couperin.org
Consortium unifié des établissements universitaires et de recherche pour l'accès aux publications numériques

DECRYPTAGE

Les entrepôts de données d'éditeurs commerciaux : quelle stratégie adopter ?

Décembre 2024

Face à la multiplication des sollicitations des éditeurs auprès de la communauté scientifique pour promouvoir les entrepôts de données privés, il est nécessaire de connaître ces outils pour accompagner au mieux les chercheurs.

1. Quels entrepôts d'éditeurs commerciaux existent ?

Si la plupart des éditeurs scientifiques se contentent d'inciter leurs auteurs à déposer leurs données dans des entrepôts thématiques ou, à défaut, généralistes ou institutionnels, des éditeurs scientifiques mondiaux, comme Elsevier et Springer Nature, proposent leur propre entrepôt de données. Cependant, leur utilisation n'est pas obligatoire pour publier dans leurs différentes revues. Tour d'horizon de ces solutions :

Nom de l'entrepôt	URL	Description	Propriétaire	Coût	Remarques
Figshare	https://figshare.com/	Entrepôt généraliste permettant le dépôt et le partage de divers types de données de recherche.	Figshare est la propriété de *Digital Science*, une filiale du groupe allemand Holtzbrinck Publishing Group (actionnaire majoritaire de *Springer Nature*).	20 Go de stockage gratuit par chaque compte avec une limite de 5 Go par fichier. La tarification est basée sur le stockage jusqu'à 100 Go, puis par tranches de 250 Go (875 $ par tranche de 250 Go). Par exemple, pour un stockage jusqu'à 1 To, il faut compter 4 x 875 $ = 3 500 $.	Option Figshare+ Les données sont stockées sur Amazon Web Services (AWS) S3, mais la localisation des serveurs n'est pas spécifiée. Pour la diffusion des *supplementary materials* ; des partenariats sont établis avec ACS, Taylor & Francis, The Royal Society. Plus d'infos sur les autres éditeurs ici.

Mendeley Data	https://data.mendeley.com/	Entrepôt accessible à l'ensemble des chercheurs qui n'auraient pas d'entrepôts institutionnels ou d'entrepôts thématiques à disposition.	Elsevier	Gratuit (mais la taille des jeux de données est limitée à 10 Go). Les institutions ont la possibilité de prendre un abonnement (la taille limite des jeux de données augmente et passe à 100 Go).	Les données publiées peuvent être supprimées par leurs auteurs. Leur pérennité n'est donc pas assurée. Les données sont stockées sur Amazon Web Service S3 (basé en Irlande) + archivage par le Data Archiving and Networked Service basé au Pays-Bas et dépendant de la Dutch Academy (KNAW) et du Conseil national de la recherche Néerlandais (NWO).
Digital Commons Data	https://elsevier.digitalcommonsdata.com/research-data/ https://www.elsevier.com/fr-fr/products/digital-commons/data	Elsevier propose une solution pour les institutions qui souhaiteraient mettre en place leur propre entrepôt de données.	Elsevier	Prix non public.	Les données peuvent être stockées à deux endroits différents. Soit sur les serveurs de l'institution ; soit sur les serveurs d'Amazon (AWS S3) avec un archivage par le DANS (comme pour Mendeley Data).
IEEE DataPort	https://ieee-dataport.org/submit-dataset	Entrepôt proposé par défaut en cas de soumission dans certaines revues d'IEEE	IEEE	1950 $ pour un accès ouvert aux données déposées.	En cas de soumission d'article, le déposant se voit proposer d'ouvrir ses données. S'il répond "oui", le DataPort lui est alors automatiquement proposé.

2. Quels sont les risques encourus lors du dépôt de données dans un entrepôt d'éditeur commercial ?

Plusieurs risques peuvent être encourus par les chercheurs qui déposent leurs données dans un entrepôt commercial. Les principaux risques identifiés, sans ordre de priorité, sont les suivants :

- La pérennité des données déposées n'est pas assurée ;
- Cela augmente notre dépendance vis-vis des grands éditeurs privés et limite ainsi notre capacité future de négociation de services plus ciblés ;

- Le devenir des données déposées dans ce type d'entrepôt pour un article qui n'est finalement pas accepté n'est pas clair ;
- La titularité des droits sur les métadonnées n'est pas toujours clairement établie ;
- Certains entrepôts ne garantissent pas le libre accès et la libre réutilisation des données ;
- Les modalités de dépôt et d'accès à la notice du jeu de données dépendant des éditeurs, les possibilités de modifier la notice, d'ajouter de nouvelles versions du jeu de données pourraient être restreintes ;
- Les données peuvent être rendues non interopérables en l'absence d'un schéma de métadonnées standard.

Le risque majeur réside dans l'appropriation de ces données par les éditeurs. Voici par exemple ce qui est indiqué dans les termes d'utilisation de Mendeley (Elsevier), consultés en octobre 2024 :

For Research Data that you make publicly available on the Site, you grant us a perpetual, worldwide, non-exclusive right and license to publish, extract, reformat, adapt, build upon, index, re-distribute, link to and otherwise use all or any part of the Research Data in all forms and media (whether now known or later developed), and to permit others to do so. Where you have selected a specific end user license under which the Research Data is to be made available on the Site, we shall apply that end user license.

En résumé, bien que ce droit ne soit pas exclusif, tout jeu de données déposé dans un outil d'Elsevier permet à ce dernier d'en acquérir l'entière propriété, alors qu'un jeu de données déposé ailleurs permet de choisir une licence de réutilisation qu'Elsevier est alors tenu de respecter.

S'ajoutent à la liste ci-dessus les critères d'exclusion définis par le Collège Données du Comité pour la science ouverte (voir partie 4) :

1. Absence de modération des dépôts ;
2. Absence d'identifiant pérenne ;
3. Absence de garantie sur la pérennité de l'infrastructure ;
4. Politique tarifaire excessive ;
5. Localisation des données hors Union européenne pour certains types de données.

3. Les chercheurs sont-ils obligés de déposer dans l'entrepôt adossé à la revue dans laquelle ils publient ?

Non, comme le précise le *Guide d'application de la Loi pour une république numérique*[1] : « *les éditeurs d'écrits scientifiques ne peuvent valablement obtenir de cessions exclusives de droits sur des données de recherche liées à la publication, qu'elles soient déposées sur un entrepôt de données ou qu'elles figurent comme* supplementary materials *de l'article* ». Si un chercheur se voit malgré tout imposer le dépôt dans un entrepôt d'éditeur, le service d'accompagnement peut se rapprocher de son ADAC (administrateur des données, algorithmes et codes) ou de son service juridique pour trouver la solution la plus adaptée. A ce jour, il n'y a pas encore de préconisation nationale sur le sujet.

L'éditeur peut en revanche imposer la mise à disposition du jeu de données associé à l'article publié, à des fins de transparence et d'intégrité scientifique. Il peut même l'imposer dès la soumission de l'article, afin que les relecteurs puissent avoir accès aux données qui justifient les conclusions de l'article.

[1] 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⟩. Le texte stipule également à la même page : "*Lorsque des éditeurs rassemblent des données de recherche liées à des publications dans des entrepôts, ils ne peuvent faire jouer leur droit de producteur de bases de données pour limiter la réutilisation*" (p. 15).

Ex. chez MPDI :

> ***Data availability statements***
>
> *Data availability statements are required for all articles published with MDPI. During the peer review and editorial decision process, authors can be asked to share existing datasets or raw data that have been analyzed in the manuscript, and whether they will be made available to other researchers following publication. Authors will also be asked for the details of any existing datasets that have been analyzed in the manuscript.*

Ex. chez ACS (pour certaines revues) :

> *The journal* ***requires****, as a condition of publication, all authors to publicly share all the data underlying the results reported in the paper, preferably via archiving in an appropriate public repository. The data will undergo peer review along with the manuscript. Authors are* ***required*** *to provide a Data Availability Statement describing the public availability of the data supporting the article's conclusions.*

L'éditeur peut demander une mise à disposition via une rubrique spécifique liée à la publication ("*supplementary data*" ou "*supplementary materials*") ou via un entrepôt externe. **Il ne peut pas imposer une modalité spécifique pour la diffusion des données** (*supplementary data*, entrepôt commercial…).

Il est donc tout à fait possible et recommandé de choisir un entrepôt adapté, thématique ou institutionnel, et de fournir à l'éditeur le DOI qui mène au jeu de données, dès la soumission.

Recherche Data Gouv permet ainsi la mise à disposition d'une URL privée pour les relecteurs, pour la phase de *peer reviewing*.

Il est essentiel, au moment du dépôt des jeux de données, de choisir une licence, de préférence CC-BY, afin que l'éditeur ne puisse pas s'approprier de droits sur les données.

4. Quel(s) entrepôt(s) faut-il privilégier ?

Afin de publier des jeux de données, qu'ils soient en accès ouvert, en accès restreint ou avec un embargo, il est donc recommandé de les déposer dans un entrepôt de données de confiance.

La liste des entrepôts thématiques de confiance établie par le Collège Données du Comité pour la science ouverte **exclut clairement les entrepôts pratiquant la cession des droits** : "*Les pratiques de certains éditeurs en matière de propriété intellectuelle ne permettent pas de garantir le libre accès et la libre réutilisation des données qui seraient déposées dans les entrepôts qu'ils développent et recommandent. C'est par exemple le cas d'ACS en chimie, qui propose le dépôt de données de résonance magnétique nucléaire sous forme de fichiers FID au sein du « research data center » sans que la politique en matière de licences ne soit explicitée. Les entrepôts pratiquant la cession de droits sont donc exclus. Cette position est cohérente avec le guide « Partager les données liées aux publications scientifiques » du Collège données de la recherche du Comité pour la science ouverte (2022), qui préconise de ne pas « rendre les utilisateurs captifs au sein d'environnements maîtrisés par de grands acteurs commerciaux de l'édition scientifique »*"[2].

[2] Frédéric de Lamotte, Véronique Stoll, Cécile Arènes, Marie-Emilia Herbet, Stéphane Debard, et al. *Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante*. Comité pour la Science Ouverte. 2024. ⟨hal-04534321⟩, p.12-13.

Dès lors, vers quel entrepôt faut-il se tourner ? Le logigramme dynamique du portail Recherche Data Gouv indique la marche à suivre :

- Quand il existe, privilégier un entrepôt thématique de confiance dans sa discipline parmi cette liste, qui évolue et s'enrichit dans le temps : https://recherche.data.gouv.fr/fr/entrepots
- Si aucun entrepôt n'existe dans la discipline, le dépôt dans l'entrepôt Recherche Data Gouv est conseillé.

5. Quelles sont les obligations des chercheurs en matière d'ouverture des données ?

Dans certains cas, l'ouverture des données est obligatoire :

- Si cela fait partie de la politique éditoriale de la revue : le dépôt des données dans un entrepôt peut être parfois nécessaire pour passer l'étape du *peer reviewing* ;
- Si les données présentent un intérêt économique, social, sanitaire ou environnemental particulier (Article L312-1 du Code des relations entre le public et l'administration) ;
- S'il s'agit de données géographiques concernées par la directive INSPIRE.

La plupart du temps, les politiques publiques encouragent la diffusion des données selon le principe "aussi ouvert que possible, aussi fermé que nécessaire".

La loi pour une République numérique 2016, art. 30 stipule ainsi que la réutilisation des données est libre : « *dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat* (…) *ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche. L'éditeur d'un écrit scientifique mentionné ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication* ».

Dans le cadre de projets financés sur fonds publics (ANR et Horizon Europe), l'ouverture est fortement encouragée quand cela est possible, en particulier quand elles sont uniques et irremplaçables ou reproductibles à des coûts dissuasifs, dans le cadre suivant :

- Les exceptions à l'ouverture totale ou partielle des données pour des raisons légales sont à justifier dans le plan de gestion de données ;
- L'utilisation de licences en CC-BY ou CC-0 ou équivalentes pour les données ouvertes est encouragé ;
- Pour les projets financés, les coûts associés à la gestion des données sont éligibles au remboursement.

Auteurs : Meriç Akdogan, Laetitia Bracco, Delphine Du Pasquier, Gaëlle Gauvrit, Cyril Heude, Benjamin Laillier, Alicia León y Barella, Céline Rousselot et Jozefina Sadowska pour le GTSO Données de Couperin.