Ouvrir la science !

Choisir un entrepôt de confiance

Mise en pratique

Véronique Stoll & Frédéric de Lamotte (pilotes du Collège des Données de la recherche)

D’après un travail coordonné par les formidables Marie-Émilia Herbet & Cécile Arènes

Poitiers

juin 2024

https://orcid.org/0000-0002-4118-3574 &. https://orcid.org/0000-0003-4234-1172

Recherche Data Gouv,
Zenodo
• et •
FigShare

Ouvrir la science !

- Introduction
  - Contexte
  - Objectifs
- Méthodologie
  - Définitions
  - Sources
  - Critères
- Mise en pratique
  - Nous
  - Vous
  - Tous

Ouvrir la science !

Contexte :

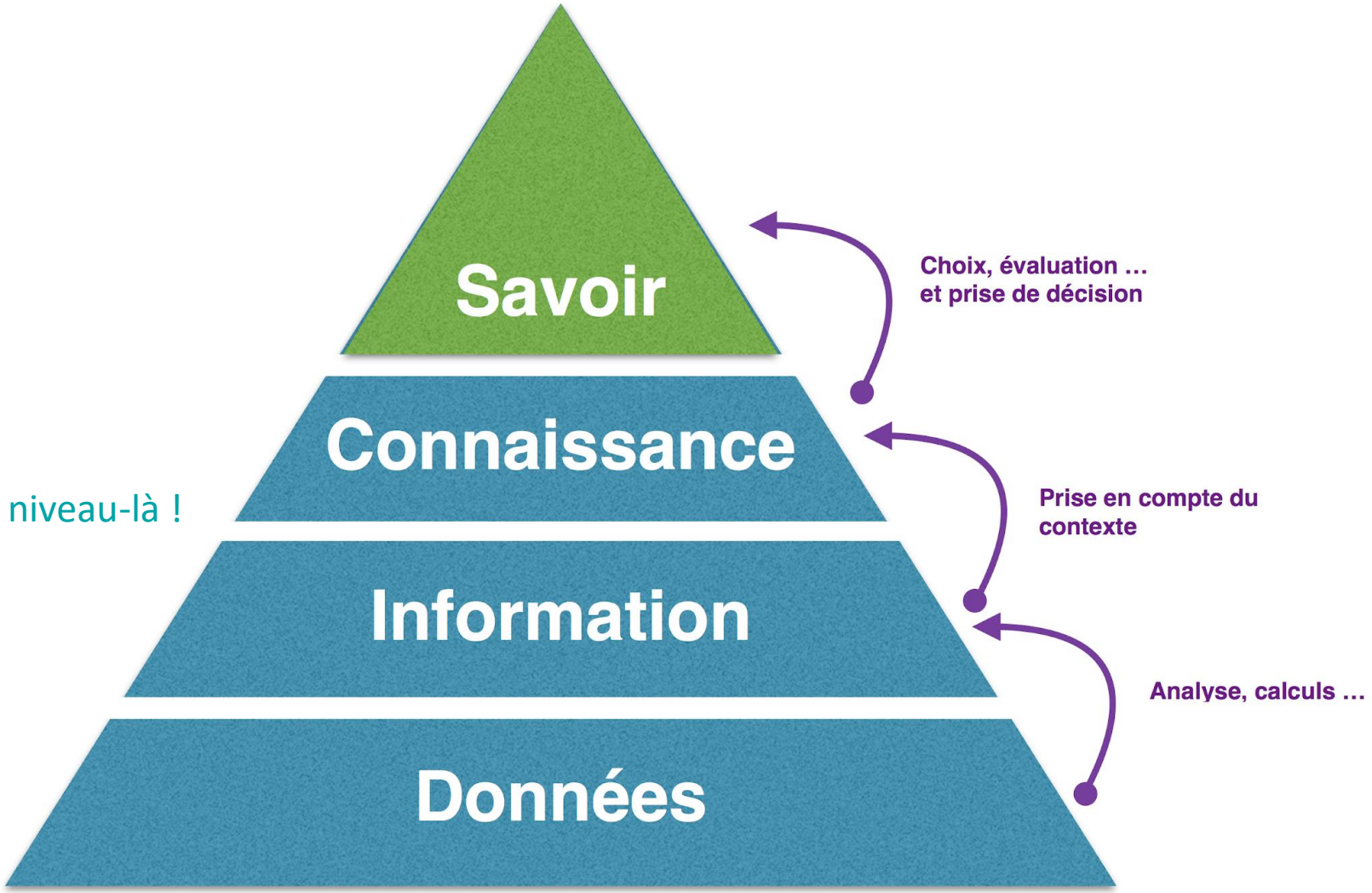
PNSO 2 ***(2021-2024) :*** **Généraliser la SO en France**

Deuxième axe :

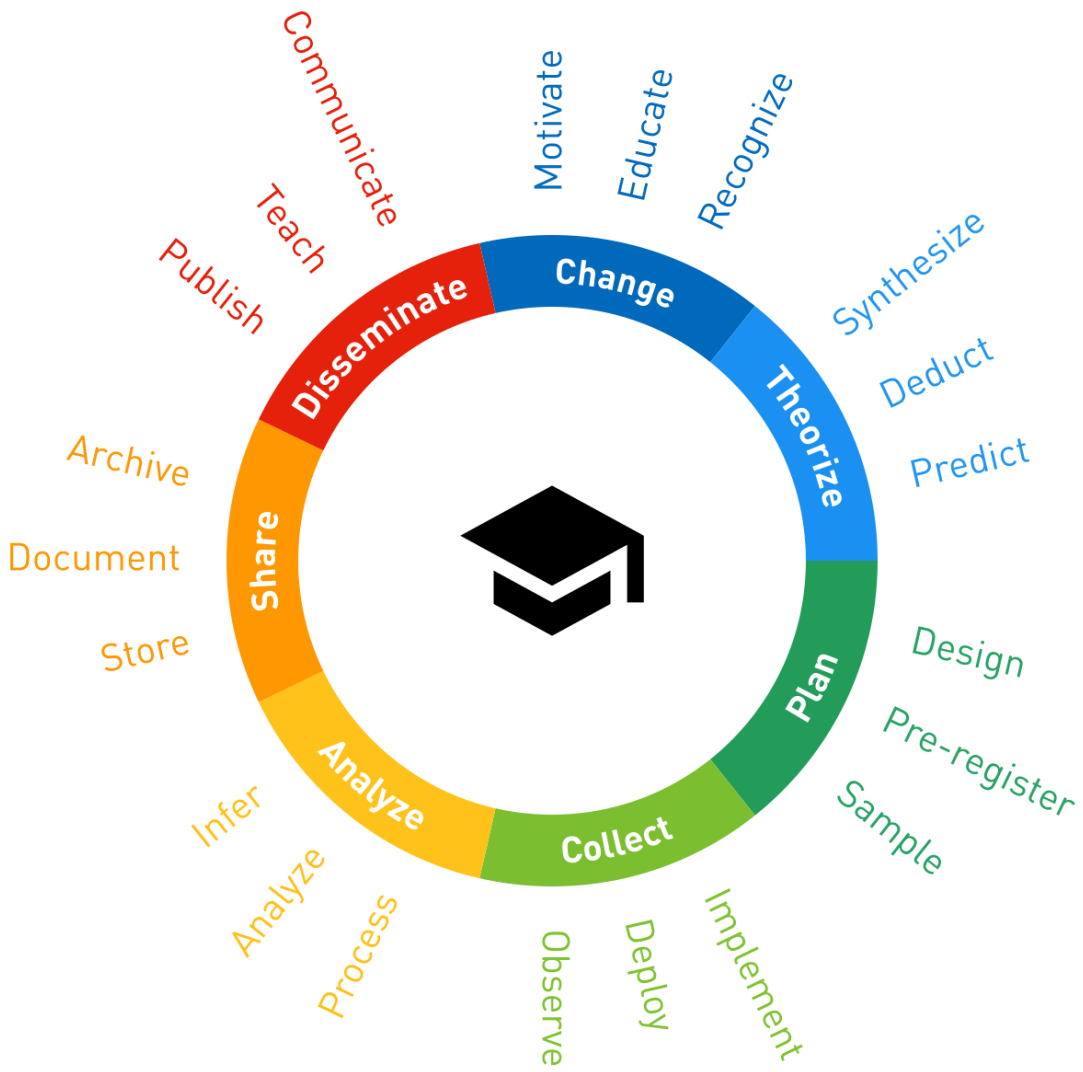
Structurer, Partager et Ouvrir les données de la recherche

Le partage des données (structurées) est une nécessité impérieuse

Les publis, c'est à peu près à ce niveau-là !



Entrepôts et Cycle de Vie

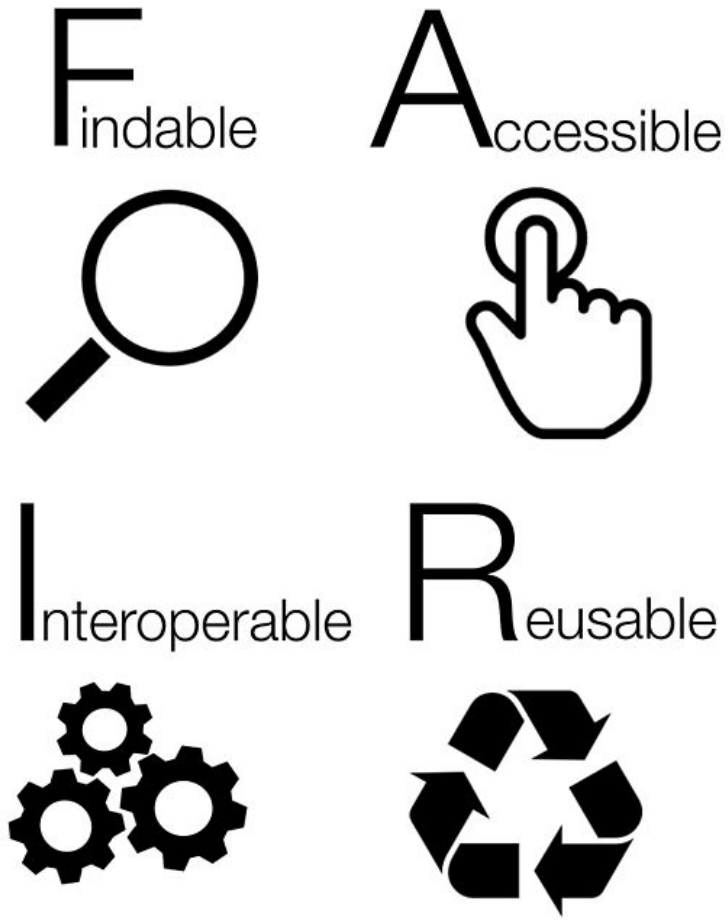


A votre avis,

où se positionnent les entrepôts ?

https://how-to-open.science

Entrepôts et FAIR



A votre avis,

où se positionnent les entrepôts ?

Entrepôt : une définition

https://www.museumtv.art/artnews/articles/rene-magritte-ceci-nest-pas-une-pipe/

Ni sauvegarde

Ni archivage

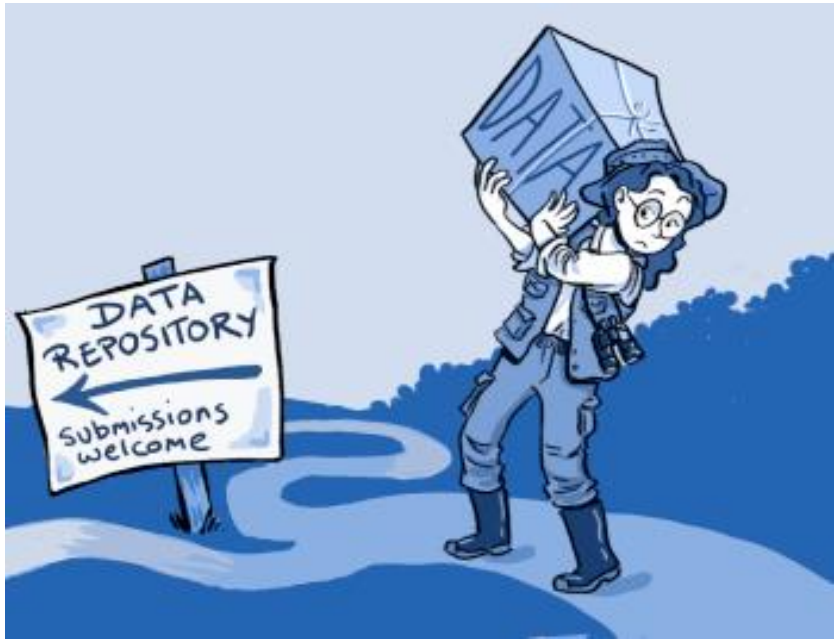
Partage

Exposition

Communication

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. »



Une triple problématique

- Du point de vue du chercheur : où déposer les données ?
- Du point de vue des équipes d'appui à la recherche : comment orienter efficacement ?
- Du point de vue de Recherche data gouv : quels entrepôts de confiance moissonner ?

Groupe de travail

- Lettre de mission mars 2023
- Aide à la sélection d'entrepôts thématiques de confiance répondant aux problématiques du chercheur
- Membres du Collège des données et experts extérieurs (UMR CNRS, DDOR, IRD, Data Terra…)
- GT piloté par Cécile Arènes (SU) et Marie-Emilia Herbet (Lyon 3)
  - Avec : Stéphane DEBARD, Françoise GENOVA, Christine HADROSSEK, Emilie LERIGOLEUR, Gaëlle LEROUX, Gilles OHANESSIAN, Christelle PIERKOT, Marie STAHL

3 Axes de travail

- Liste des critères de confiance
- Note méthodologique et première liste
- Stratégie de pérennisation

Approche cumulative

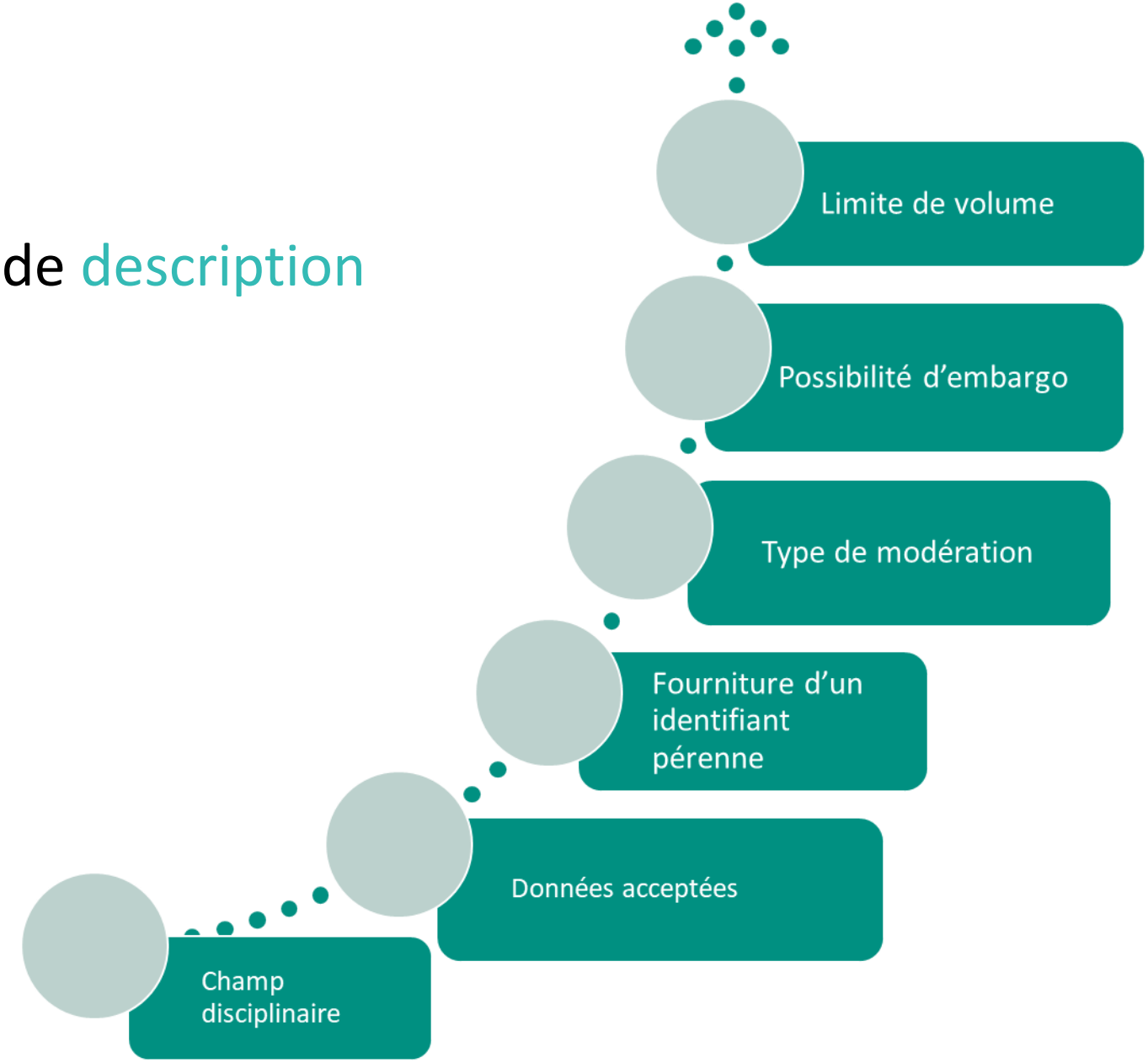
- Sources d'identification des entrepôts :
  - ✓ littérature scientifique/grise
  - ✓ annuaires d'entrepôts (CatOpidor, Re3data, Fairsharing, Opendoar)
  - ✓ plateformes disciplinaires dédiées à la gestion des données de recherche (Datacc.org, Elixir, NFDI…)
  - ✓ retours de la communauté scientifique…
- Articulation avec d'autres travaux existants (RDA Data alliance repository attributes WG, NIH…)
- Définition de critères d'exclusion
- Définition de critères de description
- Fiches descriptives
- Circuit de validation

Méthodologie

- Définition de critères d’exclusion (décembre 2022)

Méthodologie

- Définition de critères de description (décembre 2022)



Critères orientés « chercheurs »

- **Données acceptées** : termes parlants pour les communautés

Ex : microscopie super-résolutive. Images SMLM en 2D ou 3D ; corpus historiques transcrits et encodés en TEI, etc.

- **Embargo** : Oui/Non ou précision.

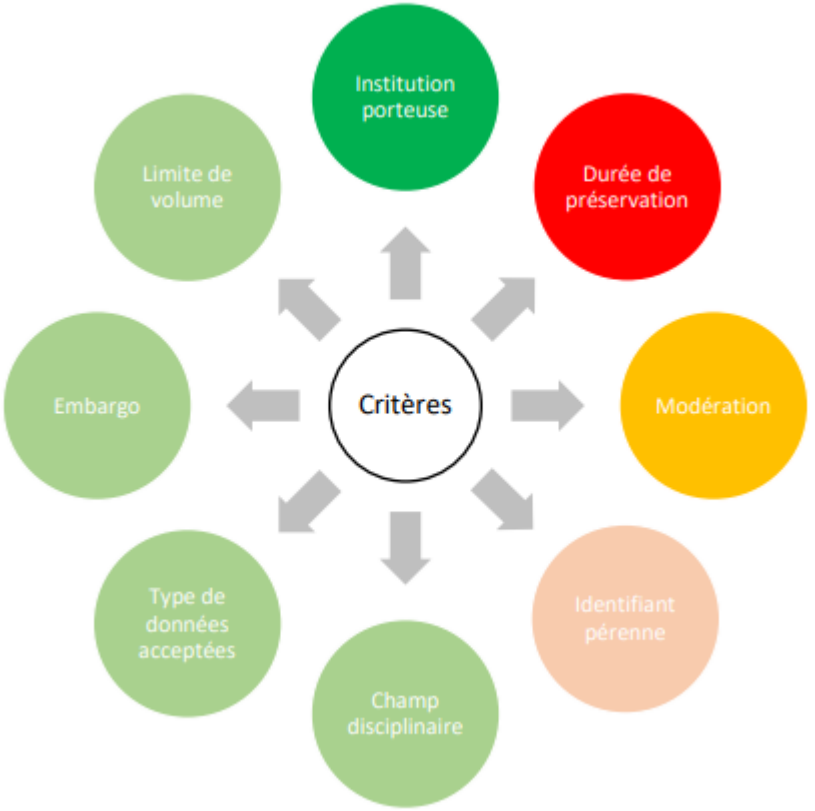
Ex : embargo possible même si les métadonnées basiques restent visibles à l'intérieur d'une collection

- **Limite de volume**

Ex : la taille maximale de téléchargement est de 50 Mo pour un fichier d'archive. Les jeux de données doivent être formatés dans un format tabulaire spécifique (basé sur YAML) et chaque tableau de données a une limite maximale de 10 Mo.

Obtenir les informations

- Des difficultés pour :
  - Identifiants pérennes
  - Modération
  - Durée de préservation



Testons !

Discipline	*Critère interne au groupe, champ disciplinaire large (ex. SHS)*
Nom de l'entrepôt	
URL	
Institution porteuse	
Modération	Détailler les types de modération si possible
Identifiant pérenne fourni	Indiquer le type d'identifiant fourni
Pérennité de l'infrastructure/Engagement sur la durée de préservation des données	Si cette information n'est pas communiquée, se baser sur l'ancienneté de l'entrepôt, ses financements actuels, etc.
Champ disciplinaire	Détail des disciplines acceptées (par exemple, au sein des SHS, histoire, linguistique, etc), s'appuyer sur la nomenclature si possible
Données acceptées	A décrire avec des mots-clés parlants pour la communauté, éviter les descriptions trop générales si possible
Embargo	Détailler les types d'embargo et les durées proposés si possible
Limite de volume	
Remarques	Toute information utile pour les déposants
Source (re3data, fairsharing, catopidor, Opendoar, littérature scientifique etc.) Contact : nom du contact, mail et fonction	Source des informations fournies, qu'il s'agisse du site web de l'entrepôt ou d'autres sites (re3data, fairsharing, catopidor, Opendoar, littérature scientifique etc.) Si contact pour obtenir des précisions : préciser le nom, le mail et la fonction de la personne contactée. Ces informations serviront pour alimenter une liste interne d'experts à contacter si besoin
Nom du curateur	*Critère interne au groupe*

Testons !

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 communauté	Embargo	Limite de volume
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 ressources	Existe depuis 1972. Référence internationale dans son domaine. Certifié CoreTrustSeal en 2019, renouvellement en cours.	Astro	Astro	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

Testons !

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 communauté	Embargo	Limite de volume
Zenodo	https://zenodo.org	CERN		DOI	8 Mai 2013 pour 20 ans au moins …	Tout	Tout	Tout	Oui	100 fichiers / 50 Go

Testons !

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 communauté	Embargo	Limite de volume
Zenodo	https://zenodo.org	CERN	**ben non !**	DOI	8 Mai 2013 pour 20 ans au moins …	Tout	Tout	Tout	Oui	100 fichiers / 50 Go

Testons !

NCBI vs ENA (même focus, ENA en Europe)

Dépôt restreint à une communauté (pays, université …)

Dépôt lié à une revue 😢

A vous de jouer !

5 Groupes

Choix d’un ou deux entrepôts (pas de doublons !)

30 minutes

https://www.ebrains.eu/data/share-data/

https://vectorbase.org/vectorbase/app/

https://ukdataservice.ac.uk/deposit-data/

https://bequali.fr/fr/les-enquetes/

Restitution en commun

10 minutes par groupe

A nous de jouer !

Quel entrepôt aviez-vous ?

Analyse complète ?

Points sensibles ?

Ouvrir la science !

Contact : entrepots-confiance@groupes.renater.fr

Merci

Envoyer votre analyse à

Veronique.stoll@obspm.fr