Initiation à la gestion des données de la recherche

Bonnes pratiques pour gérer, stocker, archiver, partager ses données tout au long de son projet de thèse

Cycl@doc – 2024

Objectifs

- Comprendre ce qu'est une donnée de la recherche (point de vue théorique)
- Savoir identifier dans son projet de thèse ce que sont les données de la recherche (point de vue pratique)
- Comprendre pourquoi un doctorant doit dès le début de sa thèse organiser, gérer, stocker, ses données de recherche pour lui même
- Puis les archiver, les diffuser, les réutiliser, les valoriser pour les autres
- Apprendre à structurer son travail : méthodes, outils, infrastructures et ressources à disposition, bonnes pratiques

Cette formation est inspirée de : Laetitia Bracco, Mathilde Barthe, Stéphanie Cheviron, Agnès Faller, Madeleine Hubert, Sylvie Steffann, … Jean-Baptiste Vu Van. (2020). Guide d'autoformation aux données de la recherche à destination des professionnels de l'information et de la documentation. http://doi.org/10.5281/zenodo.3920869

Programme

I. Pourquoi gérer les données de recherche de son projet de thèse et quelles données ?
I.1 Définitions et typologies
I.2 Pourquoi gérer ses données ?

II. Comment identifier et gérer les données de recherche de son projet de thèse : le cycle de vie de la donnée
II.1 Identifier les données de recherche dans sa thèse
II.2 Le cycle de vie des données
II.3 Les principes FAIR

III. Ressources et outils spécialisés
III.1 Ressources en autoformation
III.2 Les IR spécialisées en SHS
III.3 Les outils et services complémentaires

I. Pourquoi gérer ses données de recherche et quelles données ?

I.1 Définitions et typologies de la donnée de la recherche

- Définition OCDE
- Données collectées, données produites
- Données administratives, données de recherche, données sensibles et/ou personnelles, le RGPD.

I.2 Pourquoi gérer ses données ?

- Les politiques nationales et européennes de science ouverte
- Le décret du 3 décembre 2021 sur l'intégrité scientifique

I. Pourquoi gérer ses données de recherche et quelles données ?

I.1 Définitions et typologies de la donnée de la recherche

> « Les données de recherche sont depuis toujours le fondement de toute production scientifique. »

A partir de quand une information ou un matériau collecté sont-ils considérés comme une/des données de la recherche ?

Certains identifient informations collectées, veille et données :

> « *Moi j'utilise Zotero, mais est-ce que ça fait partie des informations de la recherche, sachant que c'est plutôt des pdf, des textes, enfin si ce n'est pas des corpus que je constitue ?»*

La donnée n'est pas figée dans le temps ou dans des états successifs bien identifiés.

> « C'est difficile de savoir où commence le traitement de la donnée collectée ou de la donnée brute. »

Source des citations : Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet. Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2 : Annexe 3 : Extraits des entretiens, [Rapport de recherche] Université Rennes 2. 2017, 26 p. ⟨hal-01635186v2⟩

Définition de la donnée de la recherche

« Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images, sons, etc.), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche.»

(OCDE, 2007, rapport « Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics »)

Autres définitions de la donnée de la recherche

« Une observation, un objet, un document ou toute autre entité devient une donnée de recherche, dès lors qu'elle est utilisée comme preuve d'un phénomène, c'est-à-dire qu'elle est collectée, analysée et interprétée »

«Les données de la recherche sont l'ensemble des informations et matériaux produits et reçus par des équipes de recherche et des chercheurs. Elles sont collectées et documentées à des fins de recherche scientifique. A ce titre, elles constituent une partie des archives de la recherche.»

SECTION AURORE

association des archivistes français

Source : Big data, little data, no data : scholarship in the networked world / Christine L. Borgman, cop. 2015

(Les deux seules) questions à se poser pour identifier ses données

- Quels sont les éléments, numériques ou non, auxquels je tiens vraiment et qui seraient irremplaçables ou très longs à remplacer en cas de perte, de vol ou de problème technique ?
- Si je devais relire et évaluer les travaux de collègues qui travaillent sur un sujet de recherche similaire au mien, de quoi aurais-je besoin pour vérifier leurs résultats ?

URFIST École nationale des

Source de cette diapositive : Introduction aux données de la recherche, Justine Ancelin-Fabre, 21 mai 2021, URFIST de Paris.

Tour de table :

Quels matériaux de recherche/données de recherche identifiez-vous dans votre thèse ?

https://app.wooclap.com/NKONNZ?from=event-page

Restitution collective :

- **Exemples pratiques**

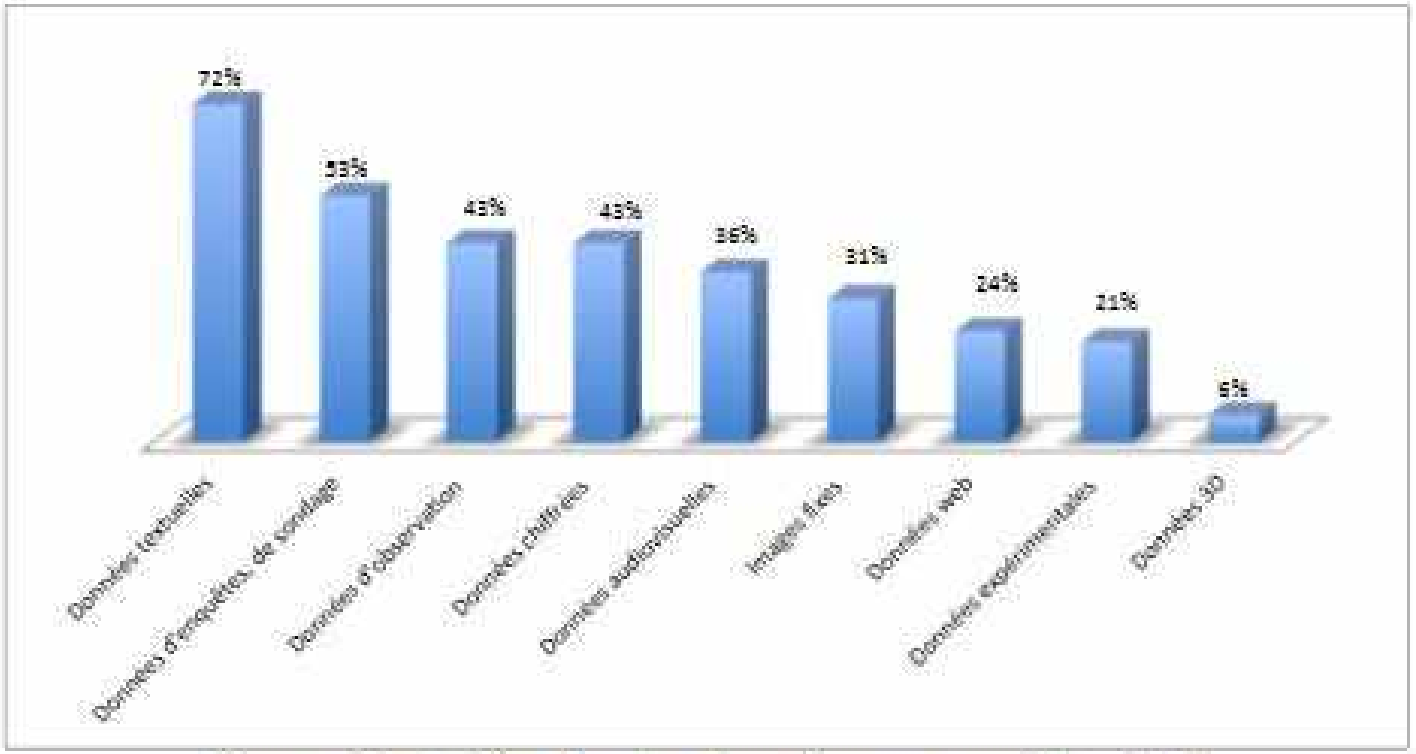


Figure 13 : Catégories des données sources (N = 127)

Source du graphique : Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet. Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2 : Rapport ; Annexe 1 : Résultats de l'enquête statistique ; Annexe 2 : Croisements statistiques ; [Rapport de recherche] Université Rennes 2. 2017, 159 p., 47 p., 114 p., 26 p., 23 p. ⟨hal-01635186v2⟩

Typologies de données 1/2

Il existe différents types de données de la recherche qui diffèrent selon la manière dont les données sont produites et selon leur valeur supposée.

- Données d'observation

- capturées en temps réel ;
- habituellement uniques et donc impossibles à reproduire ;

Ex. : neuro-imagerie, photographies astronomique, données d'enquêtes

- Données expérimentales

- obtenues à partir d'équipements de laboratoire ;
- souvent reproductibles mais parfois coûteuses ;

Ex.: chromatogrammes, puces à ADN

- Données computationnelles ou de simulation

- générées par des modèles informatiques ou de simulation ;
- souvent reproductibles si le modèle est correctement documenté ;

Ex. : modèles météorologiques, modèles de simulations sismiques, modèles économiques

Typologies de données 2/2

- Données dérivées ou compilées

- issues du traitement ou de la combinaison de données "brutes" ;
- souvent reproductibles mais coûteuses ;

Ex. : fouille de texte, bases de données compilées

- Données de référence

- collection ou accumulation de petits jeux de données qui ont été revus par les pairs, annotés et mis à disposition ;

Ex. : GenBank, base de données de cristallographie, collection de lettres ou archive d'images historiques

***

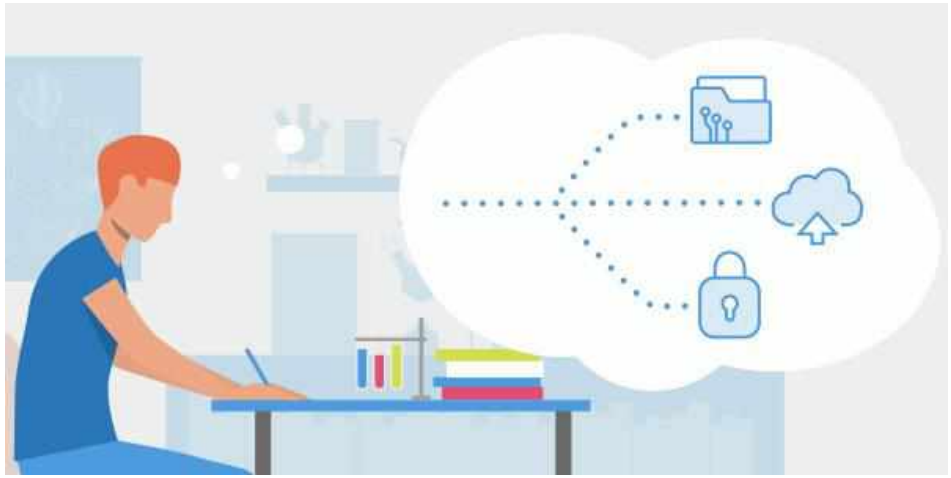
Les données de la recherche ne concernent pas les :

- Les données administratives ;
- Les publications ;
- Les supports de cours ;
- Les carnets de laboratoire ;
- Les objets matériels (échantillons de laboratoire, souches bactériennes, animaux de laboratoire...) ;
- Les communications personnelles avec des collègues.

❖ Distinguer les typologies de données

- La donnée brute
- La donnée primaire
- La donnée source
- La donnée collectée
- La donnée produite

➢ C’est à dire ?



La Donnée brute / primaire / source / collectée

- du matériau langagier [extrait d'] un terrain ethnographique [investi] sur la durée, sur la profondeur, avec une compréhension très contextualisée de la situation ;
- des conversations ;
- du témoignage ethnographique ; des entretiens ; des biographies langagières orales ou écrites ;
- des **documents d'archive** ; des documents audiovisuels ; des émissions radiophoniques ; des **corpus de presse** ; des programmes, les discours des inspecteurs généraux, des manuels ; des **corpus de dossiers** d'aide à l'enfance ; **des corpus de textes de loi** ;
- des **échanges avec les artistes** ; des interviews avec des collègues ; des relevés GPS ; des données d'enquête ;
- **des images ; des textes ; des corpus de textes** ; des **corpus visuels** ; des adresses ; des ouvrages ;
- **des articles** ; des facsimilés d'ouvrages ; **des documents** ; des brochures ; **des photos** ; des fragments de pierres, des inscriptions ; **de la donnée géographique numérique ;**

Patchwork de données brutes 1/2

« … **des cartes papier ou numériques** ; tout ce qui est image**, photographie aérienne**, image satellite ; **des information informelles glanées lors de colloques** ; des données récupérées sur Internet ; des SMS ; des enregistrements de webcams ; **des vidéos** (youtube ou autres plateformes) ; des pages de forums Internet, de discussion, sites web, ; des pages de tchat sos ; des pages ou échanges de réseaux sociaux, comptes twitter, Facebook, ; **des PdF d'œuvres sur Internet** ; des index et listes ; des photos aériennes. ; des données numériques, statistiques ; des cartes ; **des données archéologiques** ; **des captures d'écran** ; des traces GPS ; des trajectoires de marche ; des indicateurs liés à de petites tâches cognitives ; des données territorialisées...

▪ Encore plus de données brutes 2/2

… des **catalogues prosopographiques** ; des arbres généalogiques ; de recensements d'archives ; **des campagnes de photos** ; des fiches thématiques ; des bases de données géographiques ; des traces GPS ; des indicateurs ; des données biologiques, des dosages ; des électrocardiogrammes ; des consentements de participation (papier) ; des protocoles ; des fichiers Latex ; des fichiers Matlab ; des fichiers partagés ; **des bibliographies** ; des fichiers statistiques ; des nuages de points, des captures de mouvement ; **des croquis** ; des enregistrements sonores suivis sur plusieurs années ; des photographies suivies sur plusieurs années ; des corpus multilingues ; des carnets papier... »

Source : Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet. Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs: une enquête à l'Université Rennes 2: Rapport; Annexe 1: Résultats de l'enquête statistique; Annexe 2: Croisements statistiques; Annexe 3: Extraits des entretiens; Synthèse des résultats. [Rapport de recherche] Université Rennes 2. 2017 ⟨hal-01635186v2⟩

Un témoignage sur les données collectées

« *Ma collecte de données, c'est justement ce dont je parlais [...] c'est à dire* ***tous les échanges avec les artistes, mais qui ne sont pas méthodiques,*** *ce n'est pas des formes d'enquêtes, des formes d'interviews, c'est divers types d'échanges à diverses occasions qui peuvent* ***avoir l'air de rien du tout*** *mais qui sont* ***des données sur lesquelles je m'appuie réellement*** *».*

Source de la citation : Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet. Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2 : Annexe 3 : Extraits des entretiens, [Rapport de recherche] Université Rennes 2. 2017, 26 p. ⟨hal-01635186v2⟩

- **Les données produites**

C’est le résultat de recherche nécessaire pour valider votre hypothèse et qui sera décrit (jeu de données), archivé dans un entrepôt, diffusé autant que faire se peut.

Les données produites ont fait l’objet d’un traitement, d’une organisation, d’une mise en forme.

Exemples : bases de données ; corpus de textes comme un corpus de jurisprudence ; corpus d’images.

Les corpus ont été regroupés dans une optique précise et matérialisent l’obtention d’un résultat.

Les données produites permettent au jury de votre thèse ou à la communauté scientifique de valider vos conclusions finales et dans certains cas de reproduire votre démarche de recherche.

Patchwork de données produites

« …des données textuelles ; **des bases de données** ; des enregistrements audio ou vidéo d'entretiens ; des notes sur des entretiens ; des questionnaires ; des formulaires papier remplis ; des tableaux Excel ; **des fichiers Word** ; plusieurs versions de textes ; **des fiches à partir de documents d'archive** ; des sites ou plateformes internet ; des retranscriptions d'entretiens ; des analyses statistiques lexicales ; **des dossiers PDF**, **des textes** ; des livres ; des revues ; des données excel ; des données chiffrées ; des données qualitatives ; des études de cas cliniques ; des logiciels ; des tâches ; des petits programmes ; de petites épreuves cognitives ; des tests de raisonnement ; **des fichiers texte avec reconnaissance de caractères, des transcriptions** ; des documents numérisés ; **des photographies de documents originaux** ; des enregistrements vidéo ; des rushs ; des documentaires… »

- **De la donnée aux jeux de données**

Les données produites de votre projet de thèse constituent des ensembles finis, organisés qui peuvent être, pour certains ensembles, décrits avec des métadonnées, selon des normes afin de constituer des jeux de données.

Un jeu de données peut être ensuite diffusé.

« Les métadonnées sont la carte d'identité des données. Elles permettent de les identifier, les décrire, expliquer l'origine de leur création, leur utilité et leurs destinataires. »
https://www.enssib.fr/le-dictionnaire/metadonnees

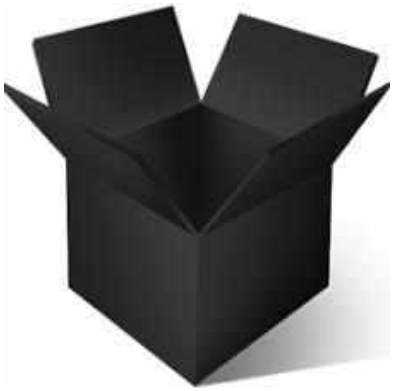
- **De la donnée aux jeux de données**

Définition de l'IRD

« Un dataset (**jeu de données ou ensemble de données en français**) est un **ensemble cohérent de données** produites dans le cadre d'un même projet, sur un **même objet d'étude** et/ou recueillies sur un même lieu. Toutes les données d'un dataset peuvent donc être décrites avec une majorité de **métadonnées** communes.

Tous les types de fichiers sont admis (**tabulaire, texte, pdf, image, vidéo, audio, SHP, etc.**), mais on choisira de préférence des formats ouverts et standards pour faciliter la réutilisation. »

Un chercheur en SHS évoque « la boîte noire » de la recherche



« Notre métier n'est pas de créer des bases de données pour créer des bases de données, c'est de publier donc ***on produit les résultats, pas la boîte noire*** *; donc* ***rendre publique la boîte noire, c'est pas une habitude****, on n'est pas valorisé scientifiquement pour avoir passé du temps à publier nos boîtes noires. »*

Source de la citation : Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet. Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2 : Annexe 3 : Extraits des entretiens, [Rapport de recherche] Université Rennes 2. 2017, 26 p. ⟨hal-01635186v2⟩

❖ Données administratives / Données de recherche / personnelles ou sensibles / RGPD

Dans votre projet de thèse, vous maniez des données administratives qui peuvent être des données personnelles (parfois sensibles) et des données de la recherche qui peuvent avoir ou non un caractère personnel.

Exemples :
- L'organisation d'un colloque génère des données administratives à caractère personnel.
- Mener des entretiens génère des données de recherche à caractère personnel.

Donnée à caractère personnel :
« … toute information [ou ensemble d'informations] relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement … »
Y compris ré-identification ultérieure

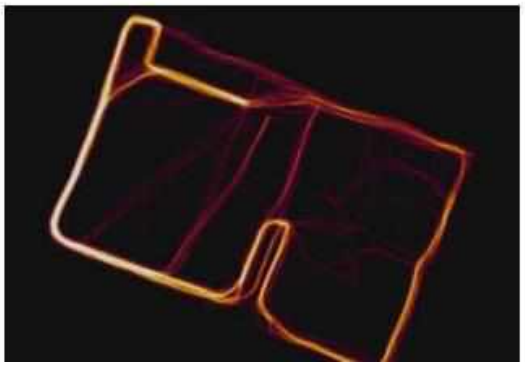
Traitement :
« … toute opération ou tout ensemble d'opérations portant sur de telles données, quel que soit le procédé utilisé… »

- **Les données à caractère personnel :**

De loi du 6 janvier 1978, dite « Informatique et Libertés « au Règlement Général sur la Protection des Données – (RGPD)

1978

2018



- **Le RGPD**

Au niveau européen :

- ✓ RGPD = Règlement Général sur la Protection des Données
- ✓ Règlement, pas directive, applicable sans transposition
- ✓ Obligatoire depuis le 25 mai 2018

Au niveau français :

- ✓ Loi Informatique et Libertés de 1978, version du 06/08/2018
- ✓ Décrets d'applications
- ✓ CNIL : Commission Nationale Informatique et Libertés
- ✓ Au niveau européen : le G29, devenu le CEPD

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Les données sensibles sont partout

Les données sensibles forment une catégorie particulière des données personnelles.

Ce sont des informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique.

Le règlement européen interdit de recueillir ou d'utiliser ces données, sauf, notamment, dans les cas suivants :

- si la personne concernée a donné son consentement exprès (démarche active, explicite et de préférence écrite, qui doit être libre, spécifique, et informée) ;
- si les informations sont manifestement rendues publiques par la personne concernée ;
- si elles sont nécessaires à la sauvegarde de la vie humaine ;
- si leur utilisation est justifiée par l'intérêt public et autorisé par la CNIL ;
- si elles concernent les membres ou adhérents d'une association ou d'une organisation politique, religieuse, philosophique, politique ou syndicale.

Pseudonymiser, anonymiser

- Inscription au registre de traitement
- Sélectionner les données personnelles à conserver et diffuser
- Minimiser les données : conserver et ne diffuser que les données
  - ✓ Essentielles
  - ✓ Pertinentes
  - ✓ Absolument nécessaires
  - ✓ Qui ont une utilité scientifique
- Pseudonymisation / anonymisation

➢ Consulter le DPD pour valider la sélection des données et leur anonymisation : **François Descubes, délégué à la protection des données (DPD) à Paris 1 -** dpo@univ-paris1.fr

I. 2. Pourquoi gérer ses données ?

- Pour vous-même

- Pour amorcer dès le doctorat une démarche de science ouverte

- Pour se conformer aux politiques nationales et européennes en faveur de la science ouverte

- Et au décret du 3 décembre 2021 sur l'intégrité scientifique

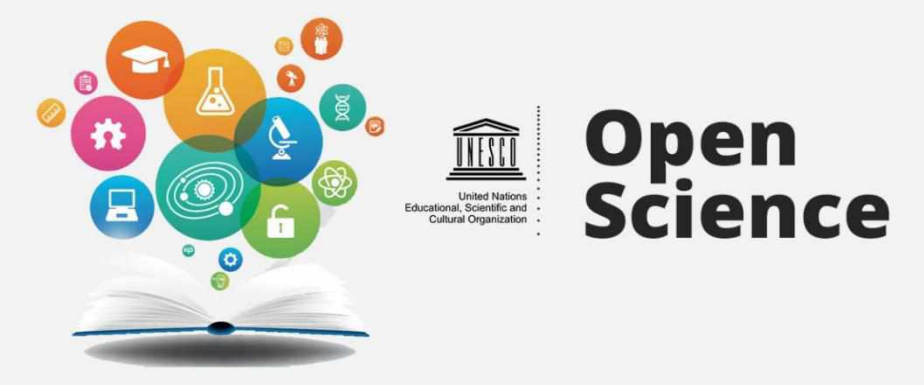
Le principe « aussi ouvert que possible, aussi fermé que nécessaire » est au cœur de la démarche.

« La science ouverte est la **diffusion sans entrave** des résultats, des méthodes et des produits de la recherche scientifique. Elle s'appuie sur l'opportunité que représente la mutation numérique pour développer l'accès ouvert aux publications et – autant que possible – aux données, aux codes sources et aux méthodes de la recherche.»

« Ouvrir la boîte noire du chercheur en partageant autant que possible les ***données et les méthodes sous-jacentes*** *aux publications. »*
(in Passeport pour la science ouverte.)

Les principes au cœur de la science ouverte :

- ✓ *La reproductibilité de la recherche*
- ✓ *Une meilleure efficacité de la recherche par **le partage** des données et des méthodes*
- ✓ *La **réutilisation** des données produites (plus d'efficacité, moins de redondance, accélération de la production de nouveaux savoirs)*
- ✓ *L'intégrité scientifique, la transparence*



❖ Les politiques nationales et européennes de science ouverte

❖ Décret relatif au respect des exigences de l'intégrité scientifique

Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique par les établissements publics contribuant au service public de la recherche et les fondations reconnues d'utilité publique ayant pour activité principale la recherche publique

NOR : ESRR2133294D
Accéder à la version consolidée
ELI : https://www.legifrance.gouv.fr/eli/decret/2021/12/3/ESRR2133294D/jo/texte
Alias : https://www.legifrance.gouv.fr/eli/decret/2021/12/3/2021-1572/jo/texte
JORF n°0283 du 5 décembre 2021
Texte n° 63

Article 2, alinéa 3.

Les établissements publics et fondations reconnues d'utilité publique mentionnés au troisième alinéa de l'article L. 211-2 du code de la recherche :

« Promeuvent la diffusion des publications en accès ouvert et la **mise à disposition des méthodes et protocoles, des données et des codes sources associés aux résultats de la recherche** afin d'en garantir la traçabilité et la reproductibilité. Ils incitent à la publication des résultats de recherche dits négatifs. »
https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044411360

❖ Ethique, déontologie et intégrité scientifique à Paris 1

« La promotion des valeurs et le respect des normes de l'intégrité scientifique est désormais une obligation pour les universités et pour tout opérateur de la recherche publique (art. L211-2 du Code de la recherche).
Il est de la responsabilité de l'université et de ses membres de veiller non seulement à éviter et à décourager des **conduites malhonnêtes** comme diverses formes de **pillage** du travail d'autrui, mais aussi à **développer une bonne recherche** selon les normes et les méthodes de la spécialité concernée.
Pour les institutions elles-mêmes – universités et organismes –, il s'agit d'une obligation de moyens. La formation et l'information sur le sujet en font partie. »

https://recherche.pantheonsorbonne.fr/politique-scientifique/ethique-deontologie-et-integrite-scientifique

Les référents sont les points de contact principaux pour les interrogations de chacun concernant l'éthique, la déontologie des agents publics et l'intégrité scientifique au sein de l'université.
https://www.pantheonsorbonne.fr/universite/referents-et-comite-ethique

EPI **Qu'est-ce que l'intégrité scientifique ?**

II. Comment identifier et gérer les données de recherche de son projet de thèse et quelles données ?

II.1 Identifier ses données de recherche
- Exercice en groupes

II.2 Le cycle de vie des données
- Les étapes du cycle de vie des données
- Approfondir les étapes du cycle de vie des données
- Les principes FAIR

Exercice en groupes :

A vous la parole …

Exercice en groupes :

A partir des données de recherche identifiées dans votre thèse, quelles sont les traitements et les démarches que vous devrez effectuer selon leur nature ?

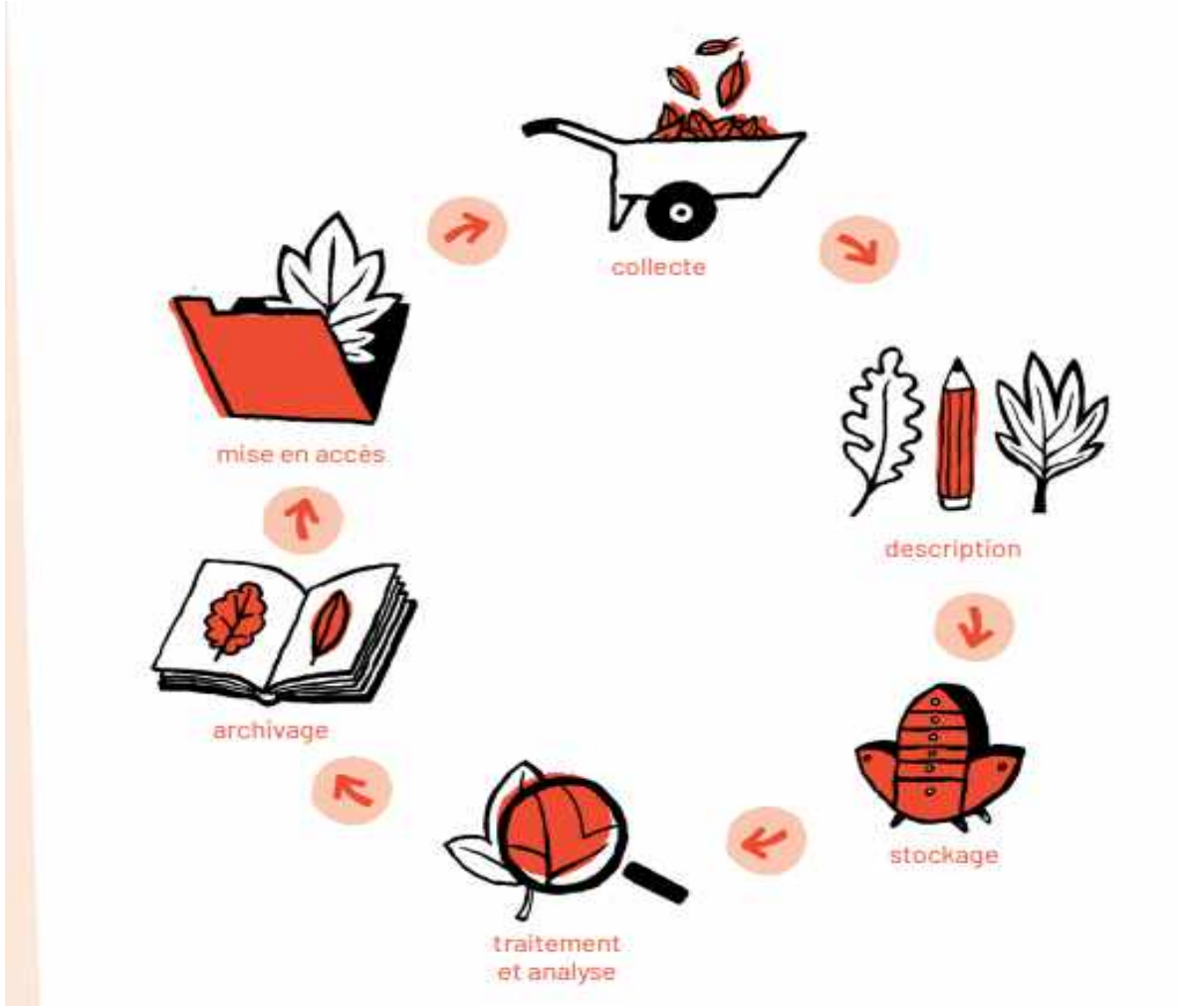
Mes données collectées / Mes données produites :	Pour quels traitements ? Pour quelles démarches ?

La gestion rigoureuse et responsable des données doit donc être un élément indispensable de tout projet de recherche, pensée dès sa conception, suivie jusqu'à son terme, anticipant le devenir des données après ce terme.

Source du graphique : https://uqo.ca/biblio/gestion-donnees-la-recherche

PAUSE de 10’

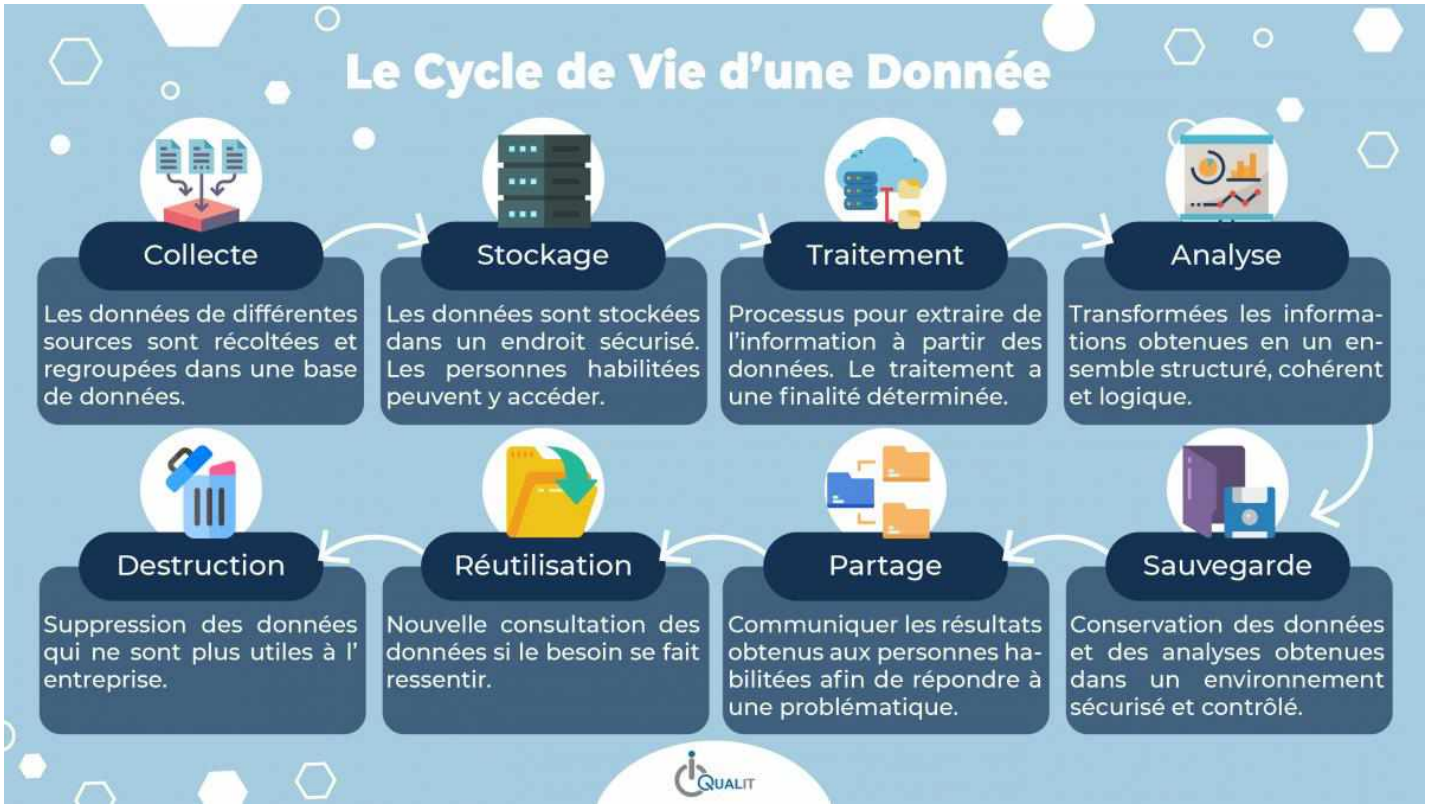
II.2 Le cycle de vie des données



Source : Passeport pour la science ouverte.

❖ Les étapes du cycle de la donnée

- Creating data (créer ou collecter)
- Processing data (traiter)
- Analysing data (analyser)
- Preserving data (conserver)
- Giving access to data (donner accès)
- Reusing data (réutiliser)



Source du graphique : https://www.iqualit.com/article/quel-est-le-cycle-de-vie-de-vos-donnees-electroniques/

❖ Approfondir le cycle de vie des données

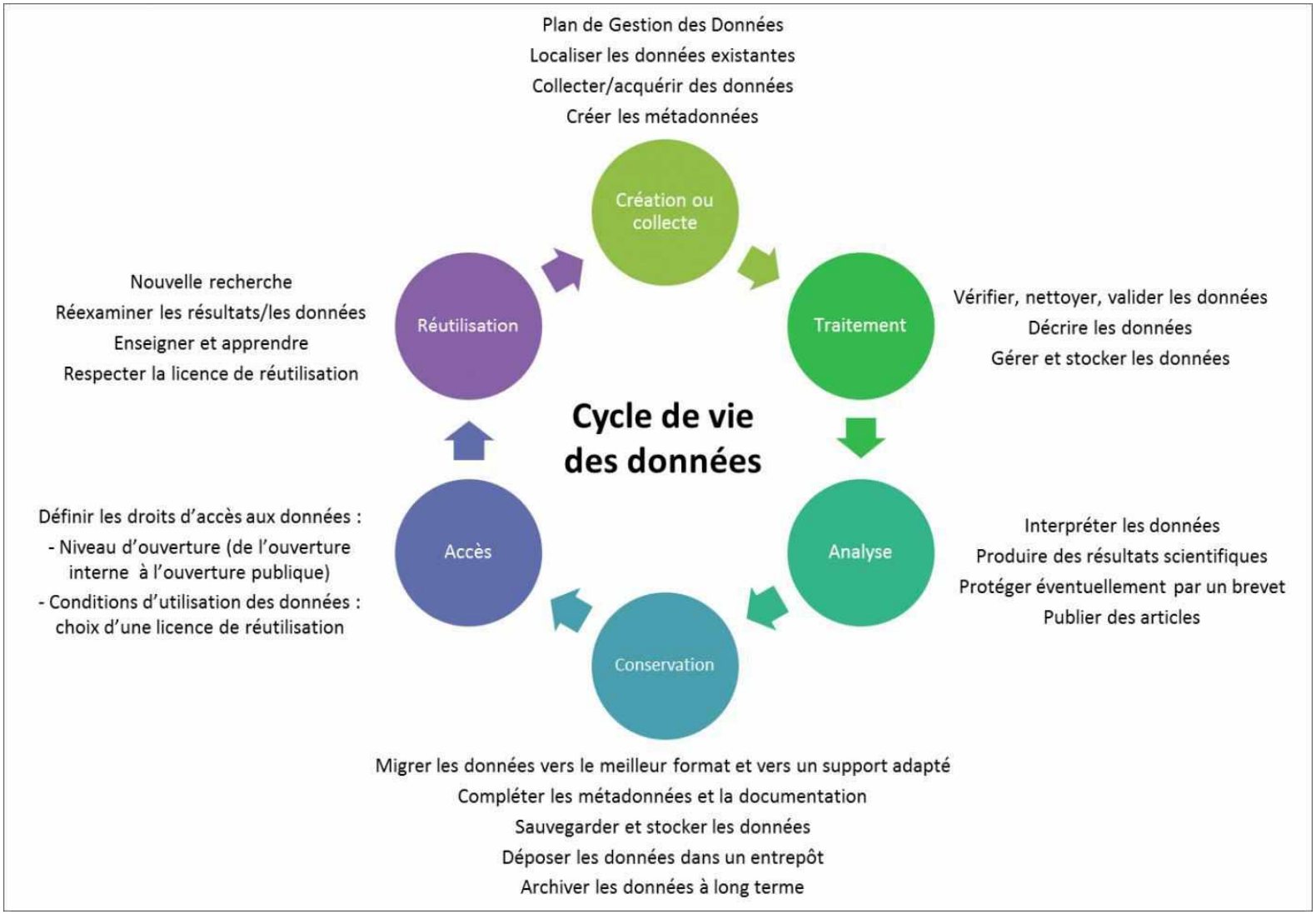


Image adaptée à partir du cycle de vie des données de UK Data Archive : https://www.ukdataservice.ac.uk/manage-data/lifecycle

« ***20 ans après publication, 80 % des données scientifiques sont perdues****, pour leurs auteurs, pour leurs équipes et leurs instituts,* ***pour la recherche mondiale… »***

La gestion des données de recherche (ou RDM : research data management) consiste à évaluer, **protéger**, documenter, **sauvegarder et partager** les données générées par l'activité scientifique durant toutes les étapes du cycle de vie de ces données. »

Source : site support IRD Data consulté le 24 mars 2021.
https://data.ird.fr/gerer/

- **Creating Data / Créer ou collecter**

Trouver et réutiliser des données existantes pour son projet de recherche

Quelques exemples de fournisseurs en open data :

International
The home of the U.S. Government's open data - catalog.data.gov/
Open Knowledge Foundation - dataportals.org
DataCite
opendatainception.io/ - OpenDataSoft
donnees.banquemondiale.org/ - La Banque mondiale
Google Dataset Search

France - Plateforme ouverte des données publiques françaises
data.gouv.fr Datainfogreffe.fr Data.culture.gouv.fr
transport.data.gouv.fr Datatourisme.gouv.fr Cadastre.data.gouv.fr
http://www.progedo.fr/ etc.

Un mot sur
Citer un jeu de données : une référence comme les autres ?

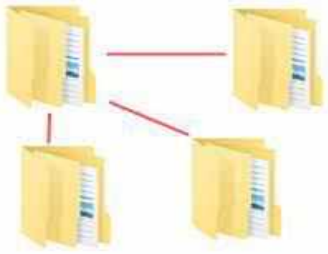
Auteur / Créateur
Date de publication / Mise en ligne
Titre
Edition (révision)
Version (toujours croissante et numérique)
Nom de la norme employée
Type de ressource
Editeur / Producteur
Identifiant (DOI)
Localisation (url)

Exemples :
P. Relvas, F. Ferreira, H. Sobreira, C. Costa et al. Robot navigation measurements driving in and out of the van [en ligne]. Dataset. Disponible sur : https://zenodo.org/record/2554950

DataCite Metadata Schema for the Publication and Citation of Research Data (2016)
https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf

Processing Data / Traiter

Un mot sur



- Définir un champ sémantique

=> Obtenir un aperçu d'un sujet de recherche avec Open Knowledge Maps
https://openknowledgemaps.org/

- Organiser les fichiers, nommer, sauvegarder

https://archives.pantheonsorbonne.fr/sites/default/files/2021-01/1_nommage.pdf

- Description des données

La description des données est une étape primordiale. Afin que les données de la recherche soient réutilisables, le contexte de leur production doit être documenté de manière précise et intelligible.

Ainsi, il peut être décrit par :

- une documentation adéquate, sous la forme d'un fichier txt ou pdf qui rapporte des informations sur le projet (hypothèses, méthodologie, échantillonnage, instruments ...), sur les fichiers ou la base de données et sur les paramètres ;
- et des métadonnées (Metadata) détaillées dans la section suivante.

- Construire un thésaurus

  - Collecte du vocabulaire à partir de référentiels existants et/ou de pratiques d'indexation existantes (approche synthétique et analytique) et identification des concepts et termes candidats.
  - Ventilation par champs sémantiques, domaines ou micro-thésaurus. / ventilation éventuelle en parallèle par facettes.
  - Choix des concepts et des termes retenus anciennement descripteurs et création des relations d'équivalence vers les termes non retenus, anciennement non-descripteurs en distinguant les types de termes équivalents.
  - Création des notes d'application et typage de ces notes pour les concepts et/ou pour les termes préférentiels.
  - Création des relations hiérarchiques entre concepts à l'intérieur d'un domaine, et éventuellement de poly-hiérarchie. Prendre en compte les différents types de relations génériques/spécifiques.
  - Création des relations associatives entre concepts de différents domaines du thésaurus.
  - Maintenance du thésaurus.

Exemples :
Thésaurus multilingue et pluridisciplinaire de l'UNESCO
http://vocabularies.unesco.org/browser/thesaurus/fr/
Thésaurus PACTOLS en archéologie https://masa.hypotheses.org/tag/thesaurus
Thésaurus Motbis multidisciplinaire et francophone du MEN https://www.reseau-canope.fr/motbis-thesagri/presentation-generale

- Décrire ses données

Comprendre le rôle des métadonnées (description, nommage, stockage, archivage)
=> Pour que les données soient utiles, elles doivent être fiables.

« Les métadonnées sont une description ou traduction numérique des données et de leur source, une version simplifiée du contenu, de sa provenance et davantage d'informations, ce qui facilite les recherches et l'utilisation d'outils et instances spécifiques sur internet, sur PC ou sur tout appareil numérique. »
(Source : https://www.purevpn.fr)

Trois principaux types de métadonnées :
- Descriptif
- Structurel
- Administratif

=> Documenter ses données permet de se repérer dans ses recherches pour ses travaux futurs, qu'elles soient repérées, consultées, réutilisées par vos collaborateurs ou par d'autres chercheurs de votre communauté.

- **Analysing Data / Analyser**

Analyser des données (format, description, méthode) 1/2

Exemples
AnalyseSHS
Service d'analyse de données pour les sciences humaines et sociales du PIREH de l'Université Paris 1 Panthéon-Sorbonne

Gephi
Logiciel pour l'analyse de réseau et la réalisation de cartographie

Ressources
Les outils numériques d'analyse de données (logiciels, bases de données),
Célya Gruson Daniel

Data Analytics Post (DAP) est un média d'information et de réflexion autour des « data sciences » porté par l'ENS Paris-Saclay

FACILE – Service de validation de formats (CINES)

L'open data et les formats de fichiers, Univ. du Littoral Côte d'Opale, 2020

Analyser des données (format, description, méthode) 2/2

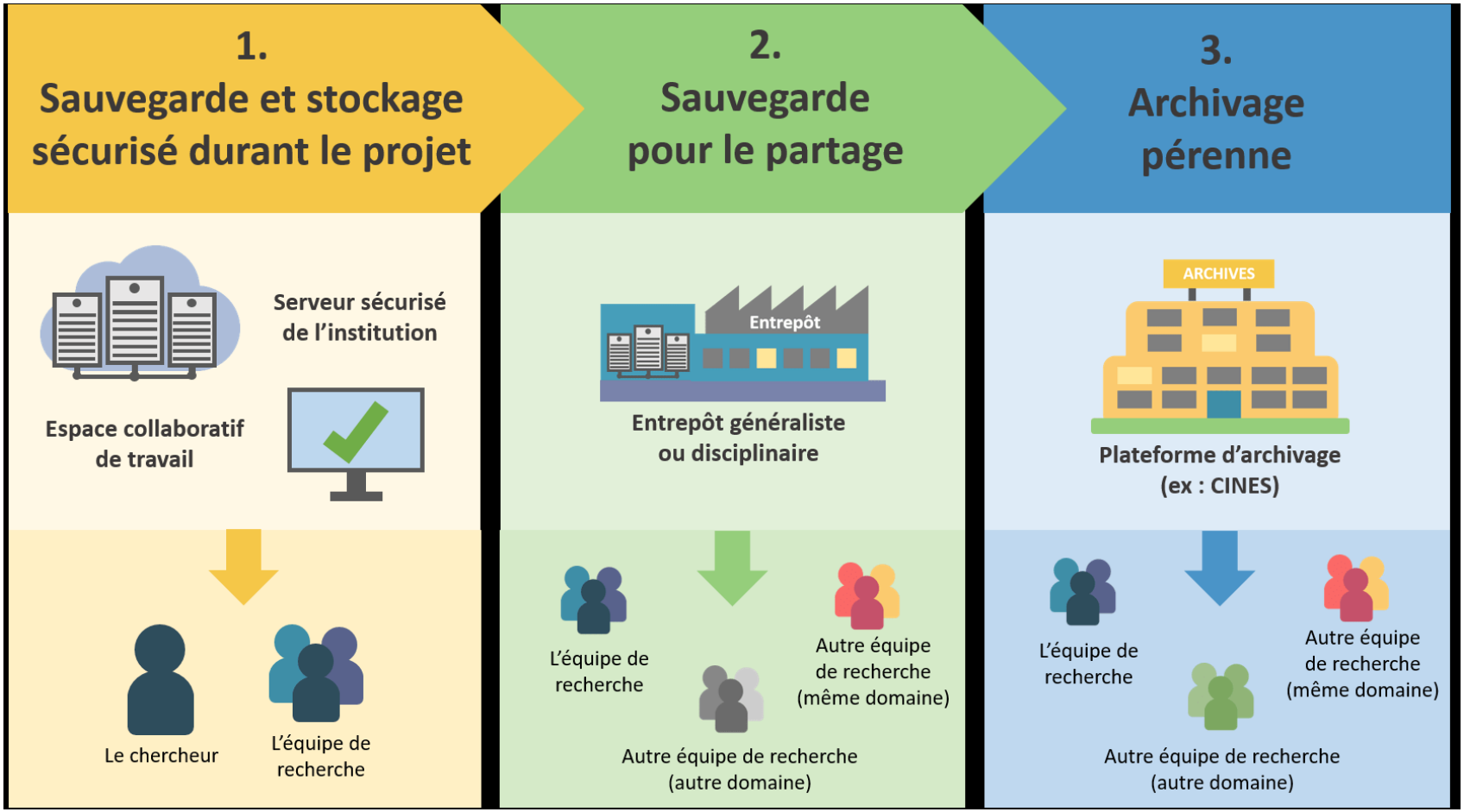
Les principaux formats sont :

Texte brut : ASCII (.txt) / sans extension
Texte formaté : TeX (.tex), OpenDocument Text (.odt), Hypertext Markup Language (.htm ou .html), XHTML (.xhtml), Feuilles de style en cascade (.css)
Tableur: OpenDocument Spreadsheet (.ods)
Document imprimable : Document PDF (.pdf)
Livre numérique : EPUB (.epub)
Données brutes : CSV (.csv) / sans extension, JSON (.json), XML (.xml)
Données géographiques : KML (.kml), SHP (.shp)

Les formats de fichier
https://www.cines.fr/archivage/des-expertises/les-formats-de-fichier/

Points d'attention :
Spécifier l'encodage utilisé pour le jeu de caractère.
=> Unicode UTF-8 est le plus utilisé.
Spécifier le séparateur.
=> La virgule et le point-virgule sont les plus utilisés, surtout en UTF-8.

- **Preserving Data / Conserver**
  Comprendre les 3 niveaux de sauvegarde des données



Source : https://doranum.fr/stockage-archivage/les-trois-niveaux-de-sauvegarde-des-donnees-de-la-recherche/

Retours d'expériences parfois malheureuses

Quatre chercheurs évoquent des incidents matériels subis par eux-mêmes ou des connaissances :
« Dans les manipulations, une sauvegarde de l'ordinateur a écrasé ce qui était déposé sur le disque, donc c'était une perte vraiment considérable pour moi. C'est comme après un incendie, je repartais à zéro »

Quatre autres chercheurs soulèvent le spectre du vol :
« Si vous me volez mon ordi portable et mon disque dur externe, qui est là et celui qui est à la maison, j'aurai perdu 6 ans de recherches avec aucun moyen de récupérer ces données-là. Donc c'est plus de ça qu'on a peur. »

Cinq chercheurs évoquent des données, ou même des outils, devenues inutilisables :
« Après j'ai eu des déboires d'enseignant-chercheur, [...] qui légitiment une réflexion sur les pratiques. [...] j'ai perdu toutes les données de ma thèse [...] à l'époque [...] j'avais enregistré et sauvegardé ça sur disquette. Et aujourd'hui, je n'ai plus les moyens de convertir ça. Données pas perdues physiquement, mais inaccessibles. Jamais pensé à les convertir »
« Je dirai que globalement avec le temps, on perd le matériau issu du recueil initial »

Source : Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet. Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2, Annexe 3 : Extraits des entretiens ; Synthèse des résultats.. [Rapport de recherche] Université Rennes 2. 2017, 26 p. ⟨hal-01635186v2⟩

Partager et diffuser ses données : les entrepôts, les licences, le DOI, le cadre légal

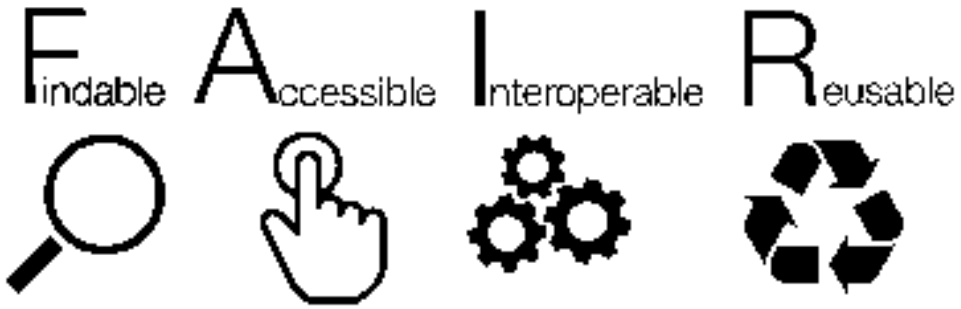
Catégories d'entrepôts de données

- Institutionnels (en France) : Dataverse Cirad, Datapartage (INRAe), DataSuds (IRD)
- Pluridisciplinaires et internationaux : Zenodo, Dryad, Figshare
- Thématiques ou disciplinaires : GenBank (séquences génétiques), TRY (caractères botaniques), GBIF (biodiversité), Pangaea (sciences de la terre et de l'environnement), WormBase (nématologie), Movebank (mobilité animale), West African Vegetation, DataFirst (enquêtes socio-économiques en Afrique), Protocols.io (protocoles), etc.
- Liés à un éditeur : GigaDB (Oxford Univ. Press), Dataverse Ubiquity Press, Dataverse *Economics*

=> Vérifier le cadre légal de diffusion de ses données avec l'outil d'aide à la décision de Andro M., Morcrette, N., Gandon, N., Système expert d'aide à la décision pour diffuser les données de la recherche
http://www.bibliotheque-numerique.fr/DonneesDiffusables.php

II.3 Les principes FAIR

Les 4 principes FAIR sont :



Findable
Les données doivent être faciles à trouver à la fois par les humains et par les systèmes informatiques.

Accessible
Les données doivent être stockées à long terme de façon à ce qu'elles puissent être facilement accessibles et/ou téléchargées.

Interoperable
Les données doivent être lisibles et utilisables par différents systèmes informatiques pour permettre le partage et la réutilisation.

Reusable
Les données doivent être prêtes à être réutilisées pour une future recherche et à être traitées en utilisant des méthodes informatiques.

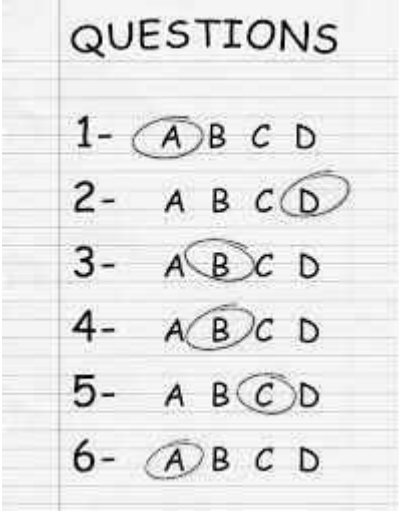
Facile à trouver	Accessible	Interopérable	Réutilisable
•Identifiant (unique et pérenne : DOI) •Versionning, organisation et nommage des fichiers •Métadonnées et vocabulaires (mot-clés) standardisés, description multilingue •Entrepôt choisi : type de recherches/moissonnages offerts (SQL, APIs,...)	•Données librement accessibles : Lesquelles ? Où ? •Données non librement accessibles : Pourquoi ? Comment ? Conditions, Embargo,... •Documentation et logiciel nécessaires : fournis, open source ? •Contacts et accès pérennes	•Formats de données standards et non propriétaires •Métadonnées standardisées et de préférence interdisciplinaires •Ou Lien (mapping) vers des vocabulaires standards •Liens standardisé vers d'autres données (relatedIdentifier, relatedPublication)	•Licence choisie •Processus d'assurance qualité •Description fine de la collecte et des traitements •Documentation et logiciel nécessaires •Partage pendant et après le projet : cf. Accessible

24

Source : Hanka Hensens, Gerer ses données avec un Plan de Gestion de Donnees (PGD/DMP), le 20 septembre 2018, dans le cadre des JeudIST de l'IRD Occitanie.
https://www.slideshare.net/IST_IRD/grer-ses-donnes-avec-un-plan-de-gestion-de-donnes-pgddmp

QCM :

https://app.wooclap.com/TEUFRI?from=event-page



III. Ressources et outils spécialisés

III.1 Ressources en autoformation

DoRANum : https://doranum.fr/
Cat OPIDoR : wiki des services dédiés aux données de la recherche https://cat.opidor.fr/
Inist : https://www.inist.fr/
CoopIST, CIRAD : https://coop-ist.cirad.fr/

III.2 Les infrastructures d'accompagnement spécialisées en SHS

Huma-Num

PROGEDO
Développer la culture des données
DATA INFRASTRUCTURE

PUDN
PLATEFORME UNIVERSITAIRE DE DONNÉES DE NANTERRE

DMP OPIDoR

HumaNum : https://www.huma-num.fr/
OPIDoR : https://opidor.fr/
Progedo : https://www.progedo.fr/
Plateformes Universitaires de Données (PUD)
https://www.progedo.fr/services/plates-formes-universitaires-de-donnees/

III.3 Quelques outils et services complémentaires

Référentiels et logiciels de traitement

Standards de métadonnées/données
Liste de standards de métadonnées ou données utilisés pour décrire les produits de recherche en application des principes FAIR.

Logiciels de documentation
Liste d'outils pour saisir les métadonnées et la documentation

Logiciels d'exploration documentaire et statistique de textes
Liste d'outils d'analyse de données textuelles (ADT), aussi appelée textométrie.
=> Dont Hyperbase (CNRS et Université Nice Sophia Antipolis), IRaMuTeQ (LERASS – EA 827, Université Toulouse), LEXICO5 (SYLED – EA 2290, Sorbonne Nouvelle), etc

III.3 Quelques outils et services complémentaires

Référentiels et logiciels de traitement

Tropy, pour organiser des corpus iconographiques
=> Logiciel de bureau installé sur la machine (comme Zotero, logiciel de gestion de références bibliographiques).

Gephi pour l'analyse de réseau et la réalisation de cartographie
=> Outil en ligne de cartographie d'ensemble de données très variés et de visualisation de réseaux.

Audacity, pour l'édition de fichiers audio et d'enregistrements.
=> Outil d'enregistrement, de manipulation, de mixage de fichiers audio sous divers formats.

Open Office Writer et Free OCR, traitement de texte et d'océrisation (reconnaissance de caractères d'un texte au format numérique).

- Offre de services DSIUN Paris 1 accessible depuis l'ENT :

Espaces collaboratifs Nuxeo / Offre Microsoft 365 / One Drive / Sharepoint
Reconnaissance de caractères et d'édition de pdf : ABBYY FineReader

III.3 Quelques outils et services complémentaires

ISTEX
Initiative d'excellence
en Information Scientifique et Technique

DataCite
FIND, ACCESS, AND REUSE DATA

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Pour aller plus loin – Ressources sur la gestion des données

GopenDoRe, un jeu de cartes coopératif sur les données par l'équipe de DoRANum

Guide de bonnes pratiques sur la gestion des données de la Recherche, groupe de travail inter-réseaux « Atelier Données », CNRS, 2021

Guide de Gestion des données de recherche, Université du Québec, (maj le 19/03/2021)

Les bonnes pratiques de la Science Ouverte appliquées aux thèses de doctorat, support de formation doctorale (Université de Lille), 2020

Passeport pour la Science Ouverte, guide pratique à l'usage des doctorant·e·s

Ressources et autoformations sur la gestion et le partage des données de la recherche (DoRANum)

Une série de capsules vidéos décline le Passeport pour la science ouverte :
· Qu'est-ce que c'est la science ouverte ?
· Des ressources ouvertes à découvrir
· La gestion des données et le cycle de vie des données de la recherche
· La thèse en accès ouvert
· La diffusion des travaux scientifiques en accès ouvert

Ancelin-Fabre Justine (2021), Formation « Introduction aux données de la recherche », 21 mai 2021, URFIST de Paris (support de formation).

Gruttemeier, Herbert, et Thérèse Hameau (2016), « Accès aux données scientifiques et contraintes juridiques – une question d'équilibre », *I2D - Information, données & documents*, 53(2), pp. 20-22. https://doi.org/10.3917/i2d.162.0020

Maurel Lionel (2018), La réutilisation des données de la recherche après la loi pour une République numérique. *La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques*, Presses Universitaires de Provence, ISBN 9791032001790. ⟨hal-01908766⟩

Rioufreyt, T. (2018). La transcription outillée en SHS. Un panorama des logiciels de transcription audio/vidéo. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 139 (1), pp. 96–133. https://doi.org/10.1177/0759106318762455

van de Weghe Tiphaine, Bessagnet Marie-Noelle, et Roose Philippe (2018), *Des données particulières : les données de la recherche en Sciences Humaines et Sociales*. 34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA 2018), Oct 2018, Bucarest, Roumanie. ⟨hal-01928548⟩

Amélie COLLIN, Service appui à la recherche et science ouverte, SCD

contacts :
appui-recherche-scd@univ-paris1.fr

Cycl@doc – 2024

UNIVERSITÉ PARIS 1

PANTHÉON SORBONNE