Ouvrir la science !

Sélectionner un entrepôt thématique de confiance pour la diffusion des données de recherche : note méthodologique

Collège Données de la recherche

Novembre 2023

Sélectionner un entrepôt thématique de confiance pour la diffusion des données de recherche : note méthodologique

———

Comité pour la science ouverte – Collège Données de la recherche

———

Pierre-Yves ARNOULD – Pilotage du Collège Données de la recherche
CNRS, OSU OTELo

Véronique STOLL – Pilotage du Collège Données de la recherche
Observatoire de Paris - PSL

Cécile ARENES – Pilotage du groupe de travail
Sorbonne Université

Marie-Emilia HERBET – Pilotage du groupe de travail
Université Jean Moulin Lyon 3

Stéphane DEBARD
IRD

Françoise GENOVA
CNRS, Observatoire astronomique de Strasbourg

Christine HADROSSEK
CNRS, DDOR

Frédéric DE LAMOTTE
INRAE

Emilie LERIGOLEUR
CNRS, UMR Géode Toulouse

Gaëlle LEROUX
CNRS, Centre de recherche en neurosciences de Lyon

Gilles OHANESSIAN
CNRS

Christelle PIERKOT
Data Terra

Marie STAHL
Ecole française d'Athènes

Novembre 2023

Conception graphique : opixido

Résumé

L’ambition des politiques du ministère de l’Enseignement supérieur et de la Recherche concernant les données de la recherche est de faire en sorte que ces données soient progressivement structurées en conformité avec les principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable), préservées et partagées ou ouvertes par des entrepôts de données de confiance.

Afin de guider les équipes de recherche vers l’entrepôt le plus adapté pour le partage et l’ouverture des données de leur domaine thématique, il est indispensable d’identifier les entrepôts thématiques nationaux et internationaux de confiance, certifiés Core Trust Seal ou non.

Dans ce cadre, le Collège Données de la recherche a défini une liste de critères d’exclusion permettant de sélectionner les entrepôts thématiques de confiance, qui pourront non seulement accepter les dépôts et la publication des jeux de données, mais aussi concourir à leur diffusion et leur réutilisation ultérieure par les communautés scientifiques.

Liste des critères d’exclusion

Absence de modération des dépôts
Absence d’identifiant pérenne
Absence de garantie sur la pérennité de l’infrastructure
Entrepôts pratiquant la cession de droits
Politique tarifaire excessive
Localisation des données hors Union européenne pour certains types de données
Dépôt restreint par l’affiliation institutionnelle

Contexte

L'enjeu de l'ouverture des données de recherche, formalisé dans les politiques publiques à l'échelle nationale et européenne, incite les chercheurs à mettre à disposition de la communauté tous les matériaux de recherche utiles à la compréhension d'un résultat scientifique. Cette mise à disposition est subordonnée au respect de certains principes visant à rendre ces données intelligibles et réutilisables grâce à une documentation étayée et des métadonnées reflétant la spécificité des disciplines. Elle suppose également la capacité à garantir l'accès à ces données dans la durée. L'atteinte de ces objectifs est largement tributaire du choix de l'entrepôt retenu. En effet, le niveau d'exigence requis lors du dépôt (politique de modération, nature des métadonnées), tout comme la pérennité de l'infrastructure, déterminent en grande partie le niveau de « fairisation » des données.

L'inscription durable de ces exigences institutionnelles dans le paysage de la communication scientifique implique que les chercheurs puissent disposer des infrastructures adéquates de dépôt et d'exposition de leurs données. Si certaines communautés se sont précocement mobilisées autour de l'enjeu du partage des données (cristallographie, astrophysique, génomique, etc.) en se dotant d'entrepôts dédiés aujourd'hui largement reconnus, d'autres ne parviennent pas à identifier facilement les entrepôts thématiques susceptibles d'accueillir leurs données. Faute de recommandations de la part des financeurs de la recherche, des sociétés savantes ou des communautés, « le choix de l'entrepôt adéquat est délégué au chercheur. »[1]

Cette absence de directive peut générer deux risques : d'une part, la multiplication de dépôts erratiques dans des entrepôts généralistes sans politique de description exigeante des données. D'autre part, la montée en puissance d'entrepôts portés par les éditeurs commerciaux vers lesquels les chercheurs pourraient être orientés, faute d'offre alternative connue ou conseillée.

Ce constat intervient dans un contexte paradoxal, où certains outils de type catalogue ou annuaire d'entrepôts de données existent. Le recours à ces outils en vue de l'accompagnement au dépôt se heurte néanmoins à plusieurs écueils :

- Présence de nombreux entrepôts nécessitant une affiliation institutionnelle précise pour déposer ;
- Affectation disciplinaire inadaptée ;
- Signalement d'entrepôts non-maintenus ;
- Parcours de navigation complexe pour l'utilisateur.

Autant de facteurs synonymes de perte de temps pour le chercheur.

Cette note propose donc une méthode d'identification des entrepôts thématiques recommandés. Elle s'appuie sur les travaux engagés dès 2022 par le Collège Données de la recherche du Comité pour la science ouverte[2]. Spécifiquement missionné par le ministère de l'Enseignement supérieur et de la Recherche, le Collège Données de la recherche a été chargé d'établir une liste de critères propres à guider la sélection des entrepôts thématiques de confiance permettant le dépôt et la publication de jeux de données, en prenant prioritairement en compte les disciplines les plus actives/structurées sur la gestion des données[3].

[1] Traduction issue de : "The selection of a suitable repository is delegated to the researcher." Source : https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf

[2] https://www.ouvrirlascience.fr/college-donnees-de-la-recherche/

[3] Lettre de mission de l'administratrice ministérielle des données, des algorithmes et des codes sources, 3 mars 2023

Définition

Dans cette note, un entrepôt de données thématique se définit comme une infrastructure de services facilitant le dépôt, la description, le partage en accès ouvert ou restreint (quand la nature des données l'impose), la découverte et la réutilisation, par des humains ou des machines, de jeux de données propres à une communauté scientifique. Ces jeux de données sont associés avec des métadonnées et sont conservés à moyen ou long terme.

Les services fournis permettent de mettre à disposition des jeux de données organisés et réutilisables grâce à l'utilisation de schémas de métadonnées, l'application d'une politique de modération, l'adoption d'identifiants uniques et la garantie d'une durée minimum de préservation des données.
Les entrepôts de données peuvent avoir des exigences spécifiques et/ou des restrictions concernant :

- Le sujet ou le domaine de recherche ;
- La qualité des données ;
- L'origine des données ;
- La réutilisation et l'accès aux données ;
- Les formats de fichiers et la structure des données ;
- Les types de métadonnées.

Un entrepôt se distingue d'un catalogue, par sa capacité à assurer l'hébergement, la gestion et la curation des données et pas uniquement le système d'information (catalogage et exposition des métadonnées moissonnées à partir d'autres systèmes).

Méthode

Source d'identification des entrepôts

Le travail de repérage et d'analyse des entrepôts s'est appuyé sur l'utilisation de cinq sources principales d'information :

- Le groupe de travail dédié du Collège Données de la recherche, composé de membres du Collège et d'experts extérieurs ;
- La littérature scientifique et la littérature grise (articles décrivant le fonctionnement des entrepôts) ;
- Les annuaires d'entrepôts (CatOpidor, Re3data, Fairsharing, Opendoar) ;
- Les plateformes disciplinaires dédiées à la gestion des données de recherche (comme le consortium allemand NFDI, ou Datacc.org) proposant un premier recensement d'entrepôts thématiques ;
- Des retours d'expérience de la communauté scientifique.

Les informations présentes sur chaque site d'entrepôt ont été systématiquement exploitées. En cas de documentation manquante ou lacunaire, des contacts ont été initiés avec les responsables des entrepôts pour obtenir des précisions (le plus souvent sur la modération ou la pérennité de l'entrepôt).

Critères d'exclusion des entrepôts

La première étape a consisté à définir un socle de critères suffisamment généraliste pour ne pas réduire l'offre disponible à celle des entrepôts certifiés.

Au terme de ses travaux, le Collège Données de la recherche a donc retenu une série de sept critères d'exclusion. Trois relèvent de critères relatifs à la qualité du service fourni, quatre relèvent de critères organisationnels.

Absence de modération des dépôts

Sont écartés les entrepôts ne pratiquant pas de modération (humaine ou automatisée) visant à assurer un niveau minimum de qualité des métadonnées renseignées, ce qui permet d'éviter le versement de données incomplètes ou mal décrites. Dans ces deux exemples, les jeux de données ont des titres peu signifiants « Supplemental table S1 »[4], sans mots-clés, contexte, ni documentation associés.

Absence de garanties sur la pérennité de l'entrepôt

Seront sélectionnés les entrepôts maintenus, proposant une durée de préservation des données d'au moins 5 ans. C'est par exemple le cas de *Recherche data gouv*. A minima, une longévité déjà démontrée de l'entrepôt toujours en activité peut fournir une crédibilité suffisante.

Absence d'attribution d'identifiant pérenne

Conformément aux principes FAIR, le recours à un identifiant unique et pérenne (PID en anglais), comme par exemple un DOI rend les jeux de données plus facilement trouvables et « citables » (dans une publication par exemple).

Entrepôts pratiquant la cession de 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. 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 des acteurs commerciaux».

Politique tarifaire excessive

Ce critère vise à exclure les entrepôts conditionnant chaque dépôt de faible volume, au versement de frais. C'est par exemple le cas de Dryad qui facture chaque dépôt 150 $ voire plus en fonction des volumes déposés[5] ou encore The Digital Archaeological Record (tDAR), qui applique des frais de conservation de 10 $ par tranche de 10 Mo, ainsi que des frais de curation (contrôle basique des métadonnées et des fichiers), à raison de 90 $ de l'heure[6]. Les entrepôts pouvant appliquer une participation financière en contrepartie du dépôt de volumes importants de données (supérieurs à 50 Go) n'ont en revanche pas été écartés.

[4] https://doi.org/10.5281/zenodo.3725604

[5] https://datadryad.org/stash/faq

[6] https://core.tdar.org/cart/add

Localisation du stockage physique des données hors de l'Union européenne pour certains types de données

Certaines données telles que des données de recherche en santé, ou issues d'enquête permettent l'identification des personnes même si des techniques de pseudonymisation et d'anonymisation sont utilisées. Dans ce cas, leur communication en accès ouvert est exclue et reste strictement encadrée par l'application du Règlement général de protection des données personnelles. Le choix a donc été fait d'exclure les entrepôts de données situés hors de l'Union européenne pour les dépôts relatifs aux données personnelles qui ne sont pas totalement anonymisables, à l'exception de la Suisse, de la Grande-Bretagne, du Japon et de l'Argentine, qui appliquent le RGPD[7]. Pour les autres types de données, le signalement d'entrepôts hors de l'Union européenne a été pris en compte, d'autant que plus les chercheurs tendent à privilégier les entrepôts ayant une dimension internationale (Prost et Schöpfel, 2015)[8]

Dépôt restreint par l'affiliation institutionnelle

Les entrepôts **thématiques** restreignant le dépôt de données à certaines communautés scientifiques où seuls les chercheurs affiliés à l'institution porteuse de l'entrepôt ou à ses partenaires sont autorisés à déposer (ex : neutrons for society de l'ILL, https://data.sciencespo.fr/ pour SciencesPo) seront écartés.

La sélection qui sera proposée à l'issue des travaux vise au signalement d'entrepôts largement ouverts et accessibles au plus grand nombre, indépendamment de l'affiliation du chercheur.

[7] https://www.cnil.fr/fr/la-protection-des-donnees-dans-le-monde

[8] https://hal.univ-lille.fr/hal-01198379v1/document

Critères de description des entrepôts

Le Collège Données de la recherche dressera pour chaque entrepôt retenu une courte fiche d'identité, reprenant les informations nécessaires aux équipes de recherche dans leur démarche de dépôt de jeux de données. Au-delà des informations descriptives d'ordre général (nom, URL, institution porteuse), le choix a été fait de se recentrer sur huit items:

Champ disciplinaire

Le champ disciplinaire s'appuie sur la nomenclature utilisée par HAL, qui propose notamment une déclinaison pertinente des disciplines en sciences humaines et sociales. De plus, cette nomenclature offre jusqu'à trois niveaux différents de granularité, ce qui permet de mieux décrire les entrepôts sélectionnés.

Données acceptées

Il s'agit ici de décrire le type de données acceptées par l'entrepôt, en prenant soin d'employer la terminologie spécifique propre à chaque discipline afin de faciliter le choix du déposant. Par exemple : spectres RMN aux formats constructeurs et JCAMP-DX pour tous types d'échantillons.

Identifiant pérenne fourni par l'entrepôt

Sont mentionnés dans cette rubrique tout identifiant pérenne (DOI, Handle…), contribuant à faciliter la « découvrabilité » des jeux de données selon les principes FAIR.

Pérennité des données

Ce critère traite des aspects liés à la pérennité du dispositif et/ou l'engagement de l'entrepôt à préserver les données déposées pendant une période de temps expressément définie.

Type de modération

Ce critère précise le type de modération pratiquée (vérification des métadonnées, contrôle scientifique des données, intervention humaine ou automatisée etc.).

Possibilité d'embargo

Certains chercheurs peuvent souhaiter poser un embargo sur leurs jeux de données. Ce critère précise donc la possibilité offerte ou non par l'entrepôt d'assortir le dépôt à un embargo.

Limite de volume

Cette information peut se révéler importante pour les chercheurs issus de disciplines générant des volumes importants de données. Elle permet également d'anticiper le coût à prévoir si l'entrepôt impose une participation financière à partir d'un certain volume.

Remarques

Ce critère signale toute information complémentaire utile au signalement et à la caractérisation de l'entrepôt (précisions sur les modalités de dépôt…).

Articulation avec les travaux menés dans le cadre de la Research Data Alliance

Un groupe de travail de la Research Data Alliance « Data Repository Attributes Working Group (DRAWG) » travaille depuis début 2022 à établir une liste de critères de « haut niveau » pour caractériser les entrepôts de données de recherche. Ce travail n'est pas terminé, donc *a fortiori* ni validé, ni publié par la RDA. Il s'agit de fournir une liste d'attributs principaux, sans critère d'exclusion ni objectif de recommandation. La liste est destinée à tous les acteurs de la recherche, pas seulement aux scientifiques producteurs de données. La démarche est donc assez différente de celle qui est décrite ici. Pour autant, tous les critères choisis ici sont présents dans la liste du DRAWG sous une forme identique ou voisine. La liste de critères retenus dans cette note est plus restreinte afin d'en faciliter l'utilisation.

Biais et limites

Lacunes disciplinaires

La liste qui sera fournie à partir des critères explicités ci-dessus pourra paraître incomplète ou partielle. Elle est le reflet du paysage actuel des entrepôts thématiques, où certaines disciplines sont très bien structurées avec de nombreux entrepôts reconnus et de confiance quand d'autres sont moins bien outillées.

Le repérage est également dépendant des disciplines représentées au sein du groupe de travail, dont la composition ne reflète pas à l'heure actuelle une couverture disciplinaire exhaustive. Cette première liste sera progressivement complétée et mise à jour, en suivant la méthodologie décrite ci-dessus.

Informations partielles ou déclaratives

Le recensement des entrepôts s'appuie en majeure partie sur les informations collectées sur les sites des entrepôts mais également auprès des porteurs des différentes infrastructures. Cette collecte d'information repose donc sur des propos déclaratifs, dont la véracité n'a pu être vérifiée.

Le critère de durée de préservation des données fait notamment partie des éléments à considérer avec prudence. Face à la difficulté d'identifier la durée d'engagement des entrepôts, la grille d'analyse a parfois été assouplie, en s'appuyant sur d'autres gages de crédibilité, comme les tutelles porteuses de l'entrepôt ou sa longévité.

Conformité avec les critères FAIR

Les travaux menés visent à rendre visible des entrepôts pertinents auprès des différentes communautés afin de faciliter l'exposition des données. L'ensemble des entrepôts proposés respecte les principes généraux FAIR mais ne précise pas le niveau de « fairisation » des métadonnées disciplinaires quand il existe ou encore de quelle manière ces métadonnées sont lisibles par des machines.