Principes de gouvernance

des données, des algorithmes et des codes de la recherche

Université de Lille

Principes de gouvernance des données, algorithmes et codes de la recherche

à l'Université de Lille

La production des données, algorithmes et codes de la recherche est en progression extrêmement rapide, avec un doublement des volumes de données produites tous les 3 ans[1].
Dans ce contexte, il s'avère nécessaire de mettre en place une politique de gouvernance de la donnée à l'échelle de chaque établissement de l'enseignement supérieur et de la recherche.

Les pratiques de la recherche ont par ailleurs profondément évolué au cours des dernières années, avec un accroissement des exigences liées à la bonne gestion des données de recherche, suivant des principes « FAIR » qui facilitent l'accès aux données, parfois considérées comme faisant partie du patrimoine scientifique et technique de la Nation.

1. *The first count of Research Data at scale. The characterisation experiences from Australia and New Zealand. Authors: Ai-Lin Soo, Rhys Francis, Luc Betbeder-Matibet, Nick Jones, Claire Rye. Presented at International Data Week, Salzburg, October 2023.* https://www.scidatacon.org/IDW-2023-Salzburg/sessions/571/paper/1152/

Ces exigences sont portées par les tutelles administratives comme par les agences de financement (Agence Nationale de la Recherche, Commission européenne, etc.) et s'inscrivent dans un triple contexte :

1 Le mouvement pour une science ouverte

qui vise à permettre la diffusion sans entrave des résultats, des méthodes et des produits de la recherche scientifique, pour répondre aux enjeux d'intégrité, de transparence et de reproductibilité de la recherche. Les enjeux et objectifs nationaux sont mis en évidence dans le Deuxième plan national pour la science ouverte (2021).

2 La politique de la donnée

priorité stratégique de l'État depuis la remise du rapport Bothorel (« Pour une politique publique de la donnée », décembre 2020) et déclinée pour le Ministère de l'Enseignement Supérieur et de la Recherche en septembre 2021 (Politique des données, des algorithmes et des codes sources). La politique de la donnée scientifique s'articule autour de trois enjeux majeurs : l'innovation, la confiance et la simplification.

3 Le rapport de la mission Gillet

sur l'écosystème de la recherche et de l'innovation (juin 2023), qui vise à une simplification de la structure et du fonctionnement de l'Enseignement Supérieur et de la Recherche, ce qui passe – entre autres – par une consolidation et une meilleure gestion des données des universités et des unités, à des fins de pilotage notamment.

Les initiatives se multiplient, à l'international comme en France, pour accompagner les communautés scientifiques, co-construire des outils et structurer des réseaux. À l'international, on pensera notamment au *European Open Science Cloud* (EOSC), à la *Research Data Alliance* (RDA), au Comité de données pour la science et la technologie / *COmmittee on DATA of the International Science Council* (CODATA) ou encore au *World Data System* (WDS). En France, un nombre croissant d'établissements d'Enseignement Supérieur et de Recherche ainsi que d'Organismes Nationaux de Recherche se dotent de politiques en la matière, ou au moins de mesures relatives à la gouvernance des données. Recherche Data Gouv[2] représente un écosystème complet autour des données de la recherche française, en articulant des outils (entrepôt, catalogue) et des services d'accompagnement (centres de référence thématiques, centres de ressources, ateliers de la donnée).

Parallèlement à ce contexte politique et institutionnel, des plateformes commerciales mettent à disposition des services pour stocker ou manipuler des données (ex. Amazon, Google, etc.), diffuser des données (ex. les éditeurs, via les revues) ou encore pour recenser les données produites par un établissement (ex. Data Monitor, Elsevier). Ces données sont stockées en dehors du cadre réglementaire européen - Règlement Général sur la Protection des Données (RGPD) - et sans prendre en compte les enjeux de souveraineté numérique.

Il apparaît donc nécessaire de formaliser dans le présent document les principes de gouvernance des données, algorithmes et codes de la recherche, générés au sein des unités de recherche de l'université de Lille. L'Université souligne par ailleurs, par ce document, l'importance de la gestion des données de recherche, son rôle incontournable dans l'excellence en matière de recherche, mais également dans la prise en compte des enjeux liés à la sobriété numérique.

Les principes de gouvernance des données, algorithmes et codes de la recherche à l'Université de Lille sont élaborés en tenant compte des exigences légales et réglementaires[3] et sont pensés en complémentarité avec les autres documents stratégiques de l'Université, et particulièrement :

- Politique du numérique à l'Université de Lille, 2022 ;
- Feuille de route pour la Science ouverte de l'université de Lille, 2021[4] ;
- Plan de transition écologique de l'université de Lille. Feuille de route 2023-2033, 2023.

2. recherche.data.gouv.fr/fr
3. voir Annexe 1 : Textes de loi et réglementations applicables aux données de recherche
4. www.univ-lille.fr/recherche/publier-et-diffuser-la-recherche/science-ouverte

Périmètre

Sont concernés ici les données, algorithmes et codes de la recherche, générés au sein des unités de recherche de l'université de Lille.
Sont exclues du périmètre : les données à caractère administratif et de pilotage.

DÉFINITIONS :

Données de recherche : ensemble des informations factuelles, collectées, observées ou créées sous une forme numérique et à des fins de recherche[5].

Jeu de données (ou ensemble de données, dataset en anglais) : 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.

Algorithme : suite d'instructions qui décrit le déroulé pour la résolution d'un problème posé qui précède l'écriture d'un code.

Code source : mise en œuvre et formalisation d'un algorithme dans un langage informatique. Il s'agit d'un texte écrit par (au moins) un ou plusieurs auteurs humains, suivant un langage (ex. : C, Java, Python, R, Ocaml, Scilab, etc.)

Logiciel : description, dans un ou plusieurs langages informatiques, d'un processus de traitement de données ou de modélisation que l'on souhaite faire réaliser par un ordinateur. C'est un ensemble comprenant le code source, l'exécutable (compréhensible par l'ordinateur), la documentation, la licence associée, etc.

On prêtera une attention toute particulière aux données, codes et logiciels associés aux publications, afin de veiller à ce que ces données ne soient pas captées par les grands éditeurs commerciaux.

Les principes ici énoncés s'appliquent à l'ensemble des personnes créatrices et utilisatrices des données de recherche au sein de l'Université de Lille (enseignants-chercheurs, doctorants, ingénieurs et personnels d'appui à la recherche, communauté étudiante, etc.).

Ce document stratégique a vocation à être complété par des procédures visant à faciliter sa mise en application. Il est par ailleurs, pour les Unités Mixtes de Recherche notamment, complémentaire des principes établis par les autres universités et organismes de recherche, quand ils existent[6].

5. voir Annexe 2 : Définitions des données de recherche
6. voir Annexe 3 : Partenaires de l'Université ayant une politique formalisée sur la gestion des données

Principe 1

Les données, les algorithmes et les codes doivent être partagés et réutilisés en respectant les valeurs de la science

L'Université de Lille encourage et soutient tous les chercheurs à adopter de bonnes pratiques en matière de gestion des données de recherche. Une saine gestion des données de recherche permet aux chercheurs d'utiliser, de conserver et, le cas échéant, de partager en toute sécurité leurs données de recherche. Elle permet en outre de renforcer la rigueur scientifique de la recherche et la collaboration des chercheurs dans leurs domaines.

La gestion responsable de ces données ainsi que leur partage répondent aux enjeux et aux attentes en termes notamment de transparence et d'intégrité scientifique.

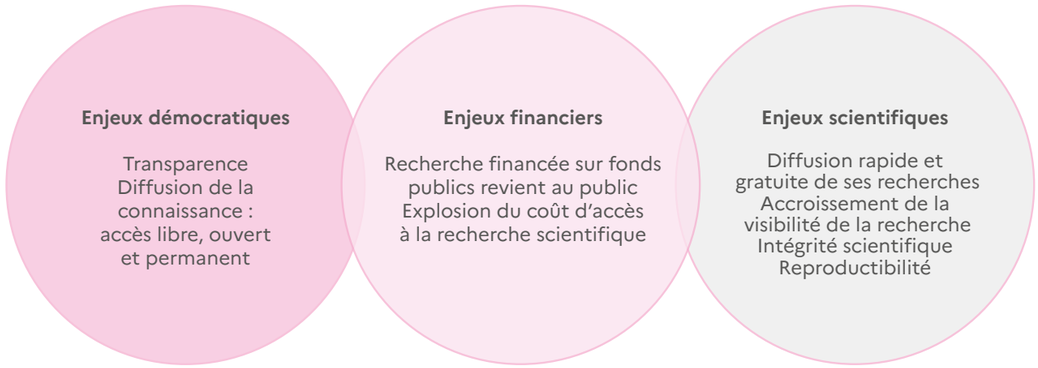

Figure 1 : Enjeux liés au partage des données

On distinguera les notions de partage des données et d'ouverture large. Les données peuvent être partagées tout en restreignant l'accès (accès sur authentification, après accord du producteur des données, etc.). Les données, algorithmes et codes doivent tous être partageables mais tous n'ont pas vocation à être ouverts pour autant (cf. Principe 3).

Enfin, quand la recherche s'appuie sur l'utilisation de données produites et mises à disposition par un tiers (un autre chercheur, une autre unité de recherche, une institution, une entreprise…), une démarche stricte d'intégrité scientifique doit être appliquée : citation des producteurs de données, diffusion du lien permettant d'accéder aux données réutilisées, respect des licences accordées pour l'utilisation de ces données, etc.

Principe 2

Les données doivent être Faciles à trouver, Accessibles, Interopérables, Réutilisables (FAIR)

Qu'elles soient destinées à être librement accessibles (ouvertes) ou partageables à usage restreint uniquement, les données doivent être gérées pour les rendre FAIR. Cela suppose d'adopter des bonnes pratiques dès leur création (voire en amont), pour être capable de gérer correctement leur partage, leur destruction ou leur archivage.

Une bonne gestion des données consiste à :

- Réaliser des plans de gestion de données (« data management plan »), y compris quand cela n'est pas requis par les agences de financement. Ces outils sont indispensables à la qualité des projets scientifiques et des résultats obtenus. Ils permettent dès la conception du projet de se poser les bonnes questions sur les modalités de gestion et de valorisation de tous les résultats, en abordant des aspects techniques, juridiques, éthiques, scientifiques et économiques (dont l'estimation des coûts de gestion et de FAIRisation qui doivent être insérés dans les budgets de projets).
- Mettre en œuvre les principes FAIR, dont :
    - le dépôt des données dans un entrepôt quand c'est possible et dans tous les cas leur description à l'aide de métadonnées descriptives (contributeurs, établissement, licences, financeurs du projet…),
    - l'identification des jeux de données (utilisation d'identifiants pérennes de type DOI),
    - la documentation des jeux de données, nécessaire pour les rendre compréhensibles dans leur contexte, via la rédaction d'un fichier Readme par exemple,
    - l'utilisation de formats et de systèmes non propriétaires si possible, et a minima de formats garantissant une forme de pérennité[7],
    - la préparation et l'organisation des données en vue de leur archivage dans un Système d'Archivage Electronique (SAE).

7. Sur le sujet, on s'inspirera notamment des préconisations des Archives nationales de France - https://francearchives.gouv.fr/fr/section/88482503 - et de la Bibliothèque nationale de France - https://bnf.hal.science/hal-03374030/document - qui ont élaboré des politiques de formats.

Principe 3

Les données doivent être aussi ouvertes que possible, aussi fermées que nécessaire, décrites dans tous les cas

Les données en cours de traitement, non finalisées, non publiées, n'ont pas vocation à être diffusées largement[8].

Les données abouties qui peuvent être diffusées largement à l'issue d'un projet de recherche seront diffusées dans des entrepôts de données publics, qu'ils soient thématiques (ex. Nakala, Gaia Data…) ou généraliste (Lillodata[9]). Elles devront être associées à une licence ouverte de type Licence Ouverte 2.0 (Etalab) ou CC-BY, afin de garantir une réutilisation aussi large que possible, permettant fouille de données ou entrainement de modèles d'apprentissage profond, si nécessaire.

Plusieurs raisons peuvent légitimement et règlementairement empêcher l'ouverture libre, immédiate et universelle des données (données personnelles, données sensibles, secret industriel, secret défense, certaines données ZRR[10], certaines données faisant l'objet d'une valorisation économique, etc.).
Néanmoins, même quand les données ne peuvent être diffusées ni partagées, elles devront être décrites dans un entrepôt, afin de faciliter l'identification des recherches produites (cf. Principe 2).

8. voir Annexe 4 : Le cycle de vie des données
9. Lillodata : entrepot.recherche.data.gouv.fr/dataverse/lillodata
10. Zones à Régimes Restrictifs. « Toutes les données produites par des laboratoires situés en zone à régime restrictif ne sont pas automatiquement exclues du principe d'ouverture par défaut ». (Guide d'application de la Loi pour une République numérique pour les données de la recherche, décembre 2022).

Principe 4

Les caractéristiques particulières des codes source et des logiciels doivent être reconnues

Le logiciel est un objet hybride au sein de la recherche, dont il est à la fois moteur et outil, mais également objet de recherche à part entière. Le logiciel peut évoluer rapidement au gré des modifications apportées par une équipe de développeurs ou par des tiers et associe une communauté de contributeurs et des actions particulières de valorisation. Il présente donc des enjeux spécifiques par rapport aux données de la recherche.

La mise en œuvre de bonnes pratiques, quelle que soit la taille du code, est essentielle pour faciliter le travail du scientifique pendant toute la durée de vie du logiciel et aider à sa pérennité. Ces bonnes pratiques consistent à :

- utiliser une forge logicielle (instance GitLab de l'université de Lille par exemple),
- associer à son code une documentation utilisateur et développeur,
- définir une licence logicielle en fonction de l'objectif de diffusion et la déployer dans le dépôt logiciel,
- signaler son code sur HAL et l'archiver sur Software Heritage pour en garantir la pérennité et la reproductibilité,
- mentionner dans son plan de gestion de données les logiciels ayant servi à créer, acquérir, transformer ou exploiter les données en tant que productions scientifiques. Lorsque le contexte de développement d'un logiciel le justifie (durée du projet, complexité particulière…) des modèles du plan de gestion de logiciel peuvent éventuellement être utilisés.

Principe 5

Les enjeux liés à la sobriété numérique doivent être pris en compte

Le numérique, dont l'ouverture des données, est source d'une empreinte environnementale non négligeable. Pour limiter cet impact et aller vers une plus grande sobriété des données, il devient nécessaire de mettre en place de bonnes pratiques sur l'ensemble du cycle de vie des données :

- évaluer la pertinence à créer de nouvelles données quand d'autres peuvent être réutilisées,
- limiter le stockage de versions obsolètes, restreindre le volume des données quand c'est possible,
- s'interroger sur les modalités de diffusion selon l'usage et l'intérêt des jeux de données,
- choisir des logiciels de traitement des données open source, ne développer des logiciels que si nécessaire, chercher à réduire l'impact du code[11]…

Ces enjeux devront néanmoins se penser en complémentarité avec le cadre réglementaire des archives, notamment concernant l'élimination des archives publiques.

11. « Je code : les bonnes pratiques en éco-conception de service numérique à destination des développeurs de logiciels », https://hal.science/hal-03009741

Rôles et responsabilités

UNIVERSITÉ

L'Université de Lille s'engage à fournir à ses producteurs de données des infrastructures et services d'accompagnement solides.

L'Université s'assure de créer un environnement sécurisé pour la gestion des données, assure la mise à jour et le bon fonctionnement dudit environnement selon les normes reconnues de sécurité. Elle met à disposition des solutions matérielles et logicielles, par exemple Nextcloud, Lillodata (espace institutionnel de Recherche Data Gouv), Gitlab, espaces de stockage du Mésocentre, Service d'Archivage Electronique, etc.

Elle mobilise par ailleurs des expertises internes

collectives :

- Comité d'éthique de la recherche ;
- Comité de protection des personnes ;
- Direction Générale Déléguée Recherche et Valorisation ;
- Direction Générale Déléguée au Numérique ;
- Direction sûreté sécurité-défense ;
- Direction des données personnelles et archives ;
- Direction des affaires juridiques ;
- Service commun de documentation.

comme individuelles :

- Administrateur des données, des algorithmes et des codes (ADAC) ;
- Délégué à la Protection des Données (DPO) ;
- Fonctionnaire de Sécurité et de Défense (FSD) ;
- Personne Responsable de l'Accès aux Documents Administratifs et des questions relatives à la réutilisation des informations publiques (PRADA) ;
- Responsable de la Sécurité des Systèmes d'Information (RSSI).

L'Université s'assure du respect de la législation et des normes applicables en matière de sécurité de l'information et de protection des renseignements personnels.

L'Université s'engage par ailleurs à fournir l'accompagnement nécessaire pour son personnel afin d'aller vers une saine gestion des données de recherche, notamment via l'atelier de la donnée Lille Open Research Data[12] :

- Un soutien apporté aux équipes, pour diffuser la culture de la donnée et faire connaître les bonnes pratiques et les outils disponibles pour gérer, diffuser, archiver ses données, codes et logiciels : accompagnement individualisé, formations doctorales, formation continue des personnels de l'université, webinaires, etc. ;
- Un guichet unique permettant l'accès facilité à l'ensemble des acteurs de la donnée au sein de l'université ;
- La mise à disposition de ressources pratiques.

Enfin, l'Université s'engage à prendre en compte des critères liés à la bonne gestion et à l'ouverture des données lors de l'évaluation de ses unités et de ses chercheurs.

L'Université appuie son action sur une cellule « gouvernance des données de recherche » autour d'un administrateur des données, des algorithmes et des codes, cellule qui a vocation à instruire les cas complexes qui ne pourraient être traités par les services, personnes ressources, référents et instances existantes.

COMPOSANTES

Dans le cadre de leur politique de recherche, les composantes doivent tenir compte des principes de gouvernance des données et prendre part à la bonne circulation de l'information. Elles ont vocation à intégrer ces enjeux dans leurs formations, en articulation avec les autres acteurs concernés.

UNITÉS DE RECHERCHE

Les directions d'unité doivent relayer et soutenir l'appropriation des enjeux liés à la bonne gestion des données, et orienter vers les services d'appui quand nécessaire. Plusieurs leviers existent :

- La désignation d'un « correspondant données » au sein du laboratoire, qui peut appartenir à n'importe laquelle des co-tutelles de l'unité ;
- L'organisation de formations au sein de l'unité, en s'appuyant sur les services support quand nécessaire ;
- La rédaction d'un plan de gestion de données de l'unité, afin de décliner la stratégie institutionnelle dans l'écosystème local, de mieux prendre en compte les préconisations et solutions techniques des différentes tutelles, et enfin de diffuser les bonnes pratiques.

ENSEIGNANTS-CHERCHEURS

Le chercheur doit veiller au cadre éthique et déontologique (rigueur scientifique, reproductibilité…) appliqué à la gestion des données. Il a la responsabilité de se former à ces enjeux autant que nécessaire dans le cadre des moyens qui lui sont alloués.

12. voir Annexe 5 : Atelier de la donnée Lille Open Research Data

Annexes

ANNEXE 1
TEXTES DE LOI ET RÉGLEMENTATIONS APPLICABLES AUX DONNÉES DE RECHERCHE

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique
- Code du patrimoine, Livre II : Archives
- Code de la recherche
- Code de la propriété intellectuelle
- Code des relations entre le public et l'administration
- Règlement général sur la protection des données - UE 2016/679
- Instruction interministérielle relative à la protection des systèmes d'information sensibles n° 901/SGDSN/ANSSI

ANNEXE 2
DÉFINITION DES DONNÉES DE RECHERCHE

(définition validée par la cellule ADAC le 16/10/2023)

Les données de recherche regroupent l'ensemble des informations factuelles, collectées, observées ou créées sous une forme numérique et à des fins de recherche.

Les données de recherche n'englobent pas les objets matériels (échantillons de laboratoire, animaux de laboratoire, objets archéologiques…) ni les sources secondaires (publications, travaux universitaires et communications scientifiques…).

Les données sont de forme et de nature variable, selon les disciplines concernées et l'objet d'étude. Elles se distinguent notamment par :

- leur **mode d'obtention** : données produites dans le cadre d'expérimentation ou d'analyse par des instruments, données d'observation, données collectées lors d'une enquête ou d'un prélèvement sur le terrain… ;
- le **moment** de leur obtention dans le processus de recherche : données brutes, intermédiaires, raffinées, publiées ;
- leur **contexte de production** : partenariat industriel, laboratoire en zone à régime restrictif… ;
- leur **régime juridique** : données sensibles (Protection du Potentiel Scientifique et Technique – ZRR), à caractère personnel, couvertes par un secret (professionnel, défense ou industriel par exemple), soumises à un accord de confidentialité…

Exemples de types de données par disciplines (liste non exhaustive) :

Santé et sciences du vivant	Ex. : Données d'échantillonnage, images numériques en microscopie, séquences génomiques, etc.
Sciences et technologies	Ex. : Relevés météorologiques, sismogrammes, images satellitaires, images numériques en microscopie, mesures, modèles, etc.
Sciences humaines, sociales et juridiques	Ex. : Résultats d'enquêtes, séries statistiques, transcription d'archives ou d'entretiens, relevés topographiques, données géospatiales, corpus annoté, photographies, captations vidéo/sonores, etc.

ANNEXE 3
PARTENAIRES DE L'UNIVERSITÉ AYANT UNE POLITIQUE FORMALISÉE SUR LA GESTION DES DONNÉES

Pour les unités de recherche partagées avec d'autres tutelles, les établissements suivants disposent d'une politique formalisée sur la science ouverte, incluant un volet sur les données de recherche :

CNRS	*Feuille de route du CNRS pour la science ouverte, 18 novembre 2019*
ENSAM	*Charte pour une science ouverte, Arts et Métiers, Mai 2023*
IMT	*Charte Science ouverte, Institut Mines Télécom, Avril 2024*
INRAE	*Gouvernance des données, algorithmes et codes, INRAE*
IRD	*Feuille de route pour une science ouverte et partagée et Plan d'action, juin 2021*
ONERA	*Science ouverte : les quatre engagements de l'ONERA, janvier 2022*
Université d'Artois	*Feuille de route de l'Université d'Artois pour une science ouverte*
Université de Lille	*Feuille de route pour la Science ouverte de l'Université de Lille, 2021*
Univeristé Paris Cité	*Feuille de route pour la Science ouverte à Université Paris-Cité, 15 juin 2021*
Université Polytechnique des Hauts de France (UPHF)	*Plan Science ouverte de l'UPHF et de l'INSA Hauts-de-France, 2021*

ANNEXE 4
LE CYCLE DE VIE DES DONNÉES

La gestion des données de la recherche concerne l'ensemble du cycle de vie des données, c'est-à-dire des opérations de collecte, production, description, stockage, traitement, analyse, mise en accès et archivage des données.

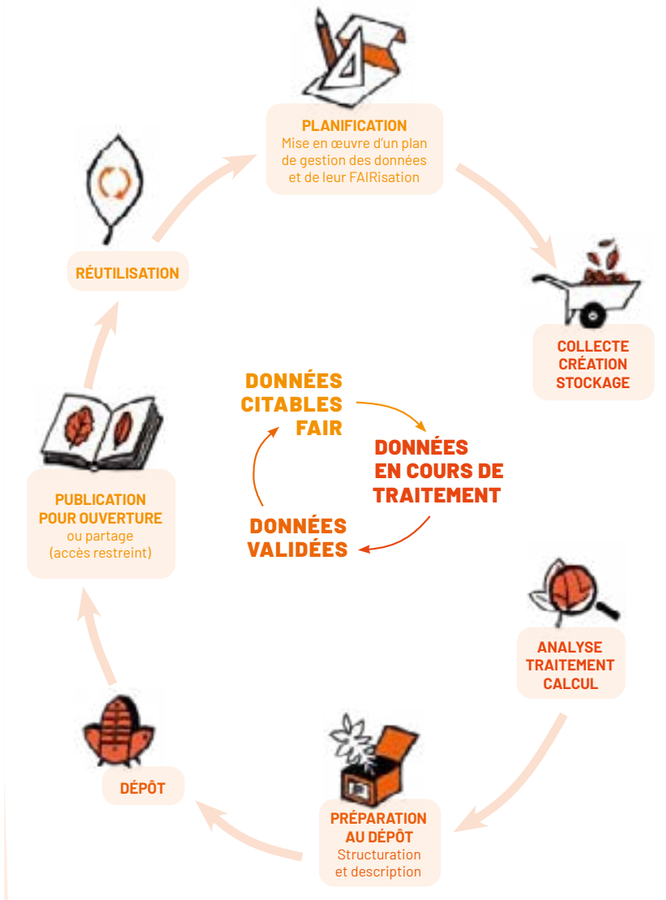

Figure 2 : Cycle de vie des données
(Crédits : 4 minutes 34 - Passeport pour la science ouverte, 2024 - CC-BY-SA)

ANNEXE 5
PRÉSENTATION DE L'ATELIER DE LA DONNÉE « LILLE OPEN RESEARCH DATA »

Lille
Open
Research
Data

LORD (Lille Open Research data) est un dispositif validé « atelier de la donnée » par le Ministère de l'enseignement supérieur et de la recherche, au sein de l'écosystème Recherche Data Gouv.

Il s'agit d'un dispositif complet d'assistance et d'accompagnement des équipes de recherche de la métropole lilloise sur l'ensemble des volets concernant la gestion des données, des codes et des logiciels de la recherche. Il se décline en plusieurs actions destinées à couvrir l'ensemble des besoins des équipes de recherche : actions de sensibilisation et de formation, accompagnement individualisé, mise en place du guichet unique.

L'atelier de la donnée LORD est porté par l'Université de Lille et associe plusieurs établissements et organismes de recherche : Centrale Lille, IMT Nord-Europe, Université de Liège, Institut Pasteur de Lille, CHU de Lille, INRAE, INRIA, CNRS, ENSAM, ONERA.

Découvrir la page LinkedIn de LORD :
www.linkedin.com/company/lille-open-research-data/

Contacter LORD via le guichet unique :
bit.ly/lord-support

Conception : Service commun de documentation - Université de Lille
Version approuvée le 6 mars 2025 par le conseil scientifique de l'Université de Lille