

Sélectionner un entrepôt thématique de confiance

Collège Données de la recherche

Cécile Arènes, Véronique Stoll

Université Grenoble Alpes

01/07/2024

Enjeux de la politique nationale des données

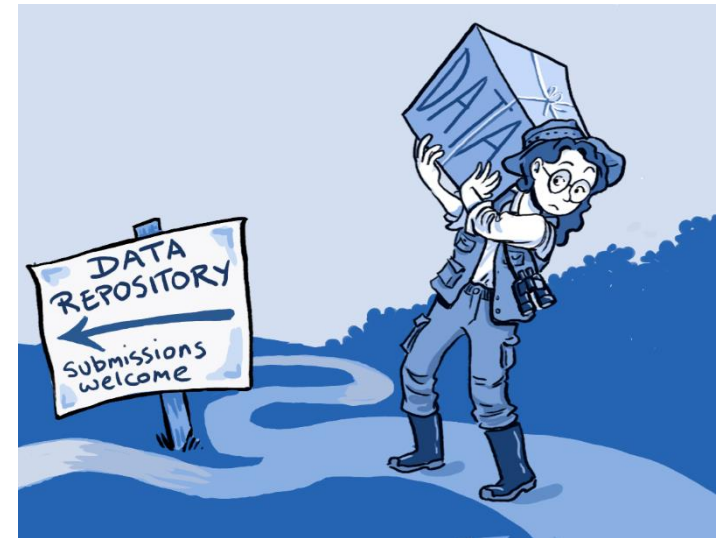


- **Innovation** : ouvrir les données
 - ✓ pour favoriser leur réutilisation par les chercheurs, les enseignants, les citoyens et par les entreprises
 - ✓ pour qu'elles deviennent créatrices de valeur scientifique et économique
- **Confiance** : pour plus de transparence de l'action publique
- **Simplification** : faire de la circulation des données un outil pour simplifier et rendre plus efficaces les actions et les processus administratifs (« Dites-le nous une fois ! »)

Pourquoi partager ses données dans un entrepôt de confiance ?



- Garantir la visibilité des données
- Garantir la qualité des métadonnées pour une bonne « trouvabilité » des données
- Lier publications, données et codes grâce aux identifiants pérennes



Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779.
<https://doi.org/10.1371/journal.pbio.1001779>

Illustration credit: Ainsley Seago

Cadre juridique des données de la recherche



- Loi Valter, 2015 : open data pour les établissements publics
- Loi pour une République numérique, 2016
- RGPD, 2018

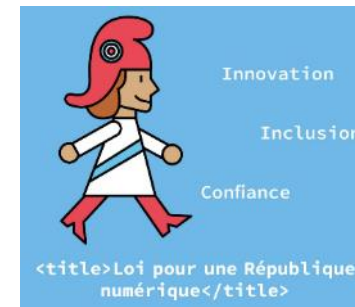
Principes

- Données de la recherche = documents administratifs, c'est-à-dire données publiques
- Open data : ouverture par défaut, lorsque le travail est achevé (à la publication par exemple)
- Des exceptions prévues à l'ouverture : données personnelles, données de santé, secret industriel et commercial, PPST, etc.



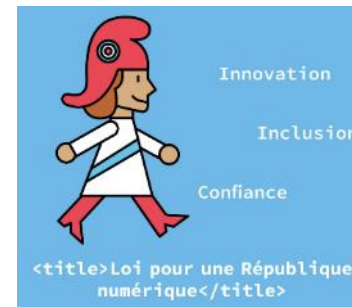
Cécile Arènes, Lionel Maurel, Stephanie Rennes. Guide d'application de la Loi pour une République numérique pour les données de la recherche. Comité pour la science ouverte. 2022. <https://hal-lara.archives-ouvertes.fr/OUVRIR-LA-SCIENCE/hal-03968218v1>

Loi pour une République numérique, article 30



II.- Dès lors que les données issues d'une **activité de recherche financée au moins pour moitié par des dotations de l'Etat**, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne 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, **leur réutilisation est libre.**

Loi pour une République numérique, article 30



III.- L'éditeur d'un écrit scientifique mentionné au I **ne peut limiter la réutilisation des données de la recherche** rendues publiques dans le cadre de sa publication.

Reproductibilité et intégrité scientifique de la recherche



- **Permettre la vérification des résultats de recherche**
 - Mise en place de data policies chez les éditeurs
- **Un cadre défini par le décret du 3 décembre 2021**
 - La science ouverte et la gestion des données relèvent de l'intégrité scientifique

Préserver les données sur le long terme



20 ans après leur publication, 80% des données seraient perdues.

Perte

... pour les auteurs

... pour les équipes, organismes, institutions

... pour la recherche mondiale

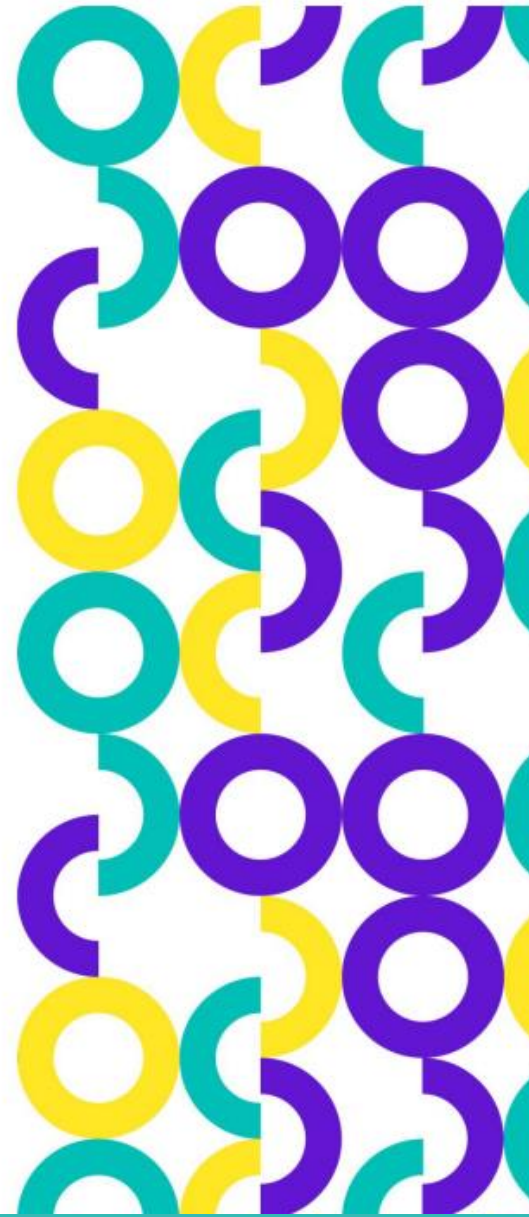
Causes

- Destruction des supports
- Problème de version, obsolescence matérielle ou logicielle
- Lieu de stockage indéfini

Conséquences

- Perte de temps, d'argent
- Pas de vérification des résultats possible
- Pas de comparaison des résultats dans le temps ou l'espace
- Pas de réutilisation

Sélectionner un entrepôt thématique de confiance pour le dépôt de données



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

Collège Données de la recherche

Mars 2024

- Des incitations à l'ouverture des données sans infrastructures clairement identifiées dans la plupart des disciplines :
*« Le choix de l'entrepôt adéquat est délégué au **chercheur**. »*
Rapport 2022 du Conseil des sciences et technologie du gouvernement américain
- L'existence de catalogues ne permettant pas d'isoler efficacement les entrepôts pratiquant **l'auto-dépôt**, indépendamment de l'institution d'affiliation et selon le type de données ou la discipline
- Demande récurrente des personnels d'appui à la recherche souhaitant fournir une réponse plus efficace et plus éclairée sur le choix des entrepôts

Proposition



**Mise en place d'un outil d'aide à la sélection
d'entrepôts thématiques de confiance répondant
aux problématiques du chercheur**

Trois axes de travail

- ☒ 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.

Déroulé des travaux

- Automne 2022 à avril 2024
- Groupe de travail composé de membres du collège Données de la recherche du CoSO, piloté par Véronique Stoll et Frédéric de Lamotte
 - *Cécile ARENES*
 - *Marie-Emilia HERBET*
 - Stéphane DEBARD
 - Françoise GENOVA
 - Christine HADROSSEK
 - Frédéric de LAMOTTE
 - Emilie LERIGOLEUR
 - Gaëlle LEROUX
 - Gilles OHANESSIAN
 - Christelle PIERKOT
 - Marie STAHL
- Appui de deux membres extérieurs
 - Thomas JOUNEAU
 - Maria-Grazia SANTANGELO

Un entrepôt thématique ?

« 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 sources

- 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) ou liste d'entrepôts certifiés ;
- Les plateformes disciplinaires dédiées à la gestion des données de recherche (comme le consortium allemand NFDI, Dataacc.org, Elixir) proposant un premier recensement d'entrepôts thématiques ;
- Des retours d'expérience de la communauté scientifique.

Un choix effectué après examen d'autres démarches comparables



- [RDA descriptive attributes](#):

Machine Interoperability (OAI-PMH, FTP, REST, SPARQL, SWORD, OpenDAP, RDFa)

Schéma de métadonnées : Dublin Core, NetCDF, DDI, etc.

Licences

- [NIH desirable characteristics](#):

Common format

Provenance

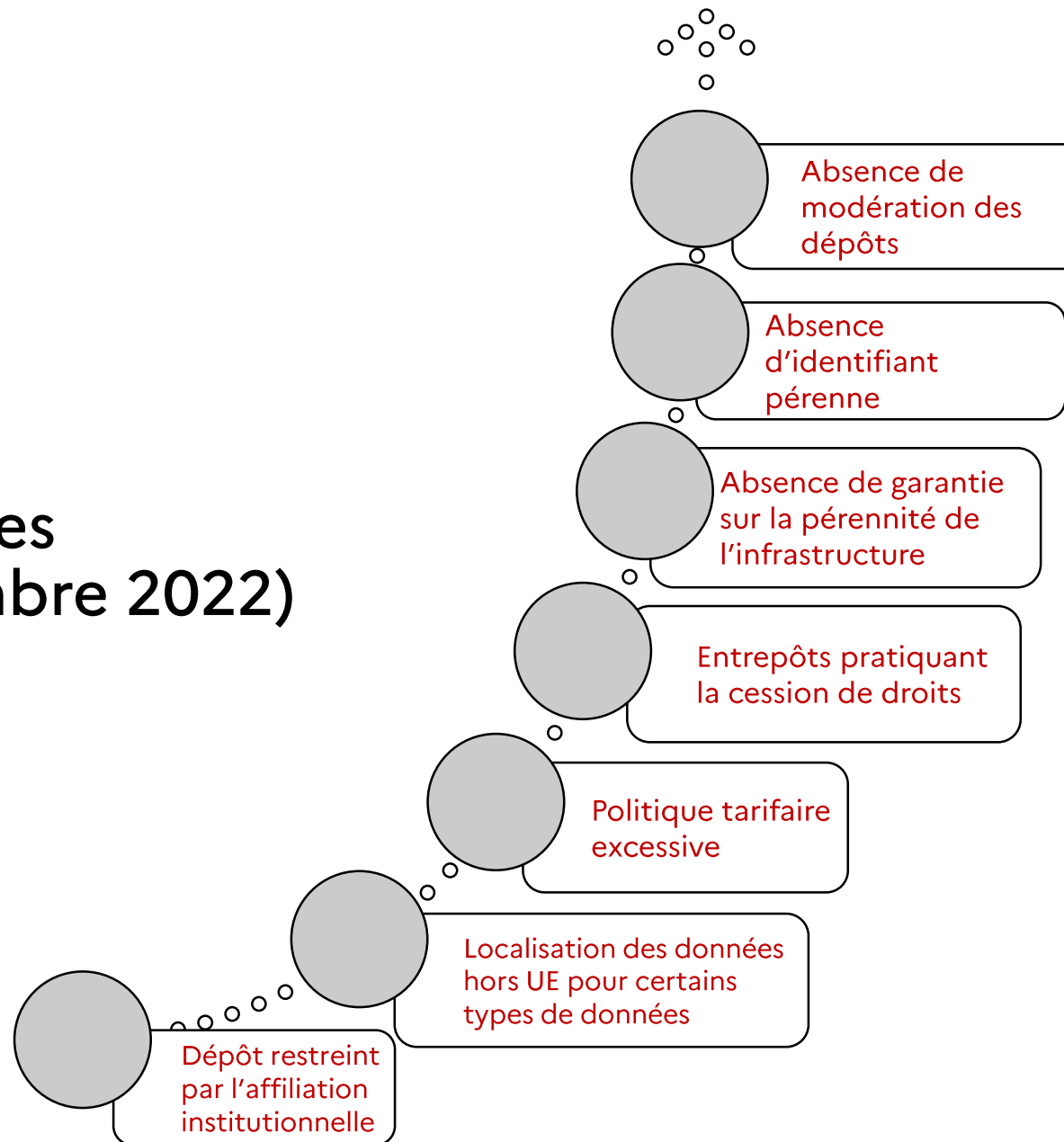
Ce que la liste ne comprend pas

- Une analyse technique des protocoles de moissonnage et API des entrepôts recensés
- Une analyse exhaustive du degré de performance des entrepôts pour chaque critère FAIR

Méthodologie



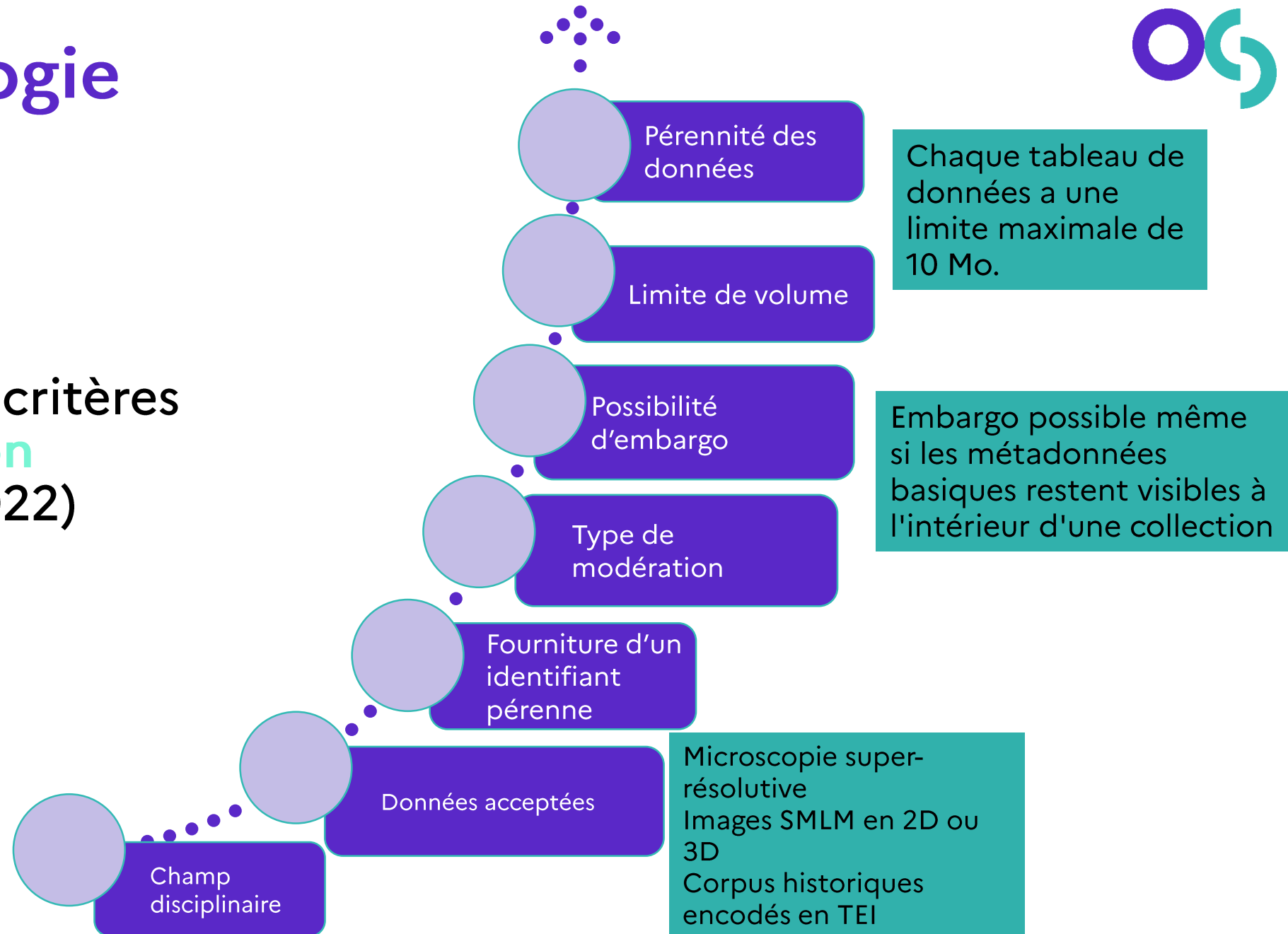
- Définition de critères d'**exclusion** (décembre 2022)



Méthodologie



- Définition de critères de **description** (décembre 2022)



Première liste, évolutive



Nom	URL	Institution porteuse	Modération	Identifiant pérenne fourni	Pérennité de l'infrastructure/Engagement sur la durée de préservation des données	Discipline	Champs disciplinaires détaillés	Données acceptées = mots-clés qui parlent à la	Embargo	Limite de volume
ArkeOpen	https://arkeopen.org/	Université de Strasbourg, Huma-Num	ArkeoGIS contrôle les formats d'import et la conformité des métadonnées descriptives via le protocole mis en place :	DOI	Hébergement assuré par Huma-Num	Archéologie	Archéologie, Histoire, Environnement	Données archéologiques géoréférencées	Oui	Pas de limite, toutefois l'équipe déconseille les jeux de plus de 10 000 lignes.
Cambridge Structural Database (CSD)	https://www.csd.c.cam.ac.uk/	CCDC et FIZ Karlsruhe	Automatisée (checkCIF) et humaine	DOI	Existe depuis les années 70	Chimie	Cristallographie	Structures cristallines (molécules organiques et	Oui	50 Mo par fichier, 100 Mo par versement
Centre de données astronomiques de Strasbourg (CDS)	https://cds.u-strasbg.fr/	Université de Strasbourg et CNRS	Les données déposées doivent être associées à une publication à comité de lecture ayant été acceptée.	Identifiants spécifiques à la discipline: numéro pour les catalogues, identificateur de relevé du ciel pour les collections d'images. Depuis quelques années, le CDS tente d'associer en plus un DOI à chacune de ces	Existe depuis 1972. Référence internationale dans son domaine. Certifié CoreTrustSeal en 2019, renouvellement en cours.	Astronomie	Astronomie	Catalogues (ou tables) astronomiques ; relevés images du ciel	Pas d'embargo en dehors de l'embargo standard d'un an pour les données produites dans les observatoires	Non
Chemotion	https://www.chemotion-repository.net/welcome	NFDI4Chem (consortium allemand)	Une équipe de 10 reviewers assure la modération "manuelle" des données. Il y a aussi une modération automatique pour certaines données spectrales. Plus d'informations ici : https://chemotion.net/docsrepo/workflowreview	DOI	L'entrepôt est intégré à l'initiative Science Data Center du Baden-Württemberg (https://mwk.baden-wuerttemberg.de/de/dienstleistungen/pressenachrichten/pid/vier-science-data-centers-in-baden-wuerttemberg/) et sera prochainement considéré comme une infrastructure nationale de recherche en Allemagne. L'accès aux données est garanti pour au moins 10 ans. Un plan de stockage de long terme et d'archivage sera développé dans ce cadre sous l'égide du Centre de calcul de Steinbuch. https://chemotion.net/docsrepo/	Chimie	Chimie organique	Molécules, réactions chimiques, données analytiques (spectres) protocoles et conditions d'expérimentation. Les résultats négatifs sont acceptés (réactions ayant échoué)	Oui	Pas de limite sur la taille de fichiers, éventuellement 50 Mo à l'avenir. Pas de limite sur le volume total par utilisateur, sauf cas de mésusage détecté.
Centre de Données Socio-Politiques (CDSP)	https://cdsp.sciences-po.fr/fr/	Sciences Po et CNRS	Le dépôt est très encadré et se fait avec l'aide d'ingénieurs du CDSP qui gèrent la curation. Voir : https://cdsp.sciences-po.fr/fr/deposer-des-donnees	DOI	Existe depuis 2005	Sciences humaines et sociales	Sciences politiques, sciences sociales	Résultats d'enquêtes nationales ou internationales, quantitatives ou	Possible en principe, ne s'est pas produit en pratique	Pas de limite formelle, les données quantitatives étant en général peu volumineuses
COLlections de Corpus Oraux Numériques (CoCoON)	https://cocoon.huma-num.fr/	UMR LACITO (Sorbonne Nouvelle Paris 3, INALCO, CNRS) et UMR LLL (Universités d'Orléans, Université de Tours, BnF, CNRS)	Vérification automatique et humaine. Syntaxe contrôlée par des schémas XML (modèle OLAC, sur-couche du Dublin-Core qualifié) et par shematron défini pour ajouter leurs propres restrictions et spécificités. Chaque demande de publication passe par une relecture des métadonnées par un administrateur. Vérification de la normalisation des champs, la complétude, la précision et la cohérence de la description. C'est lors de cette vérification qu'un alignement de certains champs sur des référentiels externes (personnes, lieux, langues, sujets, structures) est effectuée. Quelques vérifications humaines supplémentaires peuvent être faites pour s'assurer de la qualité des vidéos versées.	Identifiants OAI et DOI. Si la ressource fait l'objet d'un archivage, un identifiant ARK est attribué. Lorsque la ressource est moissonnée par ISIDORE, un identifiant HANDLE est attribué.	Démarche d'archivage pérenne (voir remarques) et partenariat Huma-Num, CINES, CC-IN2P3	Sciences humaines et sociales	Linguistique, Anthropologie, Ethnomusicologie, Histoire	Enregistrements sonores, métadonnées descriptives	Oui (via l'accès restreint)	En fonction des formats : WAV (entre 2 et 4 Go), FLAC, pour les fichiers plus volumineux, mais pas formellement précisé
Conservatoire National des Données 3D	https://3d.humanities.science/	Archeovision, Consortium "3D for Humanities", CNRS	Les métadonnées doivent être conformes au schéma déposé par le Consortium 3D SHS. Le contrôle de la qualité des métadonnées est assuré par utilisation obligatoire de l'ATAG3D. Échanges	DOI	Données hébergées chez Huma-Num	Archéologie	Archéologie	Modélisations, restitutions 3D d'artefacts, de	Par défaut les métadonnées sont publiques et le dépôt lui-même	Pas de limite

Répartition disciplinaire des entrepôts sélectionnés

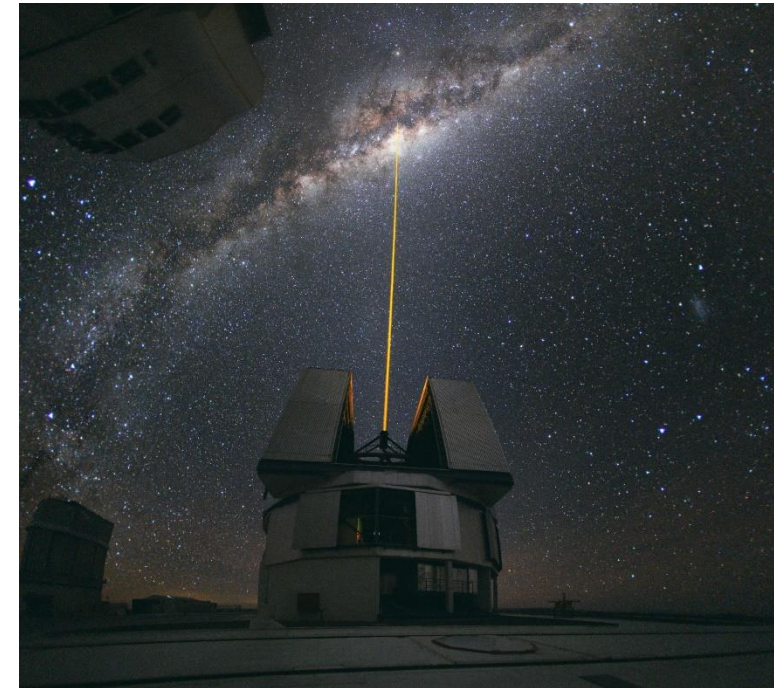


Domaine	Nombre d'entrepôts
Astronomie	2
Biologie	13
Chimie	9
Physique	2
Sciences de l'environnement	7
Sciences humaines et sociales	16 (dont 6 en archéologie et 3 en linguistique)

Astronomie



- Une discipline très organisée depuis des années
- Des habitudes d'alimentation des entrepôts par des personnels support et non par les chercheuses et chercheurs eux-mêmes
- Deux entrepôts de données retenus car ils permettent l'auto-dépôt

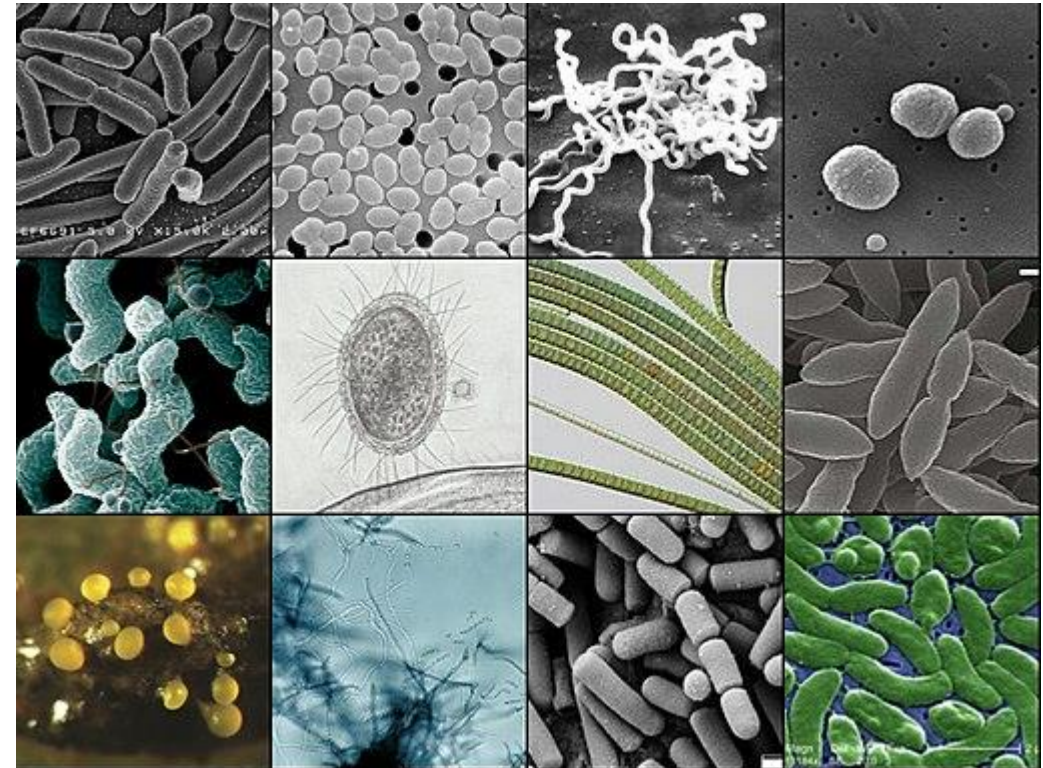


ESO/Yuri Beletsky (ybialets at eso.org), CC BY 4.0.
https://commons.wikimedia.org/wiki/File:Laser_Towards_Milky_Ways_Centre.jpg

Biologie



- Une offre large
- Une ancienneté robuste des entrepôts existants
- Sur-représentation des entrepôts situés dans des pays anglo-saxons
- Mention spécifique du lieu de conservation des données (Europe ou non) pour ces entrepôts



148LENIN, CC BY-SA 4.0.
https://commons.wikimedia.org/wiki/File:Bacteria_collage.jpg

- Des entrepôts thématiques récents, à l'exception de la cristallographie
- Répondent globalement aux critères (avec une vigilance à garder sur la question de la pérennité)
- Entrepôts relativement spécialisés en fonction d'un type de données



Joe Sullivan, CC BY 2.0.
https://commons.wikimedia.org/wiki/File:Chemicals_in_flasks.jpg

Physique



- Discipline où la production de données peut être massive et un paysage d'entrepôts encore peu développé
- Offre restreinte en ce qui concerne l'auto-dépôt de données
- Deux entrepôts signalés



Daniele Pugliesi, Public domain, via Wikimedia Commons.
<https://commons.wikimedia.org/wiki/File:CollageFisica.jpg>

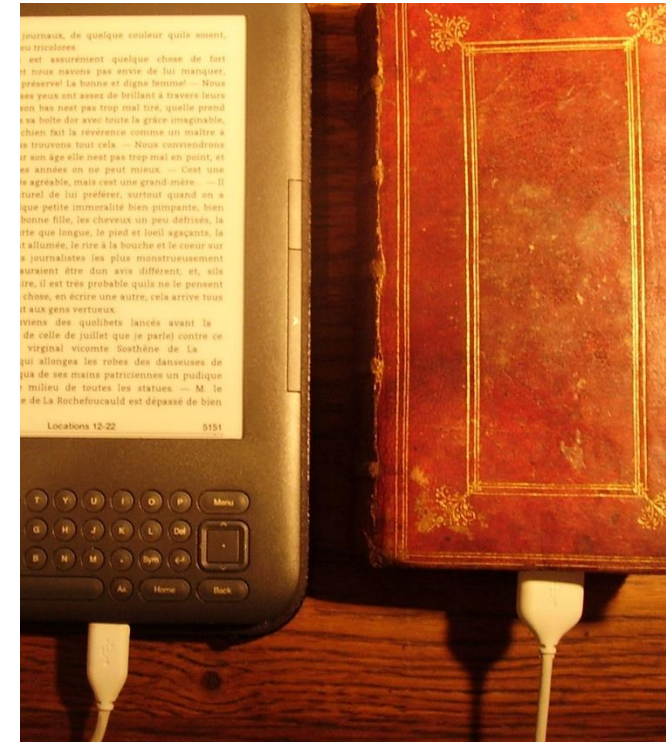
- Politique de modération bien structurée
- Présence d'entrepôts certifiés CTS
- Paysage d'entrepôts francophones assez riche
- Dépôt d'importants volumes de données possible



Andreas Weith, CC BY-SA 4.0.
[https://commons.wikimedia.org/wiki/File:Polar_bear_\(Ursus_maritimus\)_in_the_drift_ice_region_north_of_Svalbard.jpg](https://commons.wikimedia.org/wiki/File:Polar_bear_(Ursus_maritimus)_in_the_drift_ice_region_north_of_Svalbard.jpg)

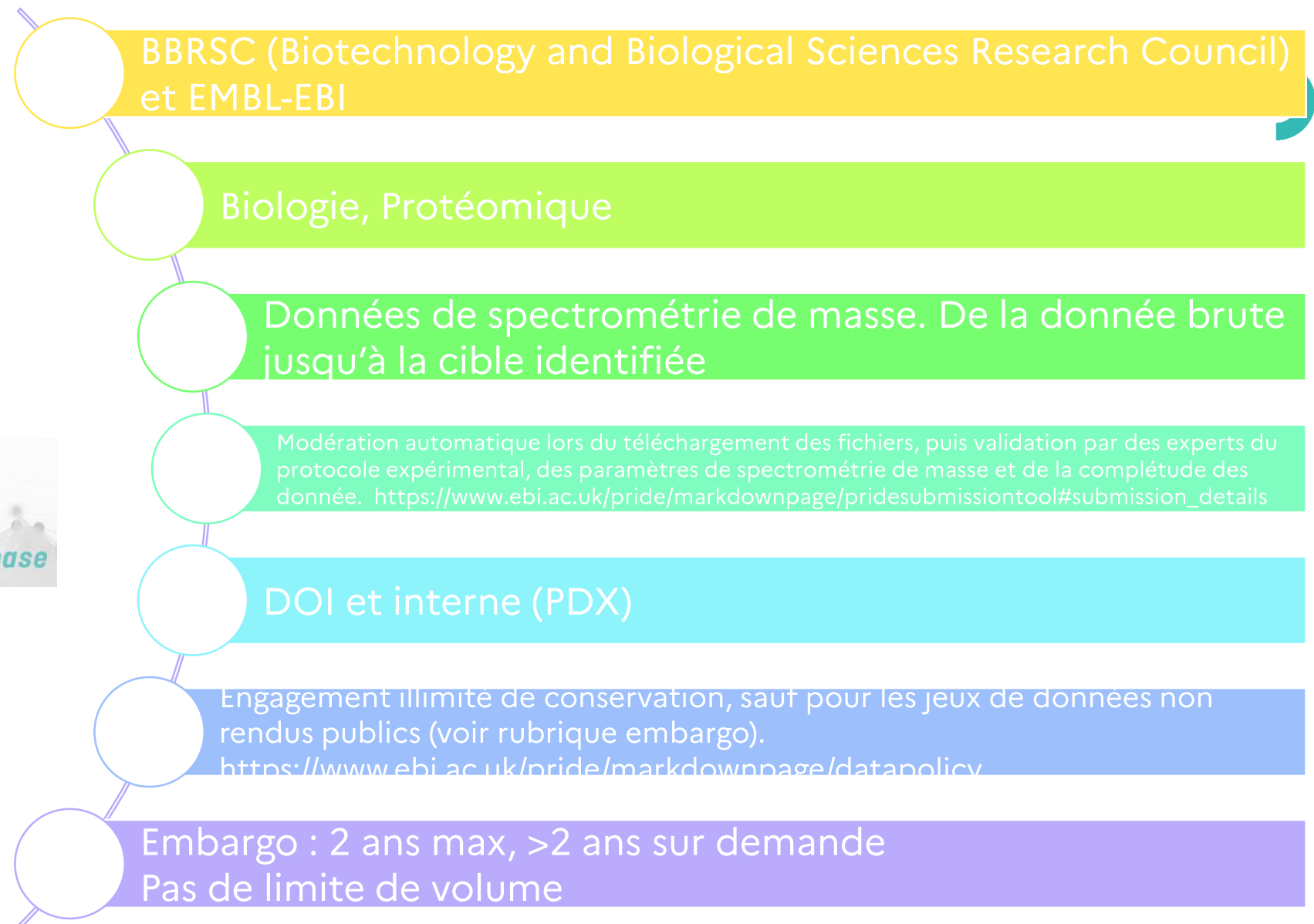
Sciences humaines et sociales

- 16 entrepôts recensés, d'un entrepôt généraliste pour les SHS à des entrepôts thématiques très pointus
- Politique de modération très différente en fonction de la discipline
- Un paysage relativement jeune



Rémi Mathis, 2004. CC:BY-SA. Flickr.
<https://www.flickr.com/photos/remimathis/5239482334/>

Exemple de notice d'entrepôt

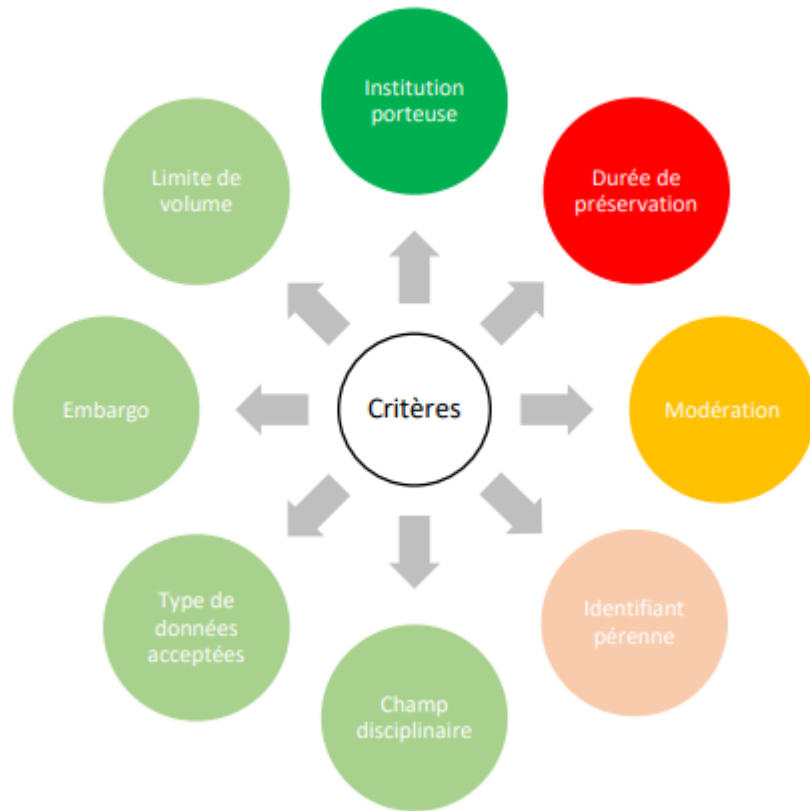


Remarques : les jeux de données sont régis par la licence CC0 depuis 2018.

Source complémentaire : Bhargava, Shruti, et Joachim Jankowski. « The PRIDE Database Resources in 2023 ».

<https://doi.org/10.1093/ndt/gfad176>.

Obtenir les informations



Des difficultés pour :

- Identifiants pérennes
- Modération
- Durée de préservation

Prochaine étape pour le collège Données de la recherche



Poursuite de la réflexion sur la stratégie de pérennisation de ce travail sur les entrepôts de confiance

- Révision périodique des critères
- Révision de la liste d'entrepôts



<https://www.dnb.de/EN/Professionell/Erhalten/LZA-System/lza-system.html>

Premiers enseignements tirés de ce travail



- **Appréciation différente de certains critères** comme la modération : modération a minima de quelques métadonnées ou modération des jeux eux-mêmes selon les disciplines
- **Caractère parcellaire des informations disponibles** publiquement sur les sites des entrepôts
- **Fragilité du paysage des entrepôts** : financements sur projets, par définition non pérennes, pour beaucoup d'entre eux

Implémentation sur le site Recherche Data Gouv



Archéologie

+

Collections de COrpus Oraux Numériques (CoCoON)



Disciplines

Sciences humaines et sociales

Institutions porteuses

« Langues et Civilisations à Tradition Orale » UMR LACITO 7107 du CNRS

Laboratoire Ligérien de Linguistique UMR LLL (7270 CNRS-Université)

Champs disciplinaires

Anthropologie

Ethnomusicologie

Histoire

Linguistique

Données acceptées

Enregistrements sonores

Métadonnées descriptives

Identifiants pérennes

DOI

ARK

Handle (via Isidore)

[Aller sur l'entrepôt →](#)

Modération assurée par l'entrepôt

La vérification est automatique et humaine. La syntaxe est contrôlée par des schémas XML (modèle ...

[Voir plus →](#)

Embargo

Oui (via l'accès restreint). Voir :
<https://cocoon.huma-num.fr/exist/crdo/missions.htm>

Limite de volume par fichier

En fonction des formats : WAV (entre 2 et 4Go), FLAC pour les fichiers plus volu...

[Voir plus →](#)

Limite de volume par jeu de données

Pas de limite précisée



[Retour en haut](#)

Merci !

- **Note** Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante
<https://www.ouvrirlascience.fr/selectionner-un-entrepot-thematique-de-confiance-pour-le-depot-de-donnees-methodologie-et-analyse-de-loffre-existante/>
- **Liste** https://www.ouvrirlascience.fr/wp-content/uploads/2024/04/ListedesEntrepotsdeConfiance_v1_202403.xlsx et <https://recherche.data.gouv.fr/fr/entrepots>

Nous contacter :

entrepots-confiance@groupe.renater.fr