Rendre publics ses jeux de données scientifiques en 6 points

1. Qu'est-ce qu'une donnée scientifique, un jeu de données, une base de données ?

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, consultez la fiche CoopIST : *S'initier en ligne aux données de la recherche et à leur gestion* .

Un jeu de données scientifiques (*data set*) est l'agrégation d'enregistrements de données organisés pour former un ensemble cohérent. Les jeux de données numériques sont formatés de telle sorte qu'ils soient communicables, interprétables et adaptés à un traitement informatisé.

Pour être utilisé et cité, un jeu de données doit être accompagné de métadonnées descriptives : titre, producteur, ….

Certains jeux de données, annotés, revus par les pairs et mis à disposition deviennent des données de référence, par exemple les données génomiques disponibles dans la base de données GenBank.

Une base de données numérique (*database*) est constituée par un ensemble de jeux de données organisés et structurés pour être accessibles et exploitables au moyen d'un programme informatique.

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

L'ouverture des données (*Open data*) a pour objectif la diffusion libre, gratuite et universelle, via internet, des données d'origine publique ou privée. Le terme *ouvert* est défini comme la liberté d'utiliser, de modifier et de redistribuer librement les données.

L'*Open data* s'inscrit dans le mouvement mondial du libre accès à la connaissance (*Open knowledge)* et plus largement de la Science ouverte (*Open science*), qui considère la science comme un bien commun dont la diffusion est d'intérêt public et général.

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

- accélerer 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. Rendre publics vos jeux de données est une décision stratégique

La question de l'ouverture des données de recherche relève de la **stratégie de recherche** d'une institution ou d'un projet scientifique. Elle se concrétise notamment par l'élaboration, lors du démarrage du projet de recherche, d'un plan de gestion des données (Voir la **fiche CoopIST :** *Découvrir les plans de gestion de données de la recherche* **).**

La décision de rendre publique tout ou partie des données de recherche d'un projet est stratégique. Elle implique l'ensemble des acteurs 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 ;
- les politiques d'ouvertures des données des partenaires du projet ;
- les droits des propriétaires des données primaires, notamment dans le cas des partenariats avec la recherche privée ;
- 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 ;
- la nécessité d'anonymisation pour la protection des données personnelles ou confidentielles ;
- le temps et l'effort nécessaires à la mise en forme des données et des métadonnées dans des formats standards adaptés à l'interopérabilité.

L'ouverture des données implique aussi la définition des modalités de leur partage. Des **licences de diffusion** spécifiques permettent de fixer les conditions de leur réutilisation.

4. Quelles sont les options pour rendre publics vos jeux de données ?

Pour rendre publics vos jeux de données, vous devez les déposer sur un site internet où ils seront accessibles. Privilégiez les options qui offriront la plus grande visibilité à vos jeux de données :

- les déposer dans un entrepôt de données (*Data repository*) certifié et accessible à tous, selon une procédure précise d'enregistrement de fichiers et de métadonnées décrivant les fichiers et les données ;

- les publier sous forme de fichiers supplémentaires associés à un article (article de recherche, étude de cas, etc.). Il s'agit alors uniquement des données qui sous-tendent la publication (*underlying data*) ;

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 Rédiger et publier un data paper dans une revue scientifique). Vérifiez les instructions de la revue dans laquelle vous souhaitez publier : elle peut imposer, ou non, le dépôt dans un entrepôt spécifique.

5. Une licence de diffusion est indispensable pour rendre public un jeu de données

Avant de rendre public un jeu de données, il faut lui apposer une licence de diffusion fixant les conditions de son utilisation : 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 qui va s'appliquer à vos données :

- **vérifiez les exigences de l'entrepôt** dans lequel vous souhaitez déposer votre jeu de données ; **il peut imposer une licence de diffusion particulière** ;
- Lorsque les données sont liées à un article scientifique, la licence de diffusion choisie pour les données doit répondre aux exigences de la revue. **Consultez les instructions aux auteurs** ;
- **Le type de licence préconisé est donc un critère de sélection important pour choisir un entrepôt ou un éditeur ;**

Voici quelques conseils supplémentaires :

- 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 ;
- assurez-vous que vous possédez tous les droits sur tous les éléments du jeu de données ou de la base de données (illustrations, …), sinon il vous est impossible de le rendre public ;
- prenez conseil auprès des juristes de votre institution.

6. Les principales licences de diffusion des jeux de données

Pour favoriser la réutilisation des jeux de données que vous rendez publics, privilégiez des licences largement utilisées.

Les principales licences utilisées pour la publication des jeux de données sont les suivantes :

Les licences Creative Commons

Les licences *Creative Commons* ont été créées en 2002 pour la diffusion de contenus numériques comme le texte, les images et les films. Elles permettent aux auteurs d'indiquer facilement les droits qu'ils veulent conserver et les droits auxquels ils renoncent afin de permettre à d'autres de réutiliser leur œuvre. Elles permettent de combiner quatre clauses, à associer selon ses besoins :

- *Attribution* (sigle **BY**) : paternité, c'est-à-dire citation de l'auteur initial (obligatoire).

- *Non Commercial* (sigle **NC**) : interdiction de tirer un profit commercial de l'œuvre sans autorisation de l'auteur.
- *No derivative works* (sigle **ND**) : impossibilité d'intégrer tout ou partie dans une œuvre composite.
- *Share alike* (sigle **SA**) : partage de l'œuvre, avec obligation de rediffuser selon la même licence ou une licence similaire.

Depuis la version 4.0 (novembre 2013), les licences *Creative Commons* prennent en compte le droit *sui generis* des bases de données. Il est donc désormais possible de les utiliser pour publier des jeux de données et de choisir d'attribuer une même licence *Creative Commons* pour l'article et les jeux de données liés, ou une licence pour l'article et une autre pour les données.

Toutefois, *Creative Commons* recommande de ne pas utiliser les clauses NC (pas d'utilisation commerciale) ou ND (pas de modification) pour des jeux de données ou des bases de données destinés à une utilisation scientifique car elles restreignent les possibilités de réutilisation.

De même, il est conseillé de ne pas utiliser la clause SA (partage à l'identique) car elle impose la licence de rediffusion et réduit donc l'interopérabilité des données.

La licence CC-by 4.0 (*Creative Commons Attribution*)

https://creativecommons.org/licenses/by/4.0/

La licence CC-by 4.0 permet de partager, copier, distribuer et communiquer les données par tous moyens et sous tous formats, de les réutiliser pour créer de nouveaux jeux de données. Toutes les utilisations, y compris commerciales, sont possibles, sous réserve de créditer les données à leurs créateurs (obligation d'attribution).

Cette licence est préconisée par un certain nombre d'entrepôts de données.

La licence CC0 (*Creative Commons Public Domain Dedication*)

http://creativecommons.org/publicdomain/zero/1.0/deed.fr

La licence CC0 a été créée en 2009 pour faciliter la réutilisation des jeux de données. Elle permet aux producteurs de données de les placer dans le domaine public, sans aucune restriction de réutilisation.

La citation du producteur du jeu de données n'est pas obligatoire, même si, d'un point de vue éthique et scientifique, il est conseillé aux utilisateurs de citer les créateurs originels des données lors de la réutilisation. Cela permet de certifier leur origine et la méthodologie associée à leur production.

La licence CC0 est imposée par quelques entrepôts de données, comme l'entrepôt pluridisciplinaire Dryad. Elle est aussi imposée par certains éditeurs de revues scientifiques, comme BioMed Central ou Nature Publishing Group. En conséquence, si vous publiez un article chez ces éditeurs en lien avec des jeux de données, vous devez les déposer dans un entrepôt sous licence CC0.

Les licences de l'Open Knowledge Fondation (OKF)

Résultats du projet *Open data commons*, lancé en 2007 au Royaume-Uni par l'*Open knowledge Foundation*, ces licences, basées sur le droit anglo-saxon et orientées bases de données, peuvent être appliquées aux bases de données et aux données qu'elles contiennent prises isolément.
Ces licences autorisent a minima la copie, l'utilisation, la redistribution, la modification des données, la réalisation de travaux dérivés de la base de données.

ODC-by (*Open Database Commons*)

http://opendatacommons.org/licenses/by/

la licence ODC-by impose d'indiquer le nom de l'auteur/créateur de la base de données originale (obligation d'attribution).

Cette licence se rapproche de la licence CC-by et est utilisée par de nombreux éditeurs scientifiques dont Pensoft.

ODC-ODBL (*Open database License*)

http://opendatacommons.org/licenses/odbl/

La licence ODC-ODBL impose d'indiquer le nom de l'auteur/créateur de la base de données originale et de la redistribuer sous les mêmes conditions (obligations d'attribution et de partage à l'identique).

Très utilisée en France, cette licence se rapproche de la licence CC-by-sa.

PDDL (*Public domain dedication and license*)

http://opendatacommons.org/licenses/pddl/

La licence PDDL est une licence libre de tout droit, de type domaine public, par laquelle l'auteur ou le créateur abandonne son droit d'auteur moral.

Cette licence se rapproche de la licence CC0.

La Licence ouverte (LO)

https://www.etalab.gouv.fr/licence-ouverte-open-licence

Créée par EtatLab pour la diffusion des données publiques françaises, la Licence ouverte autorise la réutilisation, la reproduction, la modification, la redistribution des données et leur exploitation à titre commercial sous réserve de mentionner a minima le nom du producteur et la date de dernière mise à jour.

Elle est compatible avec toute licence qui exige a minima la mention de paternité, notamment avec les licences ODC-by et CC-by.

Cette licence est utilisée notamment sur la plate-forme de mise à disposition des données publiques data.gouv.fr.

Liens utiles

Ball, A. 2012. How to License Research Data. Edinburgh : Digital Curation Centre.
http://www.dcc.ac.uk/resources/how-guides/license-research-data

Ducourneau J. 2015. L'open data : fiche synthétique
http://www.cndp.fr/savoirscdi/societe-de-linformation/tic-et-documentation/veille-technologique/lopen-data-fiche-synthetique.html

OCDE. 2007. Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics. Paris : OCDE, 27 p.
http://www.oecd.org/fr/science/sci-tech/38500823.pdf

Laurence Dedieu, Marie-Françoise Fily
Délégation à l'information scientifique et technique, Cirad
15 avril 2015

Informations

Comment citer ce document :
Dedieu L., Fily M.F. 2015. Rendre publics ses jeux de données scientifiques en 6 points. Montpellier (FRA) : CIRAD, 6 p. http://url.cirad.fr/ist/rendre-publics-ses-donnees