Choisir un entrepôt et une licence pour son jeu de données diffusable en libre accès : lignes guides et recommandations

Version au 12/07/2023

Document rédigé en collaboration avec le Service Juridique de l'Ined

Table des matières

Introduction

Ouvrir l'accès aux données de la recherche suppose de mettre à disposition ses données dans un entrepôt, sous une licence de réutilisation claire facilitant leur réutilisation. Ces pratiques, reposent sur le respect des principes FAIR (Findable, Accessible, Interoperable, Reusable) qui décrivent comment organiser les données de la recherche pour en favoriser la découverte, le partage et l'accès [1].

Ce document synthétique a pour objectif de donner des conseils pour choisir l'entrepôt dans lequel déposer les données et la licence à attribuer à son jeu de données, ainsi que des ressources pouvant aider cette prise de décision. La loi de programmation de la recherche (LPR), promulguée en 2020, a introduit l'obligation pour les établissements de recherche de donner aux chercheur·es les moyens pour ouvrir les données de la recherche. L'accompagnement au dépôt de ces données dans un entrepôt et au choix de la licence de réutilisation, représente un de ces moyens.

Ces conseils concernent les **bases de données quantitatives en démographie (hors enquêtes[1]) ayant vocation à être diffusées librement**. Toutes les données de la recherche ne sont pas réutilisables selon les mêmes modalités. Des limitations (ou interdictions) à l'ouverture peuvent exister, par exemple lorsque ces données constituent des données personnelles, ou sont protégées par le droit d'auteur de tiers (partenariat), ou encore sont couvertes par le secret défense ou encore le secret professionnel[2]. Ces contraintes ne sont pas discutées ici.

1. Les entrepôts de données

1.1 Définition, utilité

Afin d'assurer l'accessibilité et d'améliorer la visibilité des données de la recherche, il est nécessaire de les déposer dans un entrepôt de données. Les entrepôts fournissent une solution de stockage à moyen ou long terme, permettent d'attribuer un identifiant (le plus souvent un DOI) à la base de données, génèrent des formats de citations, proposent d'associer des métadonnées standardisées. L'entrepôt peut être généraliste, disciplinaire, institutionnel… Il est nécessaire de déposer les données accompagnant un article scientifique comme un *data paper* dans un entrepôt de données.

1.2 Répertoires d'entrepôts

Plusieurs ressources existent et proposent une liste d'entrepôts de données. Le répertoire re3data.org en recense plus de 3000 au niveau international[3]. Plusieurs critères sont généralement identifiés pour aider au choix de l'entrepôt le plus adapté [2, 3, 4] , e.g. :

- La discipline des données couvertes.
- Le type de données et les formats acceptés.
- La réputation : les revues, les institutions, les financeurs (ou vos collègues) peuvent recommander certains entrepôts.
- La visibilité : l'entrepôt devrait permettre d'associer un DOI au jeu de données.
- La durabilité : l'intérêt d'un entrepôt de données est d'assurer le stockage des données sur le long terme. Un entrepôt soutenu par plusieurs institutions publiques ne sera que plus durable. Les certifications Core Trust Seal sont un bon indicateur à privilégier dans le choix d'un entrepôt.
- Les licences : l'entrepôt peut proposer ou imposer des licences, vérifiez que celles-ci sont en accord avec la législation s'appliquant au jeu de données que vous devez diffuser, ainsi que les besoins de vous et de vos partenaires.
- Le prix : lorsqu'il s'agit d'entrepôts privés, leurs services peuvent être facturés, notamment à partir d'une certaine quantité de données à stocker.
- Toutes autres fonctionnalités ou services proposés peuvent être des éléments qui guident votre choix. Par exemple, tous les entrepôts n'offrent pas un catalogue d'exploration en ligne de leur contenu.

1.3 Recommandations de l'Ined

En général, les **entrepôts disciplinaires** et **institutionnels** sont à privilégier s'ils existent. En France, l'entrepôt **NAKALA** maintenu par la TGIR Huma-Num est conseillé pour les données en SHS. En

[1] L'entrepôt des jeux de données (notamment, les fichiers pseudonymisés dit « Fichiers de Production et de Recherche », FPR) issus des enquêtes de l'Ined est Quetelet-Progedo-Diffusion (QPD).

[2] Le questionnaire développé par le CIRAD et disponible ici : https://www.loginos.net/base/ylxxVO (consulté le 12/07/2023) est un outil conçu pour aider les chercheurs à comprendre si leurs jeux de données sont diffusables en libre accès ou pas.

[3] En France il y a le répertoire : https://cat.opidor.fr/index.php/Entrep%C3%B4t_de_donn%C3%A9es (consulté le 12/07/2023).

alternative, s'il n'y a ni d'entrepôt disciplinaire adapté, ni d'entrepôt institutionnel, comme c'est le cas de l'Ined, l'Ined recommande aux chercheur.es de déposer leurs données dans l'**entrepôt national Recherche Data Gouv**, un entrepôt **pluridisciplinaire des données de la recherche**. Actuellement, l'Ined ne dispose pas d'espace institutionnel dans Recherche Data Gouv (cela est en cours de réflexion), mais les chercheur·es affilié·es à un établissement de recherche française peuvent déposer leurs données dans l'espace générique de Recherche Data Gouv.

Le DataLab datalab[at]ined.fr peut être sollicité par les chercheur·es de l'Ined pour des conseils sur le choix d'un entrepôt des données et peut les accompagner dans le dépôt.

> **À noter :** les **Éditions de l'Ined** n'imposent pas d'entrepôt de données à leurs auteur·es, mais recommandent de se tourner vers des entrepôts de données disciplinaires reconnus par la communauté scientifique et vers les entrepôts publics pérennes et sécurisés. Par ailleurs, le service des éditions propose aux auteur·es de déposer les compléments en ligne et données annexes aux articles de revues et aux ouvrages dans l'entrepôt NAKALA.

2. Les licences de diffusion des données

2.1 Pourquoi adopter une licence pour ses données ?

Les données issues d'une activité de recherche financée au moins pour moitié par des financements publics doivent être librement réutilisables [5]. Attribuer une licence ouverte à sa base de données permet de garantir à tous le droit d'accéder, d'utiliser et de partager les données [6]. Elles protègent l'œuvre de l'auteur, tout en précisant explicitement les conditions à respecter pour sa réutilisation. L'apposition d'une licence à un jeu de données est recommandée car cela permet d'informer clairement les ré-utilisateurs de données de la recherche sur l'étendue de leurs droits et obligation.

2.2 Répertoires de licences existantes

Afin de choisir la licence à attribuer à ses données, ou de comprendre les implications qu'une licence peut avoir sur la réutilisation d'une base de données, certaines ressources existent en ligne, par exemple :

- Le sélecteur de licence intégré au service B2SHARE de l'infrastructure européenne EUDAT
- La liste des licences recensées par SPDX (The Software Package Data Exchange)
- Diverses ressources en français présentent des guides sur la question (comme l'université de Bordeaux, le CNRS, l'IRD…)

Au niveau international, les licences Creative Commons sont largement utilisées. Elles peuvent être choisies à l'aide d'un sélecteur spécifique.

Deux licences gratuites et ouvertes sont adaptées **et recommandées** à la diffusion des données de la recherche publique en France[4] :

- l'ODbL (Open Database License version 1.0), qui est similaire à la licence Creative Commons CC-BY-SA. Sa particularité réside dans l'obligation de redistribuer toute donnée dérivée sous la même licence que celle apposée à la base de données originale.
- la Licence Ouverte Etalab, qui est compatible avec la licence Creative Commons CC-BY. Elle est adaptée pour des données qui sont essentiellement distribuées en France et si le suivi du devenir des données n'est pas recherché. Cette licence est très permissive en ce qu'elle

[4] Décret n° 2017-638 du 27 avril 2017 relatif aux licences de réutilisation à titre gratuit des informations publiques et aux modalités de leur homologation.

permet une totale liberté de réutilisation. En revanche, elle impose le respect d'un certain formalisme notamment la mention de la base d'origine et de la licence appliquée. Cette licence est utilisée à l'Ined, par exemple, pour la base de données « La démographie des décès par Covid ».

À noter que le choix de la licence doit aussi se faire en lien avec le choix de l'entrepôt de données et/ou de la revue où l'on publie le *data paper*. Les entrepôts et les revues peuvent en fait imposer des critères en fonction, par exemple, d'exigences spécifiques.

2.3 Recommandations de l'Ined

Dès lors qu'un jeu de données est diffusable en libre accès, il est conseillé de lui attribuer une licence afin d'éclairer les possibilités de réutilisation, sans avoir à solliciter une autorisation au cas par cas.

Avant de choisir une licence, il est conseillé de :

- Vérifier que les éléments mis à disposition ne contiennent pas de données qui ne peuvent pas être diffusées en libre accès (ex. présence de données personnelles, de données soumises au secret, de droits de propriété intellectuelle appartenant à des tiers, etc.)
- Recenser l'ensemble des chercheurs et institutions partenaires ayant participé à la création des données ou bases de données pour déterminer les participations de chacun et connaître le/les titulaires des droits.
- Si les données mises à disposition ont été produites par un tiers et sont rediffusées dans la base de données (par exemple, dans le cas de données brutes, input des indicateurs calculés), il faudra s'assurer qu'il est possible de les rediffuser et respecter les éventuelles conditions de rediffusion. Dans ce cas, il faudra potentiellement contacter le producteur afin d'obtenir son autorisation pour la rediffusion.

Il est important de bien réfléchir an amont, car les licences ne sont pas révocables. Il est possible de changer de licence, mais toute personne ayant accès à une copie du matériel peut continuer à le redistribuer sous les conditions de la licence précédente.

Le DataLab datalab[at]ined.fr, en collaboration avec le service juridique de l'Ined, peut accompagner les équipes dans le choix de la licence la plus adaptée, selon le cas. Il peut également aider à contacter les producteurs de données dans les cas où il est nécessaire de demander leur autorisation préalable.

Références

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.

[2] BU, Boston University Data Service. Selecting a Data Repository. [En ligne, consulté le 12/07/2023]. Disponible ici : https://www.bu.edu/data/share/selecting-a-data-repository/

[3] IRD, Institut de recherche pour le développement. Entrepôts de données. [En ligne, consulté le 12/07/2023]. Disponible ici : https://data.ird.fr/entrepots-de-donnees-2/

[4] DoRANum. Données de la recherche : apprentissage numérique [En ligne]. France : DoRANum; 2022. Dépôt et entrepôts : fiche synthétique ; [consulté le 12/07/2023]. Disponible : https://doranum.fr/depot-entrepots/depot-et-entrepots-fiche-synthetique_10_13143_a3d4-7553/

[5] IRD, Institut de recherche pour le développement. Cadre juridique. [En ligne, consulté le 12/07/2023]. Disponible ici : https://data.ird.fr/cadre-juridique/

[6] DoRANum. Données de la recherche : apprentissage numérique [En ligne]. France : DoRANum; 2022. Aspects juridiques, éthiques, intégrité scientifique : guide des licences ouvertes [consulté le 12/07/2023]. Disponible : https://doranum.fr/aspects-juridiques-ethiques/guide-des-licences-ouvertes_10_13143_tv6f-sv31

Pour aller plus loin

- Robin, A. (2022). Droit de données de la recherche. Science ouverte, innovation, données publiques. Larcier
- Amiel, P., Frontini, F., Lacour P., et Robin, A. (2020). Pratiques de gestion des données de la recherche : une nécessaire acculturation des chercheurs aux enjeux de la science ouverte ?. Cahiers Droit, Sciences & Technologies [En ligne], 10 | 2020, mis en ligne le 27 avril 2020, consulté le 06 juin 2023. URL : http://journals.openedition.org/cdst/2061 ; DOI : https://doi.org/10.4000/cdst.2061.