Déposer des données de recherche dans un entrepôt en 6 points

1. Qu'est-ce qu'un entrepôt de données de recherche

Un entrepôt de données de recherche (Research Data Repository ou Data Repository) est une base de données destinée à accueillir, conserver, rendre visibles et accessibles des données de recherche. Son rôle est de permettre le dépôt ou la collecte de données, leur description, leur accès, et leur partage en vue de leur réutilisation. Chaque entrepôt dispose généralement d'une politique de dépôt, de description et de diffusion des données.

Ces infrastructures s'inscrivent dans une démarche de partage et d'ouverture des données selon les principes FAIR pour que les données soient "Facile à trouver, Accessible, Interopérable et Réutilisable" (en anglais : Findable, Accessible, Interoperable, Reusable).

Le dépôt d'un jeu de données dans un entrepôt s'accompagne généralement de la saisie ou de la collecte d'informations (les métadonnées) sur les données déposées facilitant la compréhension et l'interprétation de ces données (par exemple, les couvertures géographique, temporelle, etc.). Outre des métadonnées standard comme celles du format Dublin Core permettant de décrire l'auteur, le titre, l'année de création, … d'un jeu de données, un entrepôt propose généralement un ensemble de métadonnées spécifiques aux sujet, thème, discipline (par exemple données biologiques, astronomiques, environnementales, etc.) des données qu'il accueille.

2. Différentes catégories d'entrepots de données

Il existe plusieurs catégories d'entrepôts de données : institutionnels, pluridisciplinaires, disciplinaires ou thématiques.

Un entrepôt de données peut, en effet, être mis en place et géré au niveau d'un établissement de recherche, on parle alors d'entrepôt institutionnel. Il peut également être créé et géré par une entité publique, associative, ou privée, cas d'entrepôt d'éditeurs destiné à accueillir les jeux de données sous-tendant les résultats présentés dans les articles scientifiques que l'éditeur publie.

Exemples d'entrepôts de données :

- Institutionnels (en France) : Dataverse Cirad, Datapartage (INRAe), DataSuds (IRD)
- Pluridisciplinaires et internationaux : Zenodo, Dryad, Figshare
- Thématiques ou disciplinaires : GenBank (séquences génétiques), TRY (caractères botaniques), GBIF (biodiversité), Pangaea (sciences de la terre et de l'environnement), WormBase

(nématologie), Movebank (mobilité animale), West African Vegetation, DataFirst (enquêtes socio-économiques en Afrique), Protocols.io (protocoles), etc.

- Liés à un éditeur : GigaDB (Oxford Univ. Press), Dataverse Ubiquity Press, Dataverse *Economics*

3. Pourquoi déposer des données dans un entrepôt

Déposer ses données dans un entrepôt assure leur préservation, leur visibilité et leur accès, facilitant ainsi leur partage et leur réutilisation. En outre, le crédit reçu en tant qu'auteur ou contributeur crédité des données diffusées, permet d'accroître sa notoriété.

Déposer ses données dans un entrepôt apporte ainsi de nombreux avantages :

- conservation des données dans un environnement sécurisé
- visibilité des données et accès facilité pour les moteurs de recherche
- interopérabilité des données grâce à l'utilisation de standards de métadonnées
- découverte, réutilisation et citation du jeu de données facilitées par son identifiant pérenne
- gestion des modalités de partage des données par l'attribution de licences de diffusion
- respect des recommandations des financeurs et institutions sur l'ouverture des données
- reproductibilité de la recherche, intégrité et validation scientifique améliorées
- valorisation des données par leur réutilisation dans de nouvelles études et innovations.

4. Les questions à se poser avant de déposer des données dans un entrepôt

Le dépôt de données dans un entrepôt a pour objectif de partager des données, c'est-à-dire de les rendre accessibles pour qu'elles puissent être réutilisées par d'autres, que ce soit des scientifiques, des entreprises, des décideurs, ou des citoyens.

La décision de rendre publiques des données de recherche s'appuie sur des critères scientifiques, réglementaires, juridiques, humains, économiques et techniques et implique l'ensemble des contributeurs et partenaires d'un projet.

Les questions à se poser sont notamment :

- Quelles sont les obligations d'ouverture des données qui s'appliquent ? L'obligation peut être imposée par le financeur du projet, par une loi nationale, européenne ou internationale, par la politique des données de certains partenaires, par la revue dans laquelle vous publiez, etc.
- Quelle est la valeur scientifique des données et leur potentiel de réutilisation ? L'intérêt et l'utilité actuelle ou future, scientifique, environnementale, économique, ou sociale, des données peuvent guider le choix. La question du potentiel stratégique ou commercial des données peut aussi influencer la décision. Les agences de financement appliquent la recommandation « *Aussi ouvert que possible, aussi fermé que nécessaire* » pour la diffusion des données produites dans le cadre d'un projet financé.
- Avez-vous le droit de rendre publiques ces données ? En d'autres termes, avez-vous respecté :
  - les droits de propriété intellectuelle ?

    Ex : données obtenues en partenariat ou contenant des images protégées par le droit d'auteur
  - les obligations contractuelles ?

    Ex : utilisation de données préexistantes, issues d'un projet précédent ou téléchargées à partir d'un entrepôt (ex *: FAOSTAT, GBIF, Centre for Ecology & Hydrology*), éventuellement protégées par des droits spécifiques ou des licences
  - les réglementations éthiques ?

Ex : données personnelles collectées lors d'enquêtes et qui doivent être supprimées (anonymisation) pour respecter les droits des personnes

Ou données issues de ressources génétiques ou de savoirs traditionnels associés qui nécessitent de respecter la Réglementation APA sur l'accès et le partage des avantages

Ou données qui soulèvent des questions éthiques (ex : expérimentation animale, essais cliniques chez l'homme, recherches ayant un impact sur l'environnement, etc.) et requièrent la validation par un comité d'éthique

- Avez-vous obtenu l'accord de tous les contributeurs ?
- Avez-vous évalué le temps et l'effort nécessaires à la mise en forme des données et des métadonnées pour répondre aux exigences de l'entrepôt ? Si votre projet a fait l'objet d'un Plan de gestion des données, alors vos données sont quasi prêtes et leur dépôt en sera facilité.
- Avez-vous défini les conditions de réutilisation des données que vous avez produites ? Le mouvement « Open data » incite à « ouvrir » les données à tous, sans restriction aucune. En tant que chercheur, vous avez tout intérêt à imposer l'obligation de citer les créateurs des données lorsque celles-ci seront réutilisées. Ceci est possible par le choix de l'entrepôt et de la licence de diffusion que vous attribuerez à vos données.

5. Comment choisir un entrepôt de données

Le choix d'un entrepôt de données sera guidé par les pratiques de votre communauté scientifique, par la politique de votre établissement et la disponibilité d'un entrepôt institutionnel ou par les instructions aux auteurs de la revue dans laquelle vous soumettez votre article.

Plusieurs répertoires d'entrepôts de données permettent d'affiner la sélection en appliquant différents critères.

- Re3Data : répertoire d'entrepôts de données créé en 2012 par le consortium international DataCite qui œuvre pour le partage des données de recherche. Re3data renseigne sur plus de 2450 entrepôts et sur les licences proposées par chacun.
- Repository Finder : outil permettant de trouver un entrepôt de données. Cet outil, hébergé par DataCite, cherche des entrepôts dans le répertoire Re3data.
- Fairsharing : guide sur les normes, standards, entrepôts et recommandations en termes de données et métadonnées.
- Cat OPIDoR : catalogue des services français dédiés aux données scientifiques, hébergé par le CNRS. Cat OPIDoR présente une liste de 58 entrepôts de données en France.

L'identification d'entrepôts adaptés à vos données peut être facilitée en interrogeant des bases de données comme DataCite (DataCite Search) dont l'accès est gratuit (consortium international et agence d'enregistrement de DOI attribués aux jeux de données), Data Citation Index pour les institutions qui y sont abonnées (Clarivate Analytics qui produit aussi le Web of science) ou d'autres moteurs de recherche (voir la fiche CoopIST : Citer un jeu de données scientifiques).

Pour sélectionner votre entrepôt, vérifiez qu'il répond aux critères suivants :

- Adapté au type de données que vous allez déposer
- Répondant aux recommandations du bailleur de votre projet, de votre institution, ou de la revue dans laquelle vous publiez
- Reconnu dans votre discipline et par la communauté scientifique (certains entrepôts, encore peu nombreux, sont certifiés)
- Attribuant un identifiant numérique pérenne, univoque à chaque jeu de données
- Assurant la conservation des données, c'est-à-dire leur pérennité

- Gratuit (la plupart des entrepôts) ou pratiquant des coûts de dépôt de données acceptables
- Proposant les modalités d'accès aux données, adaptés à vos besoins : accès libre, après enregistrement, accès restreint, sur demande, différé par un embargo
- Attribuant la licence de diffusion des données adaptée à vos exigences Attention : 1) certains entrepôts imposent une licence alors que d'autres proposent un choix de licences et 2) lorsque les données sont liées à un article scientifique, la licence de diffusion appliquée aux données doit aussi répondre aux exigences de la revue (consultez les instructions aux auteurs).

6. Déposer ses données en pratique: exemple du Dataverse

Dataverse est un logiciel open source de création et de gestion d'entrepôts de données. Il a été développé par l'université Harvard, qui met en réseau les entrepôts Dataverse du monde entier avec son propre entrepôt Harvard Dataverse, généraliste et ouvert à tous.

Le dépôt de données comporte plusieurs étapes.

- L'identification du déposant via son compte personnel (créé à la première connexion)
- Le téléchargement du ou des fichiers, dans un format ouvert pour en assurer la lisibilité dans le temps
- La description complète des données, en renseignant des formulaires correspondant à différents standards de métadonnées :
    - Métadonnées de citation : titre, auteurs, description synthétique, discipline(s) scientifique(s), thématique(s), mots-clés, publications liées, producteur des données, contributeurs, financeur(s) du projet, type(s) de données, etc.
    - Métadonnées géospatiales, pour situer le lieu de recueil des données (si pertinent)
    - Métadonnées disciplinaires, pour décrire spécifiquement les données à l'attention de sa communauté scientifique (sciences humaines et sociales, sciences de la vie, etc.)
- Le choix des conditions de réutilisation des données, le plus souvent via l'attribution d'une licence. Dataverse propose par défaut la licence CC0-Domaine public, mais le déposant peut choisir d'affecter toute autre licence (Etalab, Creative Commons, etc.) à ses données.

L'accès aux fichiers peut être :

- Ouvert à tous
- Fermé temporairement, par exemple jusqu'à la parution de l'article tiré des résultats du projet
- Soumis à une demande d'accès, faite par l'internaute directement dans l'entrepôt et qui est transmise automatiquement à l'adresse de messagerie du déposant.

Une fois que le déposant a déposé les fichiers, qu'il a décrit les données avec les métadonnées, qu'il a défini le mode d'accès et les conditions de réutilisation, il rend publiques ses données. Elles deviennent alors visibles par tous dans l'entrepôt Dataverse concerné, avec un DOI (identifiant unique et pérenne), et peuvent être citées.

Le Dataverse du Cirad (https://dataverse.cirad.fr) permet ainsi aux chercheurs de l'institution de préserver, diffuser et valoriser les données de recherche qu'ils produisent ou coproduisent avec leurs partenaires du Nord et du Sud.

Dedieu Laurence ; Barale Martine (0000-0000-0001-5971-8402)
Délégation à l'information scientifique et technique, Cirad
Mai 2020

Information

Comment citer ce document :

Dedieu, L. Barale, M. 2020. Déposer des données dans un entrepôt, en 6 points. Montpellier (FRA) : CIRAD, 4 p.

https://doi.org/10.18167/coopist/0070