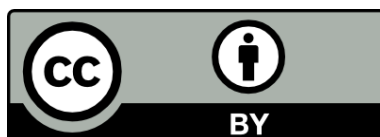


La Fresque des données de recherche

Lille
Open
Research
Data

DATAACTIVIST



Les cartes à retenir



Plan National
pour la Science
Ouverte



Open Data en
France



Données de la
recherche



Règlement Général sur
la Protection des
Données (RGPD)



Politique publique
de la donnée



Cycle de vie de la
donnée

Les cartes à retenir



Principes
FAIR



Un écosystème au service du partage
et de l'ouverture des données de recherche

Recherche Data
Gouv



*"Aussi ouvert que
possible, aussi
fermé que
nécessaire"*



Plan de Gestion
de Données (PGD)



LILLODATA

Dépôt des données de recherche
produites et collectées par la
communauté scientifique de l'Université
de Lille.
[voir la fiche](#)

Lillodata

Lille
Open
Research
Data

LORD et Ateliers
de la Données

Glossaire

- **Administrateurs des données, algorithmes et codes (ADAC):** La feuille de route 2021-2024 du MESRI sur la politique des données, des algorithmes et des codes sources prévoit la création d'un réseau d'administrateurs des données, des algorithmes et des codes sources de la recherche. Placés auprès de la gouvernance de leur établissement, ils sont des acteurs clé de la gouvernance des données de la recherche.
- **"Aussi ouvert que possible aussi fermé que nécessaire":** Le principe "Aussi ouvert que possible, aussi fermé que nécessaire" met en avant l'importance de l'accès libre aux données de la recherche, tout en reconnaissant la nécessité de certaines restrictions, en particulier en conformité avec le RGPD (Règlement Général sur la Protection des Données) qui impose des règles strictes sur la protection de la vie privée. Ce principe vise à équilibrer ouverture et protection des données, en particulier sur les données dites sensibles et/ou personnelles.
- **Bonnes pratiques - codes et logiciels :** Les bonnes pratiques en matière de logiciel libre assurent la pérennité et la réutilisation des codes et logiciels. Parmi elles, on retrouve la publication du code source d'un logiciel sous une licence libre reconnue (GNU GPL, MIT...) permettant la modification et la redistribution, le maintien d'une documentation complète et à jour ou encore l'adoption d'un système de gestion de version (comme Git).
- **Captation de données par les éditeurs scientifiques:** "Les éditeurs scientifiques associent de plus en plus les articles à un jeu de données, dans un souci d'intégrité scientifique, de reproductibilité et de transparence. Néanmoins, il arrive que les éditeurs demandent aux auteurs d'utiliser leur propre entrepôt commercial, ce qui pose de sérieux problèmes de souveraineté numérique, de respect du RGPD et de captation de données publiques par des acteurs privés."

Glossaire

- **Charte internationale de l'Open Data (Open Data Charter, ODC)** : La Charte internationale de l'Open Data a été adoptée en 2015 par 17 pays dont la France, et pose six grands principes de l'ouverture des données par les gouvernements. Ceux-ci sont l'ouverture par défaut (1), en temps opportun et de manière exhaustive (2), dans un format accessible et réutilisable (3), comparables entre pays et interopérables (4), afin d'améliorer la gouvernance et encourager la participation citoyenne (5), ainsi que pour favoriser le développement inclusif et l'innovation (6).
- **Chaire UNESCO pour la Science ouverte** : La chaire UNESCO en science ouverte a été créée à la suite des recommandations de l'UNESCO publiées en 2021, et qui préconisaient un accès gratuit et inclusif aux publications scientifiques et données de la recherche. Dirigée par Vincent Larivière, professeur en bibliothéconomie et sciences de l'information à l'Université de Montréal, elle est soutenue dans son travail par des organisations savantes de dix pays. Elle a pour objectif la conduite de recherches sur les pratiques et les effets de la science ouverte sur la production scientifique.
- **Crise de la reproductibilité** : La crise de la reproductibilité désigne un problème croissant dans plusieurs disciplines scientifiques où les résultats de nombreuses études ne peuvent pas être répliqués. Elle est alimentée par divers facteurs (échantillons trop petits, biais de publication, méthodes statistiques inappropriées, manque de transparence dans la méthodologie, fraude...). Elle est souvent invoquée dans la promotion d'initiatives de science ouverte favorisant des pratiques de recherche plus robustes.
- **Cycle de vie de la donnée** : Le cycle de vie de la donnée représente l'ensemble des étapes par lesquelles une donnée passe, de sa création à sa suppression. Ces étapes comprennent en général: la collecte, le stockage, le traitement, l'utilisation, le partage, puis l'archivage ou la suppression définitive. Le cycle de vie assure que la donnée est gérée de manière organisée et sécurisée tout au long de son existence, mais aussi dans sa réutilisation.

Glossaire

- **Documentation des données:** L'insuffisance de la documentation est un défaut majeur des données ouvertes. Sur data.gouv.fr, 22 % des jeux de données ont une description de moins de 180 caractères (soit un SMS) et 59 % moins de 1 000 caractères (une demi-page). C'est généralement insuffisant pour que les réutilisateurs comprennent le contenu et les limites des données.
- **Données brutes:** Il est essentiel d'ouvrir les données brutes, c'est-à-dire qui n'ont pas été modifiées, agrégées ou manipulées. Aux réutilisateurs après, d'effectuer les calculs et agrégations qu'ils souhaitent pour les rendre intelligibles par des humains. Par exemple, les fichiers Excel avec des lignes fusionnées ou des totaux devront être transformés en fichiers .csv structurés, avec un enregistrement par ligne
- **Données de la recherche :** "Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche." (OCDE, « Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics de 2007 »).
- **Données de la statistique publique:** Les données issues de la Statistique Publique sont constituées de grandes enquêtes et de bases de données issues de l'INSEE, de services statistiques ministériels et des institutions publiques. Elles intéressent particulièrement les sciences sociales. A Lille, la Plateforme Universitaire de Données de Lille (PUDL) accompagne les chercheurs pour l'accès à ces données sur Quetelet-Progedo-Diffusion et pour leur réutilisation.
- **Données personnelles et sensibles :** Dans le RGPD, les données personnelles sont décrites comme des informations permettant d'identifier directement ou indirectement une personne, comme le nom, l'adresse, l'e-mail, ou les données biométriques. Les données dites sensibles incluent celles révélant l'origine ethnique, les opinions politiques, les croyances religieuses, la santé ou l'orientation sexuelle. Le traitement de ces deux types de données nécessite des protections renforcées et le consentement explicite de la personne concernée.

Glossaire

- **Droit d'accès aux documents administratifs:** Prévu par loi CADA de 1978, aujourd'hui codifié dans le code des relations entre le public et l'administration, le droit d'accès concerne tous les documents sous toute forme produits ou reçus dans le cadre d'une mission de service public. Les personnes publiques ou privées sont tenues de les communiquer au public sauf si cela porte atteinte à un secret protégé. L'open data consiste à publier de manière volontaire les données sans attendre une requête mais il est nécessaire de le demander explicitement aux communes de moins de 3500 habitants.
- **Entrepôts de données:** Un entrepôt de données est une plateforme de stockage numérique qui permet de collecter, gérer et analyser des ensembles de données provenant de diverses sources. Ils peuvent être institutionnels, généralistes ou thématiques. En France, on dispose par exemple de Recherche Data Gouv (et sa déclinaison lilloise : Lillodata) ou encore Nakala pour les données en Sciences humaines et sociales. De plus en plus de revues scientifiques exigent désormais le partage des données dans ces entrepôts comme condition pour la publication.
- **Format ouvert :** La loi pour une République Numérique consacre l'utilisation de formats ouverts afin de faciliter l'interopérabilité et la réutilisation des données par les machines. Aussi, pour rendre ses données FAIR, il faudra notamment privilégier des formats de fichiers ouverts ou largement utilisés, par exemple .csv à la place de .xlsx.
- **Identification des plantes:** L'open data peut aussi servir dans vos randonnées ou balades en forêt. Par ex., Pl@ntNet est une application mobile de sciences participatives permettant d'identifier des plantes à partir d'une photo. Son service est fondé sur plusieurs bases de données ouvertes et les photos prises par les utilisateurs sont reversées en open data pour faciliter l'identification automatisée d'espèces végétales.
- **Licences ouvertes :** Les licences ouvertes sont des outils juridiques permettant aux auteurs de partager leurs œuvres (au sens légal du terme) tout en définissant les conditions de leur utilisation, facilitant ainsi la diffusion des créations tout en respectant le droit d'auteur. Parmi elles, on retrouve les licences Creative Commons, la Licence Ouverte 2.0 d'Etatlab ou encore la licence ODbL (Open Database Licence). Ces licences permettent notamment le partage des données de la recherche.

Glossaire

- **Loi Pour une République Numérique (2016):** Cette loi vise à moderniser la législation française face aux évolutions numériques. Elle impose notamment un principe d'ouverture des données par défaut à toutes les administrations de plus de 50 agents et aux collectivités locales de plus de 3500 habitants. Le périmètre est large et comprend les documents communiqués suite à des demandes CADA, les "bases de données" et les données "dont la publication présente un intérêt économique, social, sanitaire ou environnemental".
- **Métadonnées:** Une métadonnée est une donnée décrivant une autre donnée, facilitant son identification, sa gestion et sa réutilisation. Elle comprend des informations telles que le titre, l'auteur, la date de création, et le sujet d'un document. Un exemple de schéma de métadonnées (norme de description) est le Dublin Core, et propose un ensemble standard de 15 éléments permettant de décrire divers types de ressources numériques (documents textuels, images, vidéos, jeux de données...).
- **Open Data en France:** En France, l'Open Data est une démarche consistant à rendre accessibles et réutilisables les données publiques de manière libre et gratuite. Depuis la loi pour une République numérique de 2016, les administrations ont pour obligation de publier leurs données non personnelles. L'entrepôt gouvernemental data.gouv.fr centralise notamment ces données pour stimuler leur exploitation dans divers domaines et les rendre accessibles aux citoyens et aux entreprises pour leur réutilisation.
- **Pandémie de COVID-19:** Début 2020, la nécessité d'une diffusion rapide des connaissances produites pour lutter contre la pandémie a conduit à une adoption massive des données et des preprints, pour permettre aux chercheurs de partager rapidement leurs résultats. Des préoccupations ont cependant émergé concernant la qualité et la fiabilité des résultats publiés. Ce contexte a accentué les débats sur l'importance de l'accès libre, ainsi que sur la nécessité de garantir des normes élevées dans la recherche.
- **Plan de Gestion des Données (PGD):** Le plan de gestion des données (PGD ou "Data Management Plan" - DMP - en anglais) est un document décrivant la manière dont les données de recherche seront collectées, organisées, stockées, partagées et archivées. Il vise à assurer leur qualité, leur pérennité et leur réutilisation. Le PGD détaille aussi les aspects juridiques, éthiques et techniques, garantissant une gestion responsable et transparente des données. Ce document constitue un livrable pour les porteurs de projets financés par l'ANR ou la Commission européenne.

Glossaire

- **Plan National pour la Science Ouverte:** Le Plan National pour la Science Ouverte (PNSO) est un document officiel du Ministère français de l'Enseignement Supérieur et de la Recherche, lancé en 2018 et renforcé en 2021. Il définit des orientations stratégiques pour promouvoir la science ouverte en France. Il repose sur trois axes: la généralisation de l'accès ouvert aux publications, la structuration et l'ouverture des données de recherche, ainsi que la promotion de la science ouverte dans les pratiques et l'évaluation des chercheurs.
- **Politique publique de la donnée:** Le rapport au Premier ministre du député Bothorel sur la politique de la donnée (2020) a montré que les politiques liées aux données dépassent « la donnée et les codes source ne sont pas juste un enjeu « tech », mais d'abord et surtout une question politique, démocratique, scientifique et économique ».
- **Principes FAIR:** Les principes FAIR (anglais : Findable, Accessible, Interoperable, Reusable) sont un ensemble de bonnes pratiques qui visent à rendre les données de la recherche ouvertes faciles à trouver, facilement accessibles via des protocoles ouverts, interopérables (compatibles avec d'autres systèmes) et réutilisables. Ils impliquent l'utilisation de métadonnées standardisées, de formats ouverts et de licences claires pour favoriser une gestion, un partage et une diffusion efficaces des données de la recherche.
- **Qualité et intégrité scientifiques:** "La science ouverte devrait respecter la liberté académique et les droits humains et soutenir une recherche de haute qualité en rassemblant de multiples sources de connaissances et en mettant largement à disposition les méthodes et les résultats de recherche pour garantir un examen et un contrôle rigoureux, ainsi que des processus d'évaluation transparents" (UNESCO).
- **Recherche Data Gouv:** Recherche Data Gouv est la plateforme nationale fédérée dédiée au partage et à la valorisation des données de la recherche scientifique. Sa création a été annoncée dans le second Plan National pour la Science Ouverte (2021) et elle fut officiellement lancée en juillet 2022. Recherche Data Gouv propose des outils pour le dépôt des données (entrepôt, catalogue) et un réseau complet d'accompagnement, en local (ateliers de la donnée) et au niveau national (notamment : centres de référence thématiques, centres de ressources).

Glossaire

- **Règlement Général sur la Protection des Données (RGPD):** Le RGPD est une législation de l'UE entrée en vigueur en mai 2018. Ses principales dispositions incluent le consentement explicite des individus pour le traitement de leurs données, le droit d'accès, de rectification, de suppression, de notification en cas de violation, et la désignation dans chaque entreprise d'un délégué à la protection des données (DPO). Il régit le traitement des données personnelles, qu'il faut anonymiser en cas d'ouverture en s'assurant que les individus ne sont pas réidentifiables par croisement de données.
- **Secret statistique:** Le secret statistique interdit la communication de données issues de traitements à finalités statistiques dès lors qu'elles sont individuelles ou susceptibles d'identifier des personnes. Le secret statistique est opposable à toute réquisition judiciaire et toute utilisation des données à des fins de décision à l'égard des personnes concernées comme le contrôle fiscal.
- **Secrets protégés:** Certains documents administratifs (et donc des données) ne peuvent être communiqués lorsque leur contenu porte atteinte aux secrets suivants définis par la loi : défense nationale, conduite de la politique extérieure, monnaie et le crédit public, secret médical, vie privée, secrets industriels et commerciaux, secret des affaires...
- **Service public de la donnée:** Créé par la loi pour une République numérique, le service public de la donnée vise à faciliter la réutilisation des jeux de données de référence à plus fort impact économique et social (ex. : Base Sirene, Base Adresse Nationale, Code Officiel Géographique, Plan cadastral informatisé...) Les producteurs de ces données s'engagent sur la qualité, la disponibilité, la mise à jour et le traitement des signalements des usagers.
- **Software Heritage :** Software Heritage est une archive ouverte visant à collecter, préserver et partager le code source des logiciels. Fondée en 2016 par Inria et soutenue par l'UNESCO, elle vise à archiver et préserver le patrimoine des logiciels développés, en les rendant accessibles pour la recherche et la réutilisation. Le projet repose sur un dépôt durable, garantissant que le code source, les historiques de version et la documentation soient sauvegardés. Il est connecté à l'archive ouverte HAL depuis 2018.

Avec l'Université de Lille

- **Ateliers de la donnée & LORD:** Au sein de l'écosystème Recherche Data Gouv, les ateliers de la donnée apportent, en proximité des chercheurs, un premier niveau d'expertise et de services à la préparation et la diffusion des données. Lille Open Research Data est l'atelier de la donnée lillois, labellisé depuis 2022. Il regroupe 12 établissements et organismes de recherche.
- **Formations:** Lille Open Research Data propose plusieurs types de formations à destination des chercheurs et personnels d'appui à la recherche autour des données, codes et logiciels de la recherche. Webinaires science ouverte, Formation continue des personnels, Formations doctorales dans le cadre du collège doctoral, etc.
- **Lillodata** : Lillodata est l'espace institutionnel de l'Université de Lille sur l'entrepôt Recherche Data Gouv. Lillodata est destiné aux jeux de données produits et collectés par la communauté scientifique de l'Université de Lille.

Ma fresque idéale

Qu'est ce que l'ouverture des données de la recherche ?

A quoi ça sert ?

Comment on fait ?

Ma fresque idéale (libre)

Notes

Notes

La Fresque des données de recherche

Lille
Open
Research
Data

DATAACTIVIST

