Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante

Frédéric de Lamotte, Véronique Stoll, Cécile Arènes, Marie-Emilia Herbet, Stéphane Debard, Françoise Genova, Christine Hadrossek, Emilie Lerigoleur, Gaelle Leroux, Gilles Ohanessian, et al.

Ouvrir la science !

Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante

Groupe de travail du collège Données de la recherche

Juillet 2024

Sélectionner un entrepôt thématique de confiance pour l'auto-dépôt de données : méthodologie et analyse de l'offre existante

_____

Groupe de travail

_____

Frédéric de LAMOTTE – Pilotage du Collège des données de la recherche
INRAE

Véronique STOLL – Pilotage du Collège des 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 - UMR Espace-Dev

Françoise GENOVA
CNRS, Observatoire astronomique de Strasbourg

Christine HADROSSEK
CNRS, DDOR

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

Le Collège des données de la recherche remercie pour leur appui et leur expertise :

Thomas JOUNEAU, Université de Lorraine
Maria-Grazia SANTANGELO, Université Grenoble Alpes

———

Juillet 2024

———

DOI : 10.52949/52

Conception graphique : opixido

Sommaire

Table des figures

Contexte

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 *via* des entrepôts de données de confiance. Formalisé dans les politiques publiques à l'échelle nationale et européenne, l'enjeu de l'ouverture des données de recherche 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 n'identifient pas encore 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'une autre offre alternative connue ou conseillée.

Ce constat intervient dans un contexte paradoxal, où certains outils de type catalogues ou annuaires 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.

Le 3 mars 2023, le Ministère de l'Enseignement supérieur et de la Recherche a confié au Collège des données de la recherche la double mission de :

- Proposer une définition de critères de confiance permettant d'apprécier la qualité d'un entrepôt à l'aune des objectifs de « FAIRisation » des données ;
- Œuvrer à l'identification d'entrepôts thématiques de confiance propres à être moissonnés par *Recherche Data Gouv*.

[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

En effet, l'écosystème Recherche Data Gouv propose notamment aux équipes de recherche une offre pluridisciplinaire de dépôt, publication et signalement de leurs données, en complémentarité avec les entrepôts thématiques et institutionnels. Aussi apparaît-il indispensable d'identifier les entrepôts thématiques nationaux et internationaux de confiance, certifiés Core Trust Seal ou non, afin de guider les équipes de recherche vers l'entrepôt le plus adapté pour le partage et l'ouverture des données de leur champ disciplinaire.

Cette note propose donc une méthode d'identification des entrepôts thématiques recommandés permettant l'auto-dépôt des données par les communautés, assortie d'une première liste découlant de la grille de critères d'analyse retenus. Elle s'appuie sur les travaux engagés dès 2022 par le Collège des 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 des 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].

Conformément à la Charte des Collèges et groupes d'expertise du Comité pour la science ouverte (mai 2023), la liste des entrepôts de confiance établie par le Collège des données de la recherche fait l'objet d'échanges réguliers avec le bureau de la science ouverte, qui en effectue la validation finale avant transmission aux structures appropriées pour la diffusion de cette liste.

Méthodologie

Définition

Le Collège des données a défini un entrepôt de données thématique comme une infrastructure de stockage et de services facilitant le dépôt, la description, le partage en accès ouvert, 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 à des métadonnées et sont conservés à moyen ou long terme.

Les services fournis permettent d'organiser les données de manière logique et cohérente, par exemple à travers l'adoption d'identifiants uniques, l'application d'une politique de modération et la garantie d'une durée minimum de préservation des données.

Les entrepôts de données peuvent avoir des exigences spécifiques et/ou des restrictions statutaires 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 structures).

[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, en annexe.

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 des 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).

Articulation avec les autres travaux menés sur les entrepôts de données

Research Data Alliance

Le groupe de travail de la Research Data Alliance *Data Repository Attributes Working Group* (DRAWG) a travaillé en 2022 et 2023, à établir une liste de critères de « haut niveau » pour caractériser les entrepôts de données de recherche. Ce travail a été soumis à commentaires début 2024 ; il 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 ici est plus restreinte afin d'en faciliter l'utilisation.

Cat OPIDoR

Cat OPIDoR propose un wiki des services dédiés aux données de la recherche, qui recense de façon exhaustive les entrepôts de données français. Cet inventaire comprend également CNRS données de la recherche, catalogue qui recense les entrepôts et les services dédiés aux données de la recherche du CNRS et des structures qui les portent. Les objectifs de Cat OPIDoR et de la présente étude diffèrent, le premier étant dédié à un état de l'art des entrepôts français quand la seconde a défini des critères d'inclusion et d'exclusion aboutissant à une liste d'entrepôts français et internationaux.

DoRANum

En décembre 2023, une nouvelle ressource a été publiée sur DoRANum, plateforme d'auto-apprentissage consacrée aux données de la recherche qui a rejoint les centres de ressources de l'écosystème Recherche Data Gouv. Les exemples d'entrepôts de données mis en ligne portent sur les sciences humaines et sociales avec une dizaine d'entrepôts recensés, strictement disciplinaires ou plus généralistes. Une vingtaine de critères ont été retenus, représentés sous forme de pictogrammes. Le type de données acceptées, qui a constitué un critère de description de la présente étude, n'a pas été spécifié dans les fiches mises en ligne sur DoRANum.

Axes de travail et livrables

Le GT a défini trois axes de travail principaux pour réponde à la mission qui lui a été confiée :

- Établir la liste de critères de confiance permettant d'apprécier la qualité d'un entrepôt thématique ;
- Proposer une note méthodologique et une première liste d'entrepôts thématiques de confiance ;
- Proposer une stratégie de pérennisation de la liste d'entrepôts thématiques de confiance.

Liste de critères de confiance d'un entrepôt thématique

Objectifs

Établir une liste de critères appropriés permettant la sélection des entrepôts thématiques de confiance permettant l'auto-dépôt par les communautés de recherche et la publication de jeux de données.
Ces critères constitueront un cadre d'analyse homogène. Ils permettront de fournir une fiche d'identité argumentée pour chaque entrepôt étudié.

Acteurs

Membres du GT

Livrables

Liste de critères de description et d'exclusion pour une sélection d'entrepôts thématiques

Calendrier

2e trimestre 2023

Proposer une note méthodologique et une liste d'entrepôts thématiques de confiance

Objectifs

- Rédiger une note méthodologique explicitant la démarche mise en place pour constituer la liste des entrepôts thématiques (sources de signalement des entrepôts, experts consultés, prise de contact avec les administrateurs des entrepôts etc.) ;
- Sur la base des critères de confiance définis et pour guider les communautés de recherche françaises lors de la publication de leurs jeux de données, proposer une liste d'entrepôts de confiance qui sera complétée par étape. Pour valider la pertinence de cette sélection, le GT s'appuie sur des experts extérieurs issus de différentes communautés scientifiques. Établie dans le cadre de la feuille de route du Collège des données de la recherche, cette liste alimentera les travaux du groupe de travail élargi du module catalogue de Recherche Data Gouv.

Acteurs

Membres du GT et experts extérieurs

Livrables

- Note méthodologique ;
- Liste d'entrepôts thématiques recommandés.

Calendrier

- Note méthodologique ; 3e trimestre 2023 ;
- Première version de la liste des entrepôts : 3e trimestre 2023 ;
- Deuxième version de la liste des entrepôts : 1er trimestre 2024.

Stratégie de pérennisation de la liste d'entrepôts

Objectifs

- Proposer des scénarios de pérennisation de la liste des entrepôts de confiance afin d'assurer sa mise à jour ou son extension ;
- Participer à la réflexion sur la mise à jour de la liste des entrepôts de confiance moissonnés par *Recherche Data Gouv*.

Le livrable fourni par le Collège des données de la recherche pourra être actualisé durant la mandature actuelle par les membres du GT qui assurent une veille sur les entrepôts. Au-delà, il est indispensable de proposer des scénarios de révision régulière non seulement de la liste des entrepôts, mais aussi de la validité des critères de confiance, dont le niveau d'exigence est susceptible de s'accroître dans le temps.
La liste d'entrepôts thématiques sera révisée annuellement, et/ou suivant les évolutions des exigences des agences de financement de la recherche.

Acteurs

Membres du GT

Livrables

Note stratégique et scénarios pour la mise à jour de la liste des entrepôts thématiques de confiance et de la liste des critères.

Calendrier

1er trimestre 2024

Critères d'exclusion et de description des entrepôts

Critères d'exclusion des entrepôts

Le Collège des données de la recherche a retenu une série de sept 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. Trois relèvent de critères relatifs à la qualité du service fourni, quatre relèvent de critères organisationnels.

Ce socle de critères a été arrêté pour établir une première liste d'entrepôts facilement identifiables et utilisables par les équipes de recherche, tout en se voulant suffisamment généraliste pour ne pas réduire l'offre disponible à celle des entrepôts certifiés.

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

Absence de modération des dépôts

Devraient être 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 cet exemple[4], le jeu de données est doté d'un titre peu signifiant « Supplemental table S1 », sans mots-clés, ni contexte, ni documentation associée. Nous considérons que l'étape de la modération permet d'encourager le déposant à clarifier et mieux documenter son jeu de données.

Absence de garanties sur la pérennité de l'infrastructure

Nous recommandons de privilégier les entrepôts maintenus, proposant une durée de préservation des données d'au moins 5 ans, à l'instar des pratiques mises en place par 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 pérenne fourni par l'entrepôt (PID en anglais, DOI par exemple) 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.

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 ».

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

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.

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

Certaines données (données de santé, résultats 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 (RGPD). 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 anonymisables, à l'exception de la Suisse, de la Grande-Bretagne, du Japon et de l'Argentine, qui sont considérés comme étant en adéquation avec 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

Sont écartés les entrepôts thématiques restreignant le dépôt de données à certaines communautés scientifiques où seuls les scientifiques affiliés à l'institution porteuse de l'entrepôt sont autorisés à déposer. La sélection proposée à l'issue des travaux vise donc le signalement d'infrastructures largement ouvertes et accessibles au plus grand nombre, indépendamment de l'affiliation du chercheur.

Critères de description des entrepôts

Pour chaque entrepôt, une courte fiche d'identité est dressée, 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 sept items :

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

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

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

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

Liste des critères de description

Champ disciplinaire
Données acceptées
Fourniture d'un identifiant pérenne
Pérennité des données
Type de modération
Possibilité d'embargo
Limite de volume

Champ disciplinaire

Le champ disciplinaire reprend, lorsque cela a été possible, la nomenclature utilisée par HAL, qui propose une déclinaison adaptée aux sciences humaines et sociales, ce qui n'est pas toujours le cas d'autres nomenclatures existantes. De plus, elle propose 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 accepté par l'entrepôt, en prenant soin d'employer la terminologie spécifique propre à chaque discipline afin de faciliter le choix du déposant (ex : spectres RMN, structures 3D de molécules biologiques, corpus encodés en TEI, etc.).

Fourniture d'un identifiant pérenne

Sont mentionnés dans cette rubrique tout identifiant pérenne fourni par l'entrepôt (DOI, ARK, 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 de l'infrastructure 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

Certaines équipes de recherche peuvent souhaiter retarder la publication en accès ouvert de leurs jeux de données en apposant une période déterminée d'embargo. 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

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

Résultats et analyse disciplinaire

Les travaux menés par le Collège des données entre octobre 2022 et décembre 2023 ont abouti à la constitution d'une première liste comprenant 49 entrepôts thématiques relevant aussi bien des sciences exactes que des sciences humaines et sociales et répartis comme suit :

Répartition thématique des entrepôts sélectionnés

Domaine	Nombre d'entrepôts
Astronomie	2
Biologie	13
Chimie	9
Physique	2
Sciences de l'environnement	7
Sciences humaines et sociales	16 (dont 6 en archéologie et 3 en linguistique)

Le recensement de ces entrepôts s'appuie en majeure partie sur les informations fournies par les sites web des entrepôts, mais également par les porteurs des différentes infrastructures. Cette collecte d'information repose donc sur des propos déclaratifs, dont la véracité n'a pas toujours pu être vérifiée, en particulier pour les pratiques de modération impliquant une intervention manuelle.

Le critère de durée de préservation des données fait également 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'infrastructure ou la longévité de l'entrepôt.

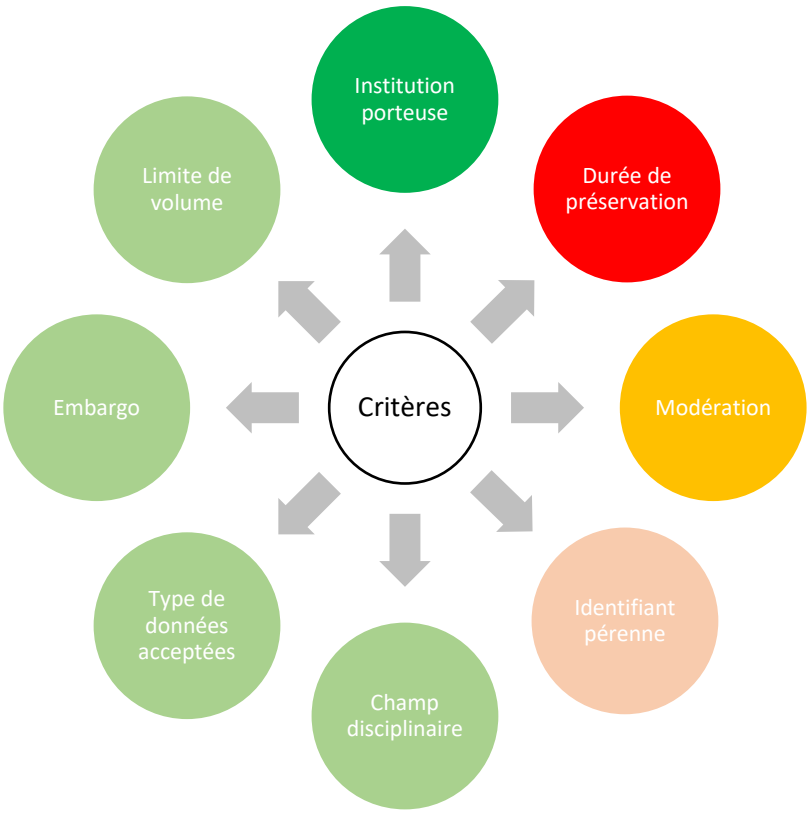
Les travaux menés cherchent à rendre visibles des entrepôts pertinents auprès des différentes communautés, afin de faciliter l'exposition de leurs données. L'ensemble des entrepôts proposés respecte les principes généraux FAIR mais ne précise pas le niveau de « FAIRisation » des métadonnées disciplinaires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

Ce recensement s'est heurté à plusieurs écueils qui reflètent les disparités de pratiques d'une discipline à l'autre mais également la fragilité inhérente au modèle économique adossé aux entrepôts de données. Les travaux ont nécessité de prendre directement contact auprès des responsables pour 31 des entrepôts signalés. Le travail mené a par ailleurs abouti à l'exclusion d'une quarantaine d'entrepôts.

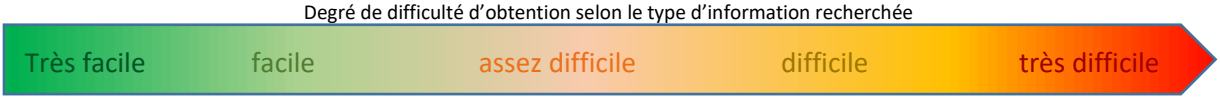
Application des critères d'exclusion et de description

Comme indiqué précédemment, le recensement s'est appuyé sur une série de critères d'exclusion qui ont permis l'identification de critères qualitatifs et organisationnels, établie après examen d'autres expériences menées en France comme à l'international[9]. La grille de critères retenue par le Collège des données constitue le socle minimum d'informations essentielles et nécessaires pour éclairer les équipes de recherche dans leur choix de dépôt. Volontairement synthétique, afin de répondre efficacement aux besoins des utilisateurs, cette liste de critères a suscité une vaste campagne de collecte d'informations, dont certaines se sont révélées difficiles voire impossibles à obtenir.

Classement des critères par difficulté d'obtention des informations



Degré de difficulté d'obtention selon le type d'information recherchée



Les critères concernant la limite de volume, l'embargo, le type de données acceptées et les champs disciplinaires détaillés ont été parmi les plus aisés à recenser. On les trouve régulièrement sur la documentation des entrepôts,

[9] Parmi les grilles de description identifiées, voir l'outil d'aide à la sélection d'entrepôts de Datacc.org (Lyon 1/UGA), les listes de critères désirables du NIH et du sous-comité à la science ouverte du gouvernement fédéral américain, la liste de critères désirables ou essentiels du COAR ou encore les travaux menés dans le cadre de RDA (RDA Fairsharing WG).

même s'il a parfois fallu un contact avec les porteurs des entrepôts pour obtenir des précisions sur les embargos ou la limite de volume, par exemple.

Obtenir des informations sur les identifiants pérennes s'est révélé plus complexe qu'attendu, pour au moins trois raisons :

- Attribution d'identifiants internes dont la valeur d'identifiant pérenne était difficile à confirmer ;
- Présence de DOI attribués aux publications associées aux jeux de données et non à ces derniers ;
- Attribution d'identifiants pérennes seulement sur demande.

La question de la modération a par ailleurs nécessité de nombreux échanges avec les porteurs des entrepôts afin de préciser la politique en place (modération automatique et/ou humaine, caractère systématique ou non de cette modération, évaluation éventuelle de la qualité des données, etc. La notion de modération étant polysémique, elle a occasionné de nombreuses discussions autour des entrepôts à retenir, d'autant plus qu'un entrepôt peut également faire évoluer sa politique de modération durant la phase de sélection.

Enfin, le critère sur la durée de préservation s'est avéré, dans de nombreux cas, impossible à obtenir. De très nombreux entrepôts sont adossés à des financements non pérennes découlant de projets de court terme, compromettant la capacité des infrastructures à s'engager sur une durée de préservation plus longue que la durée du projet lui-même. Le GT a même observé qu'un entrepôt en sciences humaines et sociales a indiqué ne plus accepter de nouvelles données, faute de nouveau financement. Dans la mesure où très peu d'entrepôts sont en mesure de s'engager sur la conservation pérenne des données, ce qui constitue une grave fragilité dans l'écosystème des entrepôts thématiques, le GT a dû amender ce critère d'exclusion, en prenant en compte l'ancienneté d'un entrepôt et ses différentes sources de financement.
La difficulté à laquelle l'équipe-projet s'est heurtée se reflète dans les études menées sur la pérennité des entrepôts de données. De récentes analyses ont pu montrer que 6,2% des entrepôts recensés dans re3data avaient cessé leur activité, après une période d'activité médiane d'environ 12 ans. Parmi eux, on constate une sur-représentation des entrepôts thématiques (136 entrepôts sur 191), principalement en sciences du vivant et sciences naturelles.[10]

Analyse disciplinaire

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

Le repérage est également dépendant des disciplines représentées au sein du groupe de travail, dont la composition ne couvre pas, à l'heure actuelle, les disciplines de manière exhaustive. Cette première liste sera progressivement complétée et mise à jour, en suivant la méthodologie décrite ci-dessus et selon une stratégie de pérennisation qui reste à valider.

Astronomie

Souvent implantés de longue date, notamment grâce à la structuration de la discipline autour de l'IVOA (*International Virtual Observatory Alliance*), les entrepôts en astronomie sont généralement alimentés par des documentalistes qui assurent le dépôt et la curation des données pour les équipes de recherche. Il peut s'agir de données de missions spatiales, dont la gestion est conçue comme partie intégrante de la mission et le dépôt organisé collectivement. Il s'agit aussi de données obtenues dans un observatoire au sol, qui peuvent être collectées auprès de l'observatoire après un embargo d'un an en général. Ces caractéristiques spécifiques à

[10] Une proportion qui peut aussi s'expliquer par la nature de l'échantillon d'entrepôts tirés de re3data, qui comprend une sur-représentation d'entrepôts thématiques dans ces disciplines. « Compared to all repositories indexed in re3data at the time of data collection, repositories with these characteristics are also overrepresented in the sample. » p.10 https://arxiv.org/pdf/2310.06712.pdf

l'astronomie expliquent que peu d'entrepôts de la discipline acceptent l'auto-dépôt, qui est l'un des critères de cette étude. D'où le nombre relativement faible d'entrepôts signalés dans cette discipline.

Si l'écosystème des entrepôts en astronomie est très international, les deux entrepôts retenus car permettant l'auto-dépôt sont français : le Centre de données astronomiques de Strasbourg et *Paris Astronomical Data Center*. Le premier a œuvré dès 1972 à la collecte de données en astronomie, quand le deuxième existe depuis 20 ans. Dans les deux cas, d'importants volumes de données peuvent être déposés. Il est toutefois à noter que certaines sous-disciplines de l'astrophysique ne disposent pas d'entrepôts dédiés et peuvent de ce fait être éparpillées dans d'autres entrepôts.

Biologie

L'offre en biologie étant plus large que dans les autres disciplines étudiées, les 13 entrepôts signalés dans le présent document ne représentent qu'une fraction des infrastructures adaptées au dépôt et partage de données. Cette richesse d'offre comprend néanmoins plusieurs biais. On constate ainsi une sur-représentation des entrepôts basés dans les pays anglo-saxons et un soutien financier marqué de la part des NIH (National Institutes of Health), identifiés comme institution porteuse de trois des entrepôts recensés. Cette réalité est compensée par la présence de l'organisation européenne EMBL-EBI (EMBL's European Bioinformatics Institute), impliquée dans les entrepôts ENA et PRIDE. Le dépôt de données d'ADN et d'ARN se structure historiquement autour de trois grandes bases internationales : GenBank aux États-Unis, ENA en Europe et DDBJ au Japon. Dans la mesure où ces trois bases sont engagées dans un processus de coopération pour faciliter le partage des séquences génétiques[11], nous avons opté pour le signalement de l'infrastructure européenne, qui apparaît comme le point d'entrée naturel des chercheurs exerçant en France.

Cette discipline se caractérise aussi par l'existence d'entrepôts ayant une forte antériorité (plus de 50 ans d'existence pour PDB et près de 40 ans pour ENA). La nature des données éligibles au dépôt dans les entrepôts recensés porte essentiellement sur la neurobiologie et la génomique. En neurobiologie, nous avons pris soin de ne recommander que le dépôt de données d'imagerie cérébrale non-humaine dans les entrepôts localisés aux États-Unis. La biologie végétale semble moins bien dotée, avec un seul entrepôt signalé dans la liste. L'utilisation de PID dans les entrepôts de cette discipline repose souvent sur des identifiants internes.

Chimie

La chimie constitue un domaine où la structuration et la diffusion des données est encore peu mature, à l'exception de la cristallographie, qui dispose d'une expertise historique en la matière. En témoigne l'ancrage de Cambridge Structural Database, dont la genèse remonte aux années 70. Le deuxième entrepôt relevant de cette spécialité, Crystallography Open Database, existe depuis 20 ans et s'est quant à lui mobilisé sur la réutilisation intégralement libre des données dès sa genèse.

Au total, nous avons identifié neuf entrepôts pertinents, dont six impliquent un portage ou co-portage institutionnel allemand.

La plupart des entrepôts décrits sont conformes aux critères qualitatifs définis, même si le caractère récent de certaines initiatives nécessite une vigilance accrue sur les politiques de préservation en place. Leur mention pouvait néanmoins se justifier par la pertinence de ces entrepôts pour la communauté. Open Reaction Database, en place depuis 2021 seulement propose une interface ergonomique de dépôt et des métadonnées structurées, dans une spécialité où l'ouverture des données est encore balbutiante et l'effort de standardisation des protocoles d'expérience relativement récent.

Dans l'ensemble, les entrepôts de chimie sont relativement spécialisés en fonction d'un type de données précis (modélisation moléculaire en chimie théorique, interactions intermoléculaires en chimie supramoléculaire, réactions chimiques en chimie organique etc.).

[11] Voir https://www.insdc.org/

Physique

Discipline où la production de données peut être massive, la physique présente une offre très restreinte d'entrepôts permettant l'auto-dépôt de données, indépendamment de l'institution porteuse. Nous recensons deux entrepôts, tous deux européens, remplissant les critères qualitatifs définis, l'un en physique des particules, porté par l'Université de Durham et le CERN, et l'autre en sciences des matériaux. Ce dernier entrepôt, NOMAD, a vocation à accueillir des données computationnelles en sciences des matériaux qui peuvent relever d'autres domaines que la physique.

Sciences de l'environnement

Avec sept entrepôts recensés, l'offre en sciences de l'environnement est satisfaisante. La liste atteste également de la présence d'un portage institutionnel français en la matière, pour plus de la moitié des entrepôts identifiés : SEANOE, SEXTANT, EasyData et Data Indores, ces deux derniers ayant été lancés récemment (2023 et 2021 respectivement).

Les entrepôts relevant des sciences de l'environnement et écologie se caractérisent par une politique de modération assez bien structurée, une possibilité de dépôt de volumes importants et la présence notable d'entrepôts certifiés. Le type de données éligibles peut être très variable : si certains entrepôts, à l'instar de Data Indores, ont une approche plutôt généraliste, d'autres infrastructures, tels que SEANOE ou encore le World Data Center of Climate, se concentrent sur un type de données en particulier (données marines géolocalisées pour le premier et simulations climatiques pour le dernier).

Sciences humaines et sociales

Avec 16 entrepôts recensés, les sciences humaines et sociales disposent d'une offre hétérogène répondant tantôt aux besoins de certaines communautés spécifiques (à l'instar de la linguistique ou de l'archéologie) tantôt à des besoins d'ordre plus générique, lorsque le type de données susceptible d'être déposé relève des SHS au sens large, sans ciblage thématique (Nakala).

Le paysage des entrepôts en SHS se caractérise par une forte représentation d'infrastructures françaises (10 entrepôts sur les 16 recensés) au sein d'une offre presque exclusivement européenne, à l'exception de Qualitative Data Repository (QDR) pour les données qualitatives et OpenContext en archéologie, basés aux États-Unis.

La conformité aux critères qualitatifs est particulièrement présente pour les entrepôts spécialisés dans les données d'enquête (CDSP, Progedo Diffusion, QDR), où la politique de dépôt est encadrée. Cela tient à la nature même des données d'enquêtes qui peuvent être mises à disposition mais doivent garantir l'anonymat dans le respect du RGPD. La contextualisation de l'enquête constitue également un travail important et spécifique à ces entrepôts. La politique de modération est également bien développée en archéologie et linguistique, où le formatage des données est largement pratiqué puis vérifié par une modération automatique et/ou humaine.

Dans d'autres cas, la politique de modération appliquée est partielle ou inexistante, mais le choix d'un signalement a néanmoins été retenu afin de répondre aux besoins de la communauté en histoire ou d'anticiper la mise en place d'une politique de modération à venir.

A l'inverse de l'astronomie par exemple, les communautés en SHS ont eu tendance à s'organiser autour d'un modèle d'auto-dépôt libre. La modération des métadonnées se met progressivement en place pour une partie des entrepôts étudiés mais elle ne fait pas partie des standards de toute la communauté des sciences humaines et sociales. L'attribution d'identifiants pérennes est en revanche assez répandue, ce qui se reflète dans le choix des entrepôts signalés, exception faite de Geovistory, que nous signalons tout de même pour sa spécificité dans le paysage des SHS (base de données relationnelle entre des figures historiques, des lieux et des organisations).

La granularité des données collectées dans les entrepôts en SHS est très variable, avec un ciblage surtout marqué en archéologie (modélisation, restitution 3D d'artefacts, échantillons bioarchéologiques, etc.) et dans une certaine mesure en linguistique (enregistrements sonores, corpus encodés en TEI ou XML, lexiques etc.). Selon le type de données visées, certaines disciplines, comme la géographie, sont amenées à émarger sur plusieurs entrepôts : Nakala pour les données d'enquête par exemple, Data Station Archaeology pour les SIG, Data Indores pour les données d'observation relevées sur le terrain, ou encore Pangaea ou EasyData pour les données de géosciences.

Les entrepôts étudiés en SHS sont relativement jeunes et il a été très difficile, voire impossible pour certains, d'obtenir des informations sur la pérennité de l'infrastructure. L'engagement sur la durée de conservation des données est fragile pour nombre d'entrepôts, en raison de financements sur des projets de courte durée ou de méconnaissance des orientations à venir des établissements porteurs.

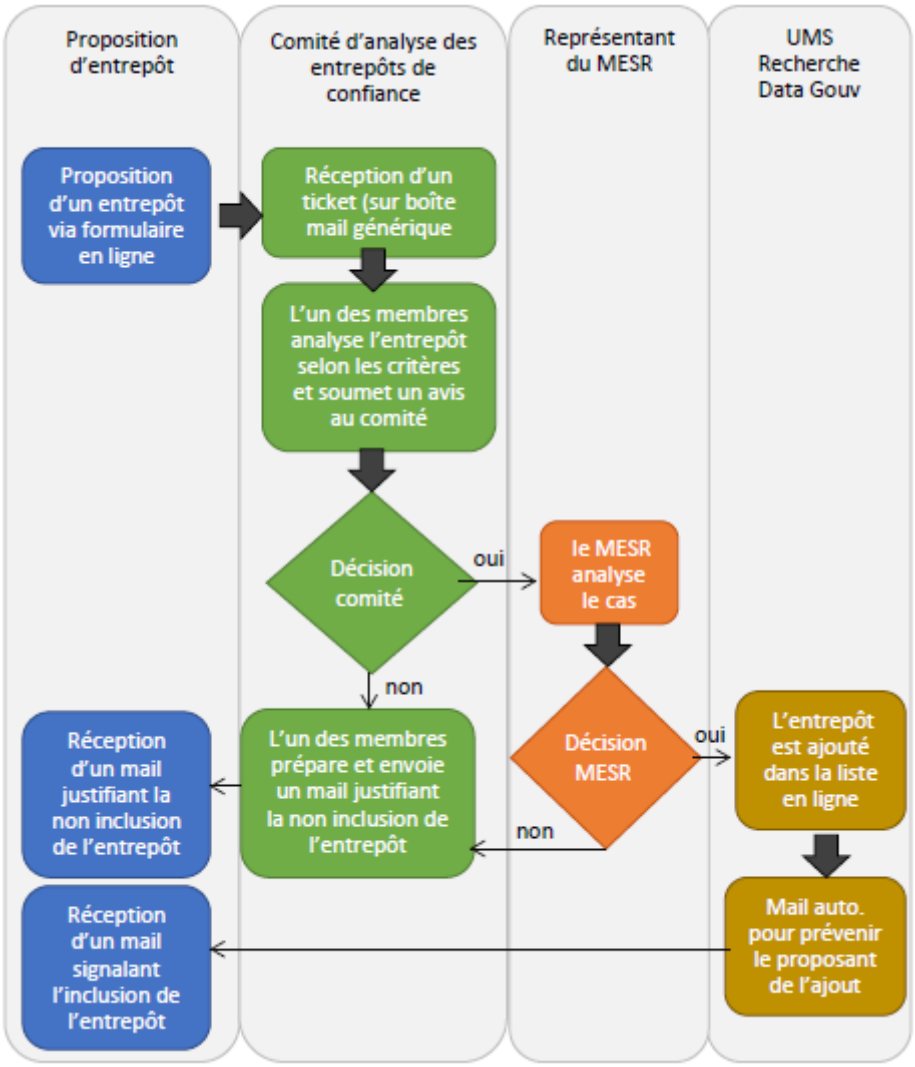
Premiers éléments pour une mise à jour de la liste des entrepôts

Le mandat actuel du Collège des données de la recherche s'achèvera à la fin du premier semestre 2025. Jusqu'à cette date, le GT constitué travaillera à la mise à jour de la liste, et à l'examen de nouvelles propositions qui seront faites par les utilisateurs ou les différents acteurs accompagnant les équipes de recherche (ateliers de la donnée, centres de ressources thématiques etc.), par exemple par le biais d'un formulaire.

Cette période pourrait également être mise à profit pour la mise en place d'un comité d'analyse pérenne, qui prendrait le relai du Collège des données dans la durée et dont la mission serait de contribuer à enrichir la liste des entrepôts, à l'aune des critères établis dans la présente note. Ce comité sera formé à l'usage des critères par le GT du Collège des données et pourra rassembler des membres issus de :

- l'équipe entrepôt-catalogue de Recherche Data Gouv ;
- Ateliers de la donnée de Recherche Data Gouv ;
- Centres de référence thématiques de Recherche Data Gouv ;
- Anciens membres du GT du Collège ;
- Tout expert disciplinaire.

Proposition de workflow pour la mise à jour de la liste des entrepôts



Conclusion

Au terme de ce travail de recensement des entrepôts, plusieurs enseignements peuvent être tirés. Si un consensus s'est formé assez rapidement au sein du GT sur l'adoption d'une série de critères qualitatifs, l'application de ces derniers à l'analyse des entrepôts thématiques a soulevé de nombreuses interrogations voire difficultés.

En premier lieu, le caractère parcellaire des informations disponibles publiquement sur les entrepôts (*via* les sites en ligne ou la littérature scientifique) a été en partie comblé par une prise de contact auprès des porteurs des entrepôts. Cette démarche n'a néanmoins pas permis de répondre à toutes les attentes, dont celle, particulièrement lancinante, de la pérennité des infrastructures et de leur engagement sur la conservation de moyen ou long terme des données collectées. Cette information, qui ne figure presque jamais sur les sites des entrepôts, reste donc totalement inaccessible aux déposants, ce qui n'est pas sans poser problème quant aux garanties apportées. Lors des échanges avec les porteurs d'entrepôts, les réponses ont souvent fait état d'un financement sur projet, donc limité dans le temps, ou d'une recherche de financements en cours. L'incertitude de ces réponses a provoqué chez les membres du groupe une certaine inquiétude concernant la robustesse de l'écosystème des entrepôts disciplinaires. Le constat, dont on espère qu'il n'est que provisoire, est que les jeux de données déposés ont une durée de vie comprise entre trois et six ans environ si l'on se fonde sur la durée des financements octroyés aux entrepôts. Cette fragilité est en partie compensée par des mesures de continuation adoptées lorsqu'un entrepôt est appelé à fermer. Le GT a pu en faire le constat pour un entrepôt en SHS dont les financements se sont éteints et qui n'a eu d'autre choix que d'organiser la migration des données vers un autre entrepôt, tout en se heurtant à des difficultés pour certains formats de données. Une étude sur l'archivage des données de la recherche, menée en lien avec les services d'archives des établissements, pourrait constituer un prolongement au présent travail pour envisager une bonne articulation entre le partage des données sur les entrepôts thématiques et leur conservation à long terme, par exemple sur des systèmes d'archivage électroniques (SAE) dont certains établissements disposent déjà. Les dispositions du Code de la Recherche sur l'intégrité scientifique rappelées par Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique font en effet porter aux établissements la responsabilité de la conservation des résultats bruts de la recherche afin de permettre leur vérification. L'enjeu est donc double : il s'agit de se conformer sans tarder au cadre légal et il est primordial de garantir la préservation à long terme de ces nouveaux objets patrimoniaux que sont les jeux de données. Il en va de la constitution du patrimoine scientifique, devenu nativement numérique, des années à venir.

En deuxième lieu, les critères d'exclusion, qui semblaient raisonnables au moment des premières recherches autour des entrepôts thématiques, se sont révélés complexes à appliquer pour deux d'entre eux. La modération n'est pas appréhendée de la même façon dans toutes les disciplines, certaines d'entre elles souhaitant garantir une grande liberté pour le déposant quand d'autres contrôlent les métadonnées du jeu de données, les métadonnées intrinsèques et les données elles-mêmes. Dans ce contexte, il n'a pas été aisé de trouver un juste équilibre et certains entrepôts recensés ne répondent pas complètement au critère de modération. Ils ont néanmoins été retenus pour leur conformité aux autres critères et pour leur importance au sein de leur discipline.

Cette première liste ne doit pas être perçue comme un aboutissement, mais comme une amorce visant à fournir un premier niveau d'aiguillage pour les équipes de recherche concernées par le partage des données. Elle a donc vocation à être actualisée selon au moins quatre axes : mise à jour des informations relatives aux entrepôts recensés, ajout de nouveaux entrepôts, suppression d'entrepôts présents dans la liste, si le cas devait se présenter, modification des critères de choix. Le paysage des entrepôts thématiques se caractérisant par des évolutions rapides, la liste fournie sera vite frappée d'obsolescence. Elle ne conservera son intérêt que si elle fait l'objet de mises à jour régulières.

Annexe : lettre de mission

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Direction générale de la recherche et de l'innovation

Direction générale de l'enseignement supérieur
et de l'insertion professionnelle

Paris, le 3 mars 2023

Madame Isabelle BLANC
Administratrice ministérielle des données, des algorithmes
et des codes sources

à

Madame Véronique Stoll et Monsieur Pierre-Yves Arnould
Co-pilotes du Collège des données de la recherche
Comité pour la science ouverte

Objet : Entrepôts thématiques de données de recherche

Madame, Monsieur,

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.
Le développement de services pour accompagner les équipes de recherche dans la gestion, le partage, l'ouverture et la réutilisation de données de recherche occupe une place centrale de ces politiques. *Recherche Data Gouv*, un écosystème au service du partage et de l'ouverture des données de recherche a été pensé pour soutenir les équipes de recherche dans leur travail de structuration des données pour les rendre conformes aux principes « FAIR ». *Recherche Data Gouv* propose notamment une offre pluridisciplinaire de dépôt, publication et de signalement des données de recherche. Cette offre complémentaire des entrepôts thématiques offre aux domaines scientifiques dépourvus d'entrepôt une solution souveraine pour le partage et l'ouverture de leurs données et ambitionne de signaler les données de recherche partagées ou ouvertes par des entrepôts tiers.
Afin, d'une part de guider les équipes de recherche vers l'entrepôt le plus adapté pour le partage et l'ouverture des données de leur domaine thématique et d'autre part développer le catalogue de données de *Recherche Data Gouv* qui fédérera les données disponibles dans des entrepôts tiers, il est indispensable d'identifier les entrepôts thématiques nationaux et internationaux de confiance, certifiés Core Trust Seal ou non.
Dans ce cadre, il est nécessaire de se doter d'une liste de critères appropriés pour sélectionner les entrepôts thématiques de confiance, qui pourront non seulement permettre les dépôts et la publication des jeux de données, mais aussi concourir à leur diffusion et leur réutilisation ultérieure par les communautés scientifiques.

Pour répondre à cette perspective, je souhaite confier les missions suivantes au Collège des données de la recherche du Comité pour la science ouverte :

1/ Dans le cadre des travaux relevant de son périmètre :

- Sur la base des travaux déjà engagés, vous établirez la liste de critères appropriés permettant 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 ;
- Vous proposerez les modalités de suivi et d'évolution de ces critères ;
- En vous appuyant sur un cercle d'experts que vous aurez préalablement identifiés, vous contribuez à identifier les besoins fonctionnels d'un outil d'aide à la sélection des entrepôts thématiques à destination des équipes de recherche ;
- Vous contribuerez à l'élaboration de supports pédagogiques pour accompagner les équipes de recherche dans l'usage de cet outil ;
- Vous proposerez des scénarios de pérennisation du projet, intégrant les évolutions futures du paysage des entrepôts.

2/ En coordination étroite avec l'équipe en charge du développement de *Recherche Data Gouv* :

- Vous proposerez une liste initiale d'entrepôts de confiance « moissonnables » par le module catalogue de *Recherche Data Gouv* issue de la liste établie dans le cadre des travaux du Collège des données de la recherche ;
- Dans le périmètre des entrepôts thématiques, vous participerez à la réflexion sur la mise à jour de la liste des entrepôts de confiance moissonnés par *Recherche Data Gouv* ;
- Vous contribuerez aux réunions de travail sur le catalogue de *Recherche Data Gouv*.

Je vous remercie d'avoir mis en place un groupe de travail et désigné des pilotes pour conduire cette mission (composition du groupe de travail en annexe). Je vous remercie par avance de préciser prioritairement au travers d'une courte note de cadrage les missions, les livrables, les modalités de travail et les différents échéanciers.

Je vous remercie pour l'investissement du Collège des données de la recherche, dont je sais compter sur le plein engagement au service de la communauté de recherche.

Isabelle Blanc