Zenodo, Figshare, etc.: rechercher et partager des données de recherche grâce aux entrepôts de données

Module 3.4



Programme

La mise à disposition des données associées aux articles ou données sous-jacentes (underlying data) est une demande croissante de la part des éditeurs et des agences de financement. Les entrepôts de données permettent de répondre à ces exigences, en assurant l'accessibilité, l'archivage et/ou la diffusion des données déposées, en conformité avec les principes FAIR. Cette session présente les différents types d'entrepôts de données et des recommandations pour identifier un entrepôt adapté.

- →Introduction
- →Trouver un entrepôt de données généraliste
- →Trouver un entrepôt de données spécialisé



Introduction

Terminologie : données de recherche Modes de partage des données de recherche Enjeux Entrepôts de données et principes FAIR

Terminologie : données de recherche

→Définition large OCDE

Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche.

→ Typologie

- > Données d'observation
- > Données expérimentales
- > Données de références
- > Données de simulation numériques
- > Données dérivées ou compilées

Terminologie : données de recherche

☑ Données d'observation

Capturées en temps réel, souvent uniques, impossibles à reproduire

- Relevés météorologiques
- Images
- Enquêtes sociales
- Fouilles archéologiques

☑ Données expérimentales

Obtenues à partir d'équipements de laboratoires, souvent reproductibles, parfois coûteuses

- Poids biomasse
- Séquences peptide
- Chromatogrammes

- Séquences de gènes
- Structures chimiques

Données de simulation numérique

Générées par des modèles informatiques, souvent reproductibles si le modèle est correctement documenté

- Modèles climatiques
- Modèles économiques

✓ Données dérivées ou compilées

Issues du traitement ou de la combinaison de données « brutes », souvent reproductibles, mais coûteuses

- Bases de données compilées
- Fouilles de texte

Source : Bracco, L., Bouchet-Moneret, F., & Jouneau, T. (2021). Que sont les données de la recherche ? https://doi.org/10.5281/zenodo.5495210

F. Flamerie

→ Répondre manuellement à des demandes



Exemple: https://doi.org/10.1186/s12888-020-03010-3

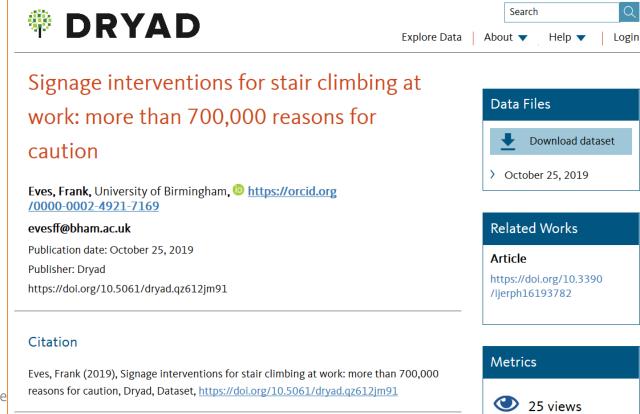
- → Annexer à un article des fichiers de Supplementary data, publiés sur le site de l'éditeur en même temps que l'article
- → Déposer les données dans un entrepôt de données



Exemple: https://doi.org/10.5061/DRYAD.QZ612JM91

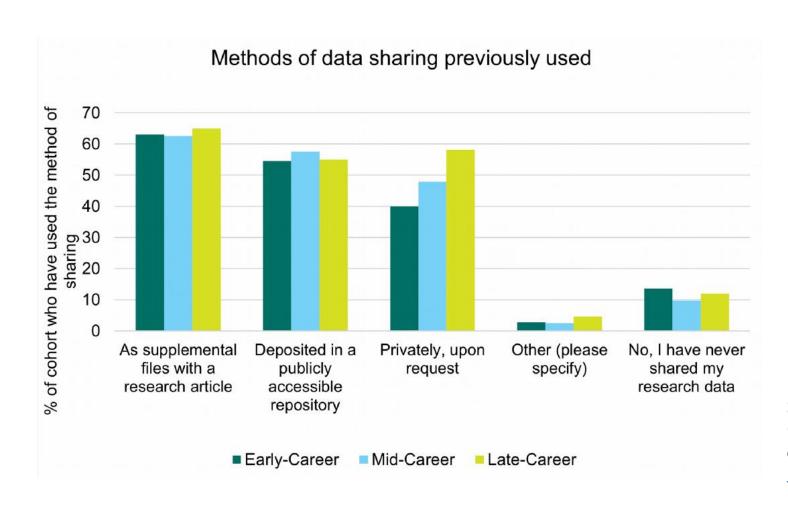
Availability of data and materials

Readers and all interested researchers may contact Runsen Chen (Email address: runsen.chen@psych.ox.ac.uk) for details.



Quel mode de partage des données vous semble-t-il le plus fréquent? - réponse dans le sondage Zoom

- → Fichiers supplémentaires à un article scientifique
- → Dépôt dans un entrepôt de données accessible publiquement
- → Privé, sur demande
- →Autre (précisez dans la discussion Zoom)
- → Aucun partage



Source: Hrynaszkiewicz, I., Harney, J., & Cadwallader, L. (2021). A survey of researchers' needs and priorities for data sharing. OSF Preprints. https://doi.org/10.31219/osf.io/njr5u

- →Enquête menée par PLOS
- →1477 réponses
 - 20,9% Médecine et sciences de la santé
 - 30,4% Biologie et sciences de la vie
 - 9,7% Sciences sociales

"The question was answered by 1150 participants. Sharing data as supplemental files alongside a research paper was the most common method for all career levels (63%), followed by deposition in a public repository (56%) and sharing privately on request (47%). Only 12% of respondents reported that they had never shared their research data - the largest proportion of whom (44%) work in Medicine and Health Science disciplines."

Source: Hrynaszkiewicz, I., Harney, J., & Cadwallader, L. (2021). A survey of researchers' needs and priorities for data sharing. OSF Preprints. https://doi.org/10.31219/osf.io/njr5u

Modes de partage des données de

recherche

NB ces modes de partage ne sont pas exclusifs les uns des autres, des fichiers supplémentaires à un article peuvent être déposés dans un entrepôt de données, parfois par la revue ellemême.

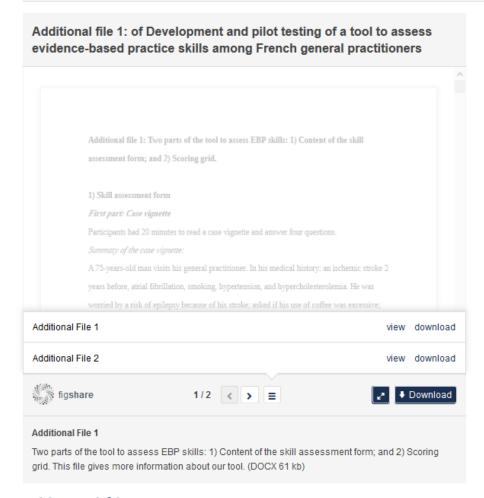
Exemple:

→ Article: Rousselot, N., Tombrey, T., Zongo, D., Mouillet, E., Joseph, J.-P., Gay, B., & Salmi, L. R. (2018). Development and pilot testing of a tool to assess evidence-based practice skills among French general practitioners. BMC Medical Education, 18(1), 254. https://doi.org/10.1186/s12909-018-1368-y



→ Fichiers déposés dans Figshare: Rousselot, N., Tombrey, T., Drissa Zongo, Mouillet, E., Jean-Philippe Joseph, Gay, B., & Salmi, L. (2018). *Development and pilot testing of a tool to assess evidence-based practice skills among French general practitioners* [Data set]. Figshare. https://doi.org/10.6084/Mg.FIGSHARE.C.4297076.V1

Additional files



Additional file 1:

Two parts of the tool to assess EBP skills: 1) Content of the skill assessment form; and 2) Scoring grid. This file gives more information about our tool. (DOCX 61 kb)

Enjeux



F. Flamerie - Données de recherche : entrepôts - 2022-06-28

Enjeux

Stage	Problem	Solution	Helps to know	Agents
Data collection and analysis	Waste of research (duplication)	Preregistration	Who is doing what, how and when?	Researchers to adopt it Journals to enforce it
	Ethical concerns	Preregistration	Is the study in line with good research practices?	
	Flawed studies	Registered reports	Is the study designed appropriately to answer the question of interest?	
Publication process	Expedite reviewing	Open reviews	Are the findings verified by independent researchers?	Researchers to adopt it Journals to enforce it
	Distrust of published results	Data and code sharing	Can scientist obtain similar results from the same data?	Institutions and funding agencies to value it Policymakers to allow it
	Conflict of interests	Disclosure of conflict of interests, including editorial roles	Are the authors incentivized to publish these findings and to publish in this journal?	Researchers to adopt it Journals to enforce it
Communication	Misuse of preprints	Collaboration between journalists and scientists	Can the public trust scientific news in the media?	Journalists and news editors Institutions
	Misleading headlines, exaggerations			
	Paywalled manuscripts	Open Access on all manuscripts	The exact content of manuscripts used as a source. Can accelerate research through universal access to scientific findings	Researchers to favor it Policymakers and institutions to enforce it

Source: Besançon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., Billy, E., Deforet, M., & Leyrat, C. (2021). Open science saves lives: Lessons from the COVID-19 pandemic. *BMC Medical Research Methodology*, 21(1), 117. https://doi.org/10.1186/s12874-021-01304-y

Enjeux - France

2 nouveaux groupes au sein de l'instance nationale <u>Comité pour la science ouverte</u> portée par le MESRI

- → Plans de partage des données issues des essais cliniques
 - > Pilotage CHU Rennes
 - > Optimiser la gestion des données collectées à l'occasion des essais cliniques
 - > Elaborer des modèles de plans de partage des données (data sharing plans) pour favoriser le partage responsable de ces données
- → Portail des études individuelles en santé
 - > Pilotage Iresp
 - Déclarer et caractériser les projets de recherche en santé impliquant des individus, quels que soient leur état d'avancement et leur issue, au travers d'un portail national FReSH (France Recherche en Santé Humaine)

Enjeux - politique de l'ANR

"Est-ce qu'il existe une obligation à l'ouverture des données?

Non. En cas de financement le coordinateur ou la coordinatrice s'engage à fournir un plan de gestion de données. Le PGD a pour but de préparer la diffusion potentielle des données mais il ne constitue pas une obligation à l'ouverture. Le principe « aussi ouvert que possible, aussi fermé que nécessaire » est au cœur de la démarche de l'ANR.

Source: FAQ Plan de Gestion des Données (PGD) de l'ANR

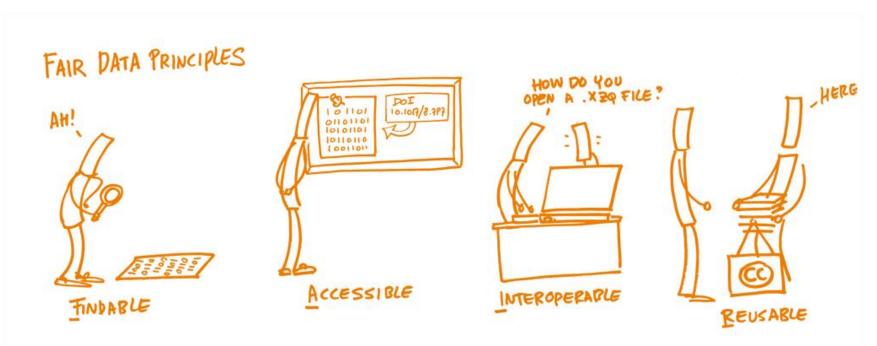
A noter que cette politique est commune au réseau des agences de financement françaises (ADEME, ANR, ANRS-MIE, Anses et INCa)

Entrepôts de données et principes FAIR

Principes FAIR -> les données doivent pouvoir être trouvées, comprises et réutilisées. **Cela ne signifie pas qu'elles doivent être nécessairement en accès entièrement libre.** Ces principes s'appliquent également aux métadonnées associées aux données.

ces principes s'appliquent egalement aux metadonnees associees aux donne

Principes FAIR en détail -> « Les principes FAIR » sur le site DORANUM



Source image: The Open Science Training Handbook

Entrepôts de données et principes FAIR

DOI et citation:

Findable - Accessible

Documentation : Reusable

Une documentation plus structurée et standardisée permettrait d'augmenter les scores Interoperable - Reusable



Search

Explore Data | About ▼ | Help ▼ | Login

Signage interventions for stair climbing at work: more than 700,000 reasons for caution

Eves, Frank, University of Birmingham, @ https://orcid.org/0000-0002-4921-7169

evesff@bham.ac.uk

Publication date: October 25, 2019

Publisher: Dryad

https://doi.org/10.5061/dryad.qz612jm91

Citation

Eves, Frank (2019), Signage interventions for stair climbing at work: more than 700,000 reasons for caution, Dryad, Dataset, https://doi.org/10.5061/dryad.qz612jm91

Abstract

Increased stair climbing reduces cardiovascular disease risk. While signage interventions for workplace stair climbing offer a low-cost tool to improve population health, inconsistent effects of intervention occur. Pedestrian movement within the built environment has major effects on stair use, independent of any health initiative. This paper used pooled data from UK and Spanish workplaces to test the effects of signage interventions when pedestrian movement was controlled for in analyses. Automated counters measured stair and elevator usage at the ground floor throughout the working day. Signage interventions employed previously successful campaigns. In the UK, minute-by-minute stair/elevator choices measured effects of momentary pedestrian traffic at the choice-point (n = 426,605). In Spain, aggregated pedestrian traffic every 30 min measured effects for 'busyness' of the building (n = 293,300). Intervention effects on stair descent (3 of 4 analyses) were more frequent than effects on stair climbing, the behavior with proven health benefits (1 of 4 analyses). Any intervention effects were of small magnitude relative to the influence of pedestrian movement. Failure to control for pedestrian movement compromises any estimate for signage effectiveness. These pooled data provide limited evidence that signage interventions for stair climbing at work will enhance population health.

Methods

Observations from automated counters followed by bootstrapped regressions

Usage Notes

Read me file for UK interventions

Up from ground floor

The coding used to allow analysis of the file with logistic regression is explained below. The dependent variable is stair and escalator choice. If you have any queries, please do not hesitate to contact me directly (evesff@bham.ac.uk).

For ascent, file IJERPHMultiSiteUKUp, the coding is as follows below.

INTSTAGE: 1 =baseline, 2=first intervention, 3=second intervention.

BASEOINT1; the coding allows comparison of the baseline (0) with the first intervention (1).

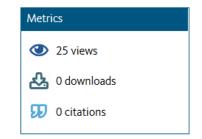
PH2ASE1; the coding allows comparison of the first intervention (0) with the second (1).

Data Files Download dataset October 25, 2019 IJERPHBarce...iSite.xlsx 276.45 kB IJERPHMulti...KDown.xlsx 5.92 MB IJERPHMulti...eUKUp.xlsx 6.48 MB

Related Works

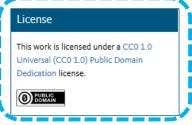
∆rticle

https://doi.org/10.3390/ijerph16193782





Licence CC: Reusable



Critères de choix d'un entrepôt de données

- → L'entrepôt est-il certifié?
 - > En savoir plus sur la certification *CoreTrustSeal* : RDA France. (2019). Entrepôts de données de confiance : Critères de conformité. Repéré à https://www.rd-alliance.org/system/files/documents/CoretrustsealFR.pdf
- → Gère-t-il différents **types d'accès**? Permet-il par exemple un accès restreint ou sous **embargo**?
 - > Y compris la gestion des demandes d'accès pour les données en accès restreint?
 - > Y compris la possibilité de définir le type d'accès fichier par fichier? [cf. exemple ICPSR ci-dessous]
- → Prend-il en charge un **format de données** particulier?
- → Quelles sont les exigences ou possibilités en termes de précision et structuration de la description des données déposées? Devrez-vous ou pourrez-vous fournir des métadonnées riches?
- → Propose-t-il d'autres services particulièrement adaptés à mes besoins? [cf. exemple connexion Zenodo-Github ci-dessous]

ICPSR gestion fine des accès

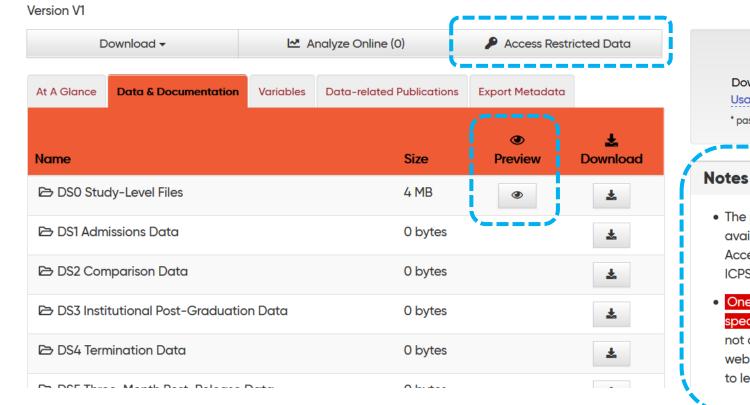
Outcome Evaluation of the Wisconsin Residential Substance Abuse Treatment (RSAT) Program: The Mental Illness Chemical Abuse (MICA) Program at Oshkosh Correctional Institution, 1997-2000 (ICPSR 3082)

Version Date: Mar 30, 2006 Cite this study | Share this page

Principal Investigator(s): (2)

D. Paul Moberg, University of Wisconsin. Medical School. Department of Population Health Sciences. Center for Health Policy and Program Evaluation; University of Wisconsin, Medical School, Department of Population Health Sciences, Center for Health Policy and Program Evaluation

https://doi.org/10.3886/ICPSR03082.v1



Downloads *

Usage Report

* past three years

- The public-use data files in this collection are available for access by the general public. Access does not require affiliation with an ICPSR member institution.
- One or more files in this data collection have special restrictions. Restricted data files are not available for direct download from the website; click on the Restricted Data button to learn more.

Data-related

Publications

Zenodo connexion GitHub



Software Open Access

Software Open Access

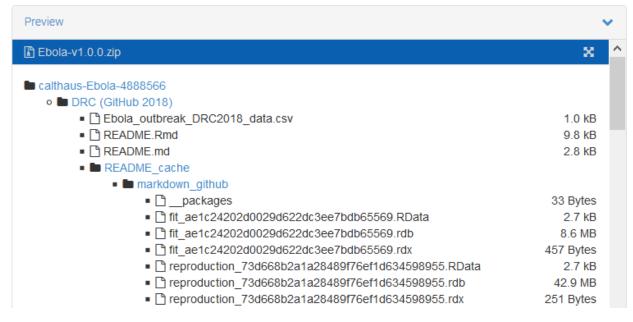
Analyses of Ebola virus disease outbreaks

(b) Christian Althaus

June 8, 2018

This release contains data and R code files from the following analyses of Ebola virus disease (EVD) outbreaks:

- Althaus CL. (2014) Estimating the reproduction number of Ebola virus (EVOB) during the 2014 outbreak in West Africa, PLoS Curr. 6.
- 2. Althaus CL, Low N, Musa EO, Shuaib F, Gsteiger S. (2015) Ebola virus disease outbreak in Nigeria: Transmission dynamics and rapid control. Epidemics, 11:80-4.
- Althaus CL. (2015) Rapid drop in the reproduction number during the Ebola outbreak in the Democratic Republic of Congo. PeerJ, 3:e1418.
- 4. Althaus CL. (2015) Ebola superspreading70135-0). Lancet Infect Dis, 15:507-8.
- Abbate JL, Murall CL, Richner H, Althaus CL. (2016) Potential impact of sexual transmission on Ebola virus epidemiology: Sierra Leone as a case study. PLOS Negl Trop Dis, 10:e0004676.
- 6. Althaus CL. (2018) Real-time analysis of the 2018 Ebola outbreak in the Democratic Republic of Congo). GitHub.



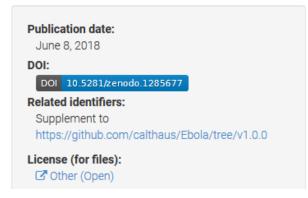
69 7

w views

See more details...

DLog in





Exemple:

https://doi.org/10.5281/zenodo.128567



Trouver un entrepôt de données généraliste

Outil d'aide à la décision en ligne UBx : « Trouver un entrepôt de données » Limites des entrepôts de données généralistes

Trouver un entrepôt de données généraliste

Outil d'aide à la décision en ligne UBx : Trouver un entrepôt de données

- →Pour identifier l'entrepôt de données généraliste qui réponde le mieux aux besoins, à partir de la réponse aux 4 questions suivantes :
 - > Allez-vous publier des données dont l'accès doit être restreint?
 - > Recherchez-vous un entrepôt avec des options de dépôt gratuit?
 - > Souhaitez-vous pouvoir définir librement les conditions d'utilisation de vos données?
 - > Souhaitez-vous que votre dépôt de données soit relu avant d'être mis en ligne?
- →Une brève fiche descriptive accompagne chacun des huit entrepôts de données comparés : <u>4TU.ResearchData</u>, <u>Figshare</u>, <u>Harvard Dataverse</u>, <u>Dryad</u>, <u>Mendeley Data</u>, <u>Open Science Framework (OSF)</u>, <u>B2SHARE</u>, <u>Zenodo</u>

Trouver un entrepôt de données généraliste

Trouver un entrepôt de données

Vous cherchez un entrepôt de données généraliste pour rendre accessibles vos données de recherche ?

Cet outil peut vous aider à trouver l'entrepôt qui réponde le mieux à vos besoins. Pour identifier un entrepôt de données spécialisé pour votre domaine de recherche, interrogez l'annuaire re3data ou contactez-nous.

Pour plus d'informations concernant les données de recherche, consultez les pages de notre site web Les données de recherche.

Pour toute question, contactez-nous à l'adresse doc-soutienrecherche@u-bordeaux.fr

r cours de parametrage et de priditsation, les contenus et les jonctionnalites sont susceptibles d'évoluér.	
2. Recherchez-vous un entrepôt avec des options de dépôt	
gratuit ?	
Oui Pas nécessairement	
4. Souhaitez-vous que votre dépôt de données soit relu	
avant d'être mis en ligne	
Oui Pas nécessairement	

Réinitialiser

Limites des entrepôts de données généralistes

Les entrepôts de données généralistes ne permettent généralement pas de bénéficier des avantages suivants des entrepôts spécialisés

- →Prise en compte des spécificités d'un ou plusieurs domaines disciplinaires
- → Richesse et précision des informations descriptives
- → Interopérabilité : renforcée si l'entrepôt de données requiert la conformité à un format de données
- → Réutilisation des données par des applications informatiques (machine readability) cf. point ci-dessus



Trouver un entrepôt de données spécialisé

Recommandations des éditeurs et des organismes de financement

Annuaire re3data

Service FAIRsharing

Recommandations

- → Editeurs : exemples <u>PLoS</u> et <u>Springer-Nature</u>
- → Organismes de financement
 - > <u>ERC European Research Council</u> : liste commentée d'entrepôts spécialisés par discipline p. 9 et suiv.
 - > <u>Commission européenne</u> : liste d'entrepôts généralistes et spécialisés par discipline ces *Data Guidelines* comportent en outre des recommandations pour la préparation des données (notamment tabulaires).
 - NIH: Open Domain-Specific Data Sharing Repositories : liste commentée d'entrepôts spécialisés, incluant les liens vers leurs politiques de soumission et d'accès
 - NIH: voir aussi NOT-OD-21-016: Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research

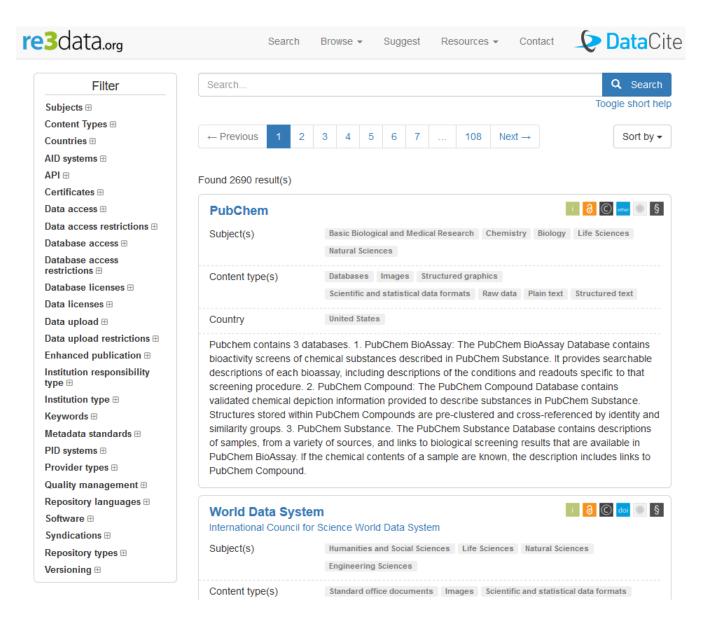
Annuaire re3data

Registry of Research Data Repositories.

https://www.re3data.org/

Il s'agit de la ressource de référence, indexant plus de 2000 entrepôts. On peut rechercher par mot-clé et naviguer dans re3data par :

- \rightarrow sujet,
- →type de contenu,
- \rightarrow pays.



Annuaire re3data

De nombreux filtres permettent ensuite d'affiner les listes de résultats, en fonction par exemple des critères suivants.

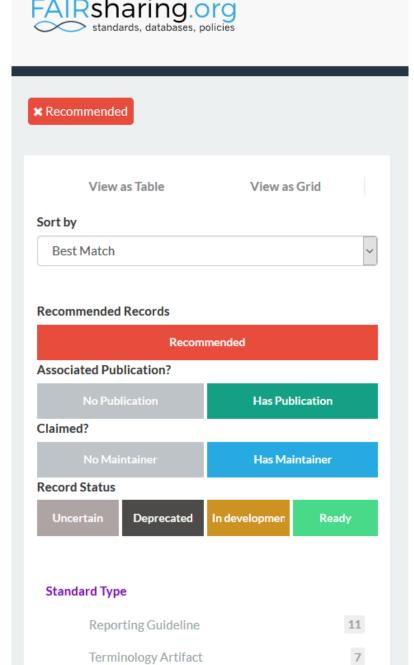
- → Certificates : par quelle certification l'entrepôt est-il qualifié?
- → Accès : gradient de valeurs de fermé à ouvert se décompose en 3 types d'accès :
 - > **Database access** : accès à l'entrepôt de données lui-même : sous quelles conditions un utilisateur peut-il accéder à la base de données en général?
 - Data access : accès aux jeux de données déposés dans un entrepôt de données spécifique : sous quelles conditions un utilisateur peut-il accéder à un jeu de données?
 - > **Data upload** : accès à la soumission de données : sous quelles conditions un utilisateur peut-il soumettre des données ?
- → **Versioning** : les jeux de données peuvent-ils être versionnés?

Service FAIRsharing

<u>FAIRsharing.org</u> répertorie non seulement des entrepôts mais également des standards, des méthodes, des vocabulaires, etc. Pour chaque ressource sont notamment spécifiés les critères suivants.

- Le statut : en développement / opérationnel / incertain / déprécié
- La recommandation : nom de l'éditeur, de la revue, etc. qui recommande

Voir: Sansone, S.-A. et al. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4), 358-367. https://doi.org/10.1038/s41587-019-0080-8





Merci pour votre attention

Des questions?

Contact: