Ouvrir ses jeux de données scientifiques en 6 points

1. Qu'est-ce qu'un jeu de données scientifiques ?

Selon l'OCDE, les **données scientifiques (ou *données de la recherche*, *research data***) sont « *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. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche. Ce terme ne s'applique pas aux éléments suivants : carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs, examens par les pairs, communications personnelles avec des collègues et objets matériels (par exemple, les échantillons de laboratoire, les souches bactériennes et les animaux de laboratoire tels que les souris)*. »

Pour vous familiariser avec le concept de données de la recherche, voir la fiche CoopIST : *S'initier en ligne aux données de la recherche et à leur gestion.*

Un jeu de données de recherche (*research dataset*) est l'agrégation d'enregistrements de données organisés pour former un ensemble cohérent. Il est mis en forme de façon à ce qu'il puisse être diffusé et soit compréhensible et réutilisable par les humains et les machines. Pour être facilement découvert et correctement interprété, un jeu de données doit être accompagné de métadonnées (informations sur les données elles-mêmes) : titre du jeu de données, producteurs, lieu de collecte, année, méthodes de collecte et de traitement des données, etc. qui précisent le contexte et les conditions d'obtention du jeu de données.

2. Qu'est-ce que l'ouverture des données (*Open data*) ?

L'ouverture des données (*Open data* en anglais) vise la diffusion gratuite et à tous publics via internet des données d'origine publique ou privée et leur libre réutilisation. Le terme ouvert désigne la liberté d'utiliser, de modifier et de redistribuer les données, même si cette liberté peut être encadrée par l'usage d'une licence de diffusion.

L'ouverture des données de recherche (*Open research data)* s'inscrit dans le mouvement mondial du libre accès à la connaissance (*Open knowledge*) et plus largement de la Science ouverte (*Open science* – voir fiche CoopIST *Découvrir des dictionnaires sur la science ouverte*), qui considère la science comme un bien commun dont la diffusion est d'intérêt général public.

L'ouverture des données de recherche répond ainsi à cinq enjeux :

- accélérer les découvertes scientifiques, les innovations et le retour sur investissement en recherche et développement ;
- encourager la collaboration scientifique et les possibilités de recherche interdisciplinaire ;
- éviter la duplication des expériences, favoriser la réutilisation des données et minimiser le risque de perte des données ;
- assurer l'intégrité et la reproductibilité de la recherche (meilleure qualité des résultats, transparence des méthodologies) ;
- accéder librement à une masse de données ouvrant de nouveaux champs d'analyse non envisagés par le producteur des données (gain de temps et de ressources)

3. Ouvrir ses jeux de données est une décision stratégique

La décision d'ouvrir tout ou partie des données issues de travaux de recherche est stratégique. Elle doit être pensée dès le début du projet et inscrite au fur et à mesure, dans un plan de gestion de données (PGD) (Voir la fiche CoopIST : *Se familiariser avec les plans de gestion de données de la recherche*). Elle implique l'ensemble des membres et partenaires du projet et s'appuie sur des critères scientifiques, juridiques, humains, économiques et techniques, comme :

- l'obligation légale éventuelle d'ouverture des données ;
- l'interdiction légale éventuelle d'ouverture des données, sauf application de traitements spécifiques (Pour la France, voir sur CoopIST : *Avez-vous le droit ou l'obligation de diffuser vos données ?*) ;
- les droits de diffusion et de réutilisation associés aux données primaires, notamment dans le cas de la réutilisation de données existantes ou de partenariats avec la recherche privée (à définir dans le contrat de projet en début de recherche);
- les politiques d'ouvertures des données des partenaires du projet ;
- la demande des bailleurs de fonds du projet ;
- les priorités de réutilisation accordées à certains utilisateurs (mise en place de périodes d'embargo) ;
- la valeur et le potentiel stratégique ou commercial des données ;
- le risque concurrentiel ou la sensibilité des données.

Ces éléments peuvent être pris en compte dans la définition des modalités de partage des données issues de vos travaux de recherche. Des **licences** spécifiques associées aux données à diffuser permettent de fixer les conditions de leur réutilisation (voir section 6.). Le temps et l'effort nécessaires à l'anonymisation et à la mise en forme des données et des métadonnées dans des formats standard adaptés à l'interopérabilité doivent être pris en compte dans le montage et la gestion du projet. Le coût engendré peut être intégré aux besoins de financements du projet.

En France, la Loi pour une République numérique (LRN, 2016) rend obligatoire la publication des données issues d'une activité de recherche financée au moins pour moitié par l'Etat. Ces données devront être librement réutilisables, à l'exception des restrictions liées à d'autres règlementations (données personnelles, etc.). Pour les projets sur financement de l'Union européenne, le programme Horizon Europe (2021 - 2027) rend obligatoire l'accès libre aux données de la recherche selon le

principe « aussi ouvert que possible, aussi fermé que nécessaire » ("*as open as possible* and as *closed as necessary*"). Une loi locale peut également s'appliquer dans le pays ou la zone géographique où vous conduisez vos travaux de recherche : renseignez-vous auprès de vos partenaires sur place ou des juristes de votre établissement.

4. Quelles sont les options pour diffuser ses données de recherche ?

Pour diffuser vos jeux de données scientifiques, vous devez les déposer dans un entrepôt de données de recherche (Research data repository), selon une procédure d'enregistrement des fichiers et des métadonnées associées propre à l'entrepôt. Privilégiez un entrepôt de confiance, disciplinaire ou thématique, c'est-à-dire qui est largement reconnu par la communauté scientifique qu'il dessert et qui possède un processus de curation.

On désigne par curation le processus de mise en forme, relecture et documentation du jeu de données. La curation scientifique (mise en forme des données, nettoyage, traitements appliqués aux données etc.) reste le plus souvent à la charge seule des auteurs du jeu de données. La curation documentaire peut être conjointe avec l'entrepôt. Elle consiste à s'assurer du niveau suffisant de description des données pour leur découverte, leur compréhension et leur réutilisation. L'existence d'un processus de curation sur l'entrepôt signifie qu'une exigence de qualité des données et/ou des métadonnées peut conditionner l'acceptation du dépôt sur l'entrepôt.

Une certification internationale (CoreTrustSeal) peut être attribuée à un entrepôt s'il répond à des critères de qualité et de transparence. Le répertoire international des entrepôts de données de recherche Re3data permet de recherche et d'identifier un entrepôt correspondant à vos données. A défaut d'un entrepôt disciplinaire ou thématique adapté, vous pouvez vous tourner vers l'entrepôt institutionnel de votre établissement ou vers un entrepôt national, s'ils existent. Certains entrepôts généralistes ont une vocation internationale, comme Zenodo (https://zenodo.org/) soutenu par l'Union européenne, mais sont dépourvus de processus de curation.

S'il était autrefois possible de publier les données sous-tendant une publication (underlying data) sous forme de fichiers supplémentaires (supplementary files) associés à l'article (article de recherche, étude de cas, etc.) sur les sites des éditeurs, cette pratique n'est plus recommandée et de moins en moins pratiquée par les éditeurs.

Pour valoriser les données déposées dans un entrepôt, vous pouvez publier un article scientifique, notamment un article de type Data paper qui informe la communauté scientifique de l'existence, de la disponibilité, de la qualité et du potentiel de ces données pour la recherche et l'innovation (voir la fiche CoopIST *Publier un data paper*). Vérifiez les instructions de la revue dans laquelle vous souhaitez publier : elle peut imposer une liste d'entrepôts.

5. Préparer ses données pour leur diffusion

Avant de déposer vos données dans un entrepôt, vous devez vous assurer qu'elles sont correctement mises en forme et documentées :

- **Rassemblez vos données et toutes les informations associées** nécessaires pour permettre leur réutilisation et/ou la reproduction de la recherche. Les données diffusées peuvent être brutes (telles qu'enregistrées par l'instrument) ou dérivées (nettoyées, formatées, organisées). Pour être comprises, elles doivent être accompagnées de documentation (fichier README, dictionnaires de variables, méthodes et protocoles) qui décrivent le contexte de production des données et donnent les clés pour interpréter les fichiers (noms des variables, instrument et unité de mesure utilisés etc.). Le code des scripts d'analyse et de traitement des

données peut également être nécessaire pour interpréter les résultats ou reproduire la méthode.

- **Vérifiez la qualité et la complétude des données :** avant de déposer les fichiers de données dans l'entrepôt, assurez-vous qu'il n'y ait pas de données manquantes ou incorrectement présentées, ni de variables mal libellées. Les fichiers (scripts, fichiers d'archives etc.) doivent également pouvoir être exécutés et les images proposées dans une résolution suffisante pour permettre leur exploitation.
- **Éliminez les éléments non communicables dans vos fichiers de données :** enlevez le contenu sous propriété intellectuelle, effacez les mentions de sites de collecte ou de localisation des espèces en danger ou vulnérables et retirez ou anonymisez les données sensibles ou personnelles permettant d'identifier les individus. Si ces actions ne sont pas possibles, certains entrepôts vous permettront de placer les fichiers qui ne peuvent pas être librement communiqués en accès restreint ou de créer une fiche descriptive des données pour signaler leur existence, sans y joindre de fichiers. Discutez-en avec les membres de votre projet et les services d'appui de vos établissements.
- **Organisez les données et documents :** les noms de fichiers doivent être descriptifs et cohérents, afin que les futurs utilisateurs des données puissent s'y retrouver. Le principe de construction des noms de fichiers peut être explicité dans le fichier readme si nécessaire. Les fichiers peuvent également être organisés de manière logique dans des dossiers (si l'arborescence est gérée par l'entrepôt de données choisi) ou via la construction des noms de fichiers.
- **Convertissez si possible vos fichiers dans des formats ouverts** plutôt que propriétaires, afin de permettre l'interopérabilité et une meilleure accessibilité aux fichiers dans la durée (par exemple, utilisez du .csv plutôt que du .xls, du .txt, .odt ou .pdf plutôt que du .doc etc.)
- **Complétez les métadonnées demandées par l'entrepôt choisi**, en respectant le format demandé. La plupart des entrepôts demandent des informations générales, mais certains entrepôts comme le GBIF peuvent vous imposer de suivre un standard disciplinaire.
- **Utilisez les métadonnées pour faire le lien entre différentes ressources :** les métadonnées permettent de faire le lien entre les jeux de données et les articles de recherche associés ou d'autres jeux de données liés. Elles permettent également d'y associer du code déposé sur une forge logicielle ou un protocole déposé sur un entrepôt spécialisé etc. Renseignez les liens et identifiants pérennes vers ces ressources accroît la visibilité de vos données et leur citation, et de futures collaborations. N'oubliez pas de mentionner le jeu de données dans les publications liées également : de plus en plus d'éditeurs vous demanderont un Data Availability Statement (DAS) pour que les lecteurs sachent comment obtenir l'accès à vos données (voir fiche CoopIST *Distinguer dans un article les jeux de données produits des jeux de données cités*).

6. Choisir une licence de diffusion pour la réutilisation des données

Avant de diffuser un jeu de données, il faut lui apposer une licence fixant les conditions de l'utilisation des données : droits d'utilisation et de modification de la donnée, droits de réutilisation commerciale et non commerciale, obligations éventuelles comme la mention de la source des données ou le partage à l'identique.

Vous n'avez pas toujours le choix du type de licence à applique à vos données :

- assurez-vous que vous possédez tous les droits sur tous les éléments du jeu de données ou de la base de données (données préexistantes, illustrations, …) ou que ceux-ci sont diffusés selon

une licence compatible avec celle que vous souhaitez apposer à votre jeu de données. Dans le cas contraire, il vous est impossible de le diffuser librement : vous devrez alors choisir une licence compatible avec ces restrictions. Si ce n'est pas possible, vous ne pourrez pas diffuser le jeu de données ;

- le bailleur peut stipuler le type de licence souhaité pour la diffusion des données, hors règlementations spécifiques ;
- lorsque les données sont liées à un article scientifique, la licence de diffusion choisie pour les données doit répondre aux exigences de la revue (toujours en respectant la réglementation en vigueur). Consultez les instructions aux auteurs ;
- vérifiez les exigences de l'entrepôt dans lequel vous souhaitez déposer votre jeu de données : il peut imposer une licence de diffusion particulière.

En cas de doute sur la licence à appliquer, prenez conseil auprès des juristes de votre institution.

Si vous pouvez choisir la licence à appliquer à vos données, privilégiez une licence largement utilisée et compatible avec les autres licences existantes, afin de faciliter la compilation de vos données avec d'autres données mises à disposition sous d'autres licences. Les licences les plus largement utilisées pour la diffusion des données sont les licences Creative Commons, mais il en existe d'autres (Voir la fiche Coopist *Connaître et utiliser les licences Creative Commons*).

Liens utiles

Ministère de l'enseignement supérieur et de la recherche. 2021. La feuille de route 2021-2024 du MESRI sur la politique des données, des algorithmes et des codes sources. France : MESR.

Ministère de l'enseignement supérieur et de la recherche. 2021. Deuxième plan national pour la science ouverte (2021-2024). France, MESR.

Open Knowledge. Qu'est-ce que l'Open Data ? In : Open Data Handbook.

Cécile Arènes, Lionel Maurel, Stephanie Rennes. Guide d'application de la Loi pour une République numérique pour les données de la recherche. Comité pour la science ouverte. 2022. ⟨hal-03968218⟩

Organisation de coopération et de développement économique. 2007. Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics. Paris : OCDE, 27 p.

Union européenne. Directive (UE) 2019/1024 du Parlement européen et du Conseil du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public (refonte). PE/28/2019/REV/1

Céline Barthélemy (OrcID : 0000-0002-4676-495X)

Délégation à l'information scientifique et à la science ouverte, Cirad, juillet 2024

Mise à jour de la version du 15 avril 2015 : Dedieu L. ; Fily M.F. 2015. Rendre publics ses jeux de données scientifiques, en 6 points. Montpellier (FRA) : CIRAD, 6 p.

Comment citer ce document :

Barthélemy C. Ouvrir ses jeux de données scientifiques, en 6 points. 2024 . Montpellier (FRA) : CIRAD, 6 p. https://doi.org/10.18167/coopist/0059