entrepôt national des données de la recherche, financé par le Ministère de l'enseignement supérieur et de la recherche, nommé Recherche Data Gouv : https://recherche.data.gouv.fr/fr



On l'a compris : le choix du bon entrepôt pour ses données est donc fondamental. Il obéit à trois critères principaux : les caractéristiques techniques de l'entrepôt, l'adéquation aux injonctions politiques des financeurs ou des éditeurs, et le critère disciplinaire.

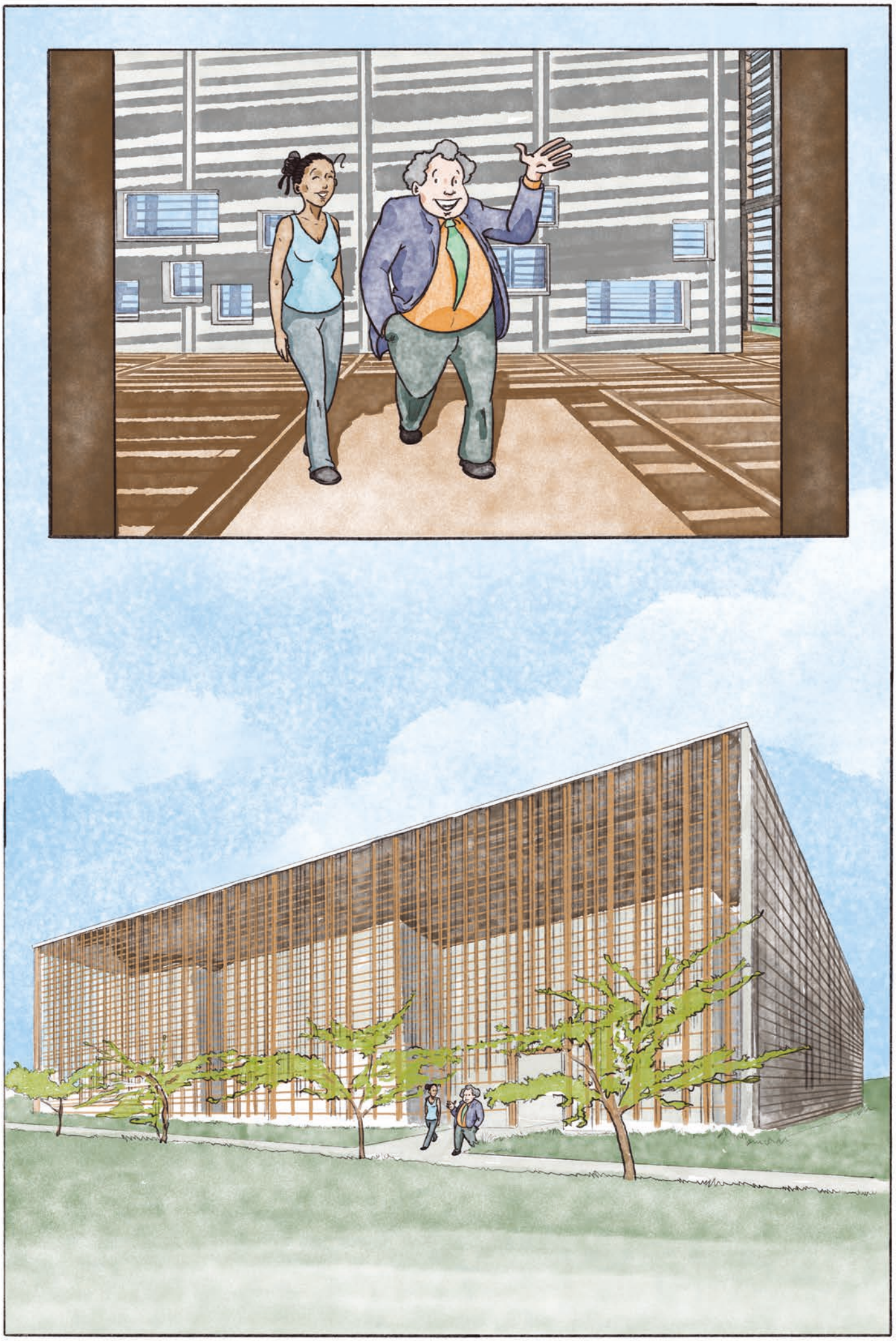


En ce qui concerne les caractéristiques techniques des entrepôts, un système de certification a été mis en place afin de permettre aux chercheurs de s'y retrouver. Ces certifications se basent sur l'évaluation de l'entrepôt par un jury indépendant sur une base déclarative. L'accès au dossier d'évaluation est rendu public afin de permettre un contrôle effectif par les utilisateurs

La certification permet notamment d'attester de la conformité de l'entrepôt avec les principes FAIR. Il en existe actuellement trois niveaux : le « CoreTrustSeal », s'obtient après avoir répondu favorablement à 16 critères (134 centres étaient certifiés dans le monde en 2023). Le Nestor Seal requiert, lui, un engagement sur 34 critères (5 centres certifiés dans le monde en 2023). Enfin, pour décrocher l'ISO + , il faut se conformer à une centaine de critères (un seul centre certifié dans le monde en 2023) !

Leur but est d'attirer le chercheur vers une suite de services de la marque qui l'accompagnera dans tous les stades de sa recherche : du logiciel de gestion de références bibliographiques qui permettra la constitution de sa bibliographie, à la base de données bibliographiques pour ses recherches documentaires, en passant par ses solutions de partage de données (entrepôts et *data-journals*) jusqu'aux outils de *datavisualisation* ou de *datamining*.

Le continuum ainsi proposé peut s'avérer très séduisant pour le chercheur du fait de sa praticité.



RÉFÉRENCES BIBLIOGRAPHIQUES :

Borgman, C. L. 2020. *Qu'est-ce que le travail scientifique des données ?* (traduit par C. Matoussowsky).
OpenEdition Press. 10.4000/books.oep.14692

Callisto Formation. Fondation UNIT.
https://callisto-formation.fr/?theme=boostplus_c06&redirect=0

CoopIST : délégation à la formation scientifique et technique, CIRAD.
Gérer les données de la recherche.
https://doi.org/10.18167/COOPIST/0005

DoRANum – Données de la recherche : Apprentissage Numérique.
https://doranum.fr/

Ouverture des données de recherche – Guide d'analyse
du cadre juridique en France – V2. (2017). Ouvrir la science !
Comité pour la science ouverte.
www. ouvrirlascience.fr/ouverture-des-donnees-de-recherche-guide-danalyse-du-cadrejuridique-en-france-v2

Partager les données liées aux publications scientifiques –
Guide pour les chercheurs. (2022). Ouvrir la Science !
Comité pour la science ouverte.
www. ouvrirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guidepour-les-chercheurs

AUTEUR
Marie Latour, directrice adjointe du
Service commun de la documentation de l'Université de Guyane

RESPONSABLE SCIENTIFIQUE
Annaïg Mahé, enseignante-chercheuse à l'URFIST de Paris

DESSINATEUR
Olivier Copin

GRAPHISME
Bénédicte Sauvage (BCOM)

RELECTEURS SCIENTIFIQUES
Cyril Heude (data librarian à SciencePo Paris*)*,
Romain Féret (directeur de Média Normandie),
Amélie Barrio (Co-responsable de l'URFIST Occitanie)

Projet co-financé par l'URFIST de Paris
(Ecole nationale des Chartes - PSL)