Trouver des jeux de données via des bases pluridisciplinaires et des moteurs de recherche en 8 points

1. L'intérêt des jeux de données

Les données (*data*) collectées, observées, créées ou compilées dans le cadre de projets de recherche sont précieuses :

- elles valident ou invalident des hypothèses de recherche ;
- elles étayent les résultats de recherche présentés dans une publication ;
- elles sont le point de départ de nouvelles recherches ;
- elles constituent un apport complémentaire à d'autres travaux.

Trouver des jeux de données susceptibles d'enrichir les connaissances dans un domaine est essentiel à l'avancée des sciences. Réutiliser des jeux de données est un facteur d'efficience en recherche.

La publication de jeux de données (*datasets*) se développe mais reste inégale. Certaines disciplines comme l'astronomie en ont une longue expérience. Dans d'autres disciplines, comme l'agronomie ou les sciences sociales, la collecte des données et leur valorisation restent confinées au poste de travail du scientifique, du doctorant ou du laboratoire qui a conduit les recherches.

Mais les pratiques évoluent avec la science ouverte, la promotion des principes FAIR pour des données *« faciles à trouver, accessibles, interopérables et réutilisables »* et l'exigence de qualité et de transparence en recherche. Les scientifiques sont amenés à penser, planifier et assurer la gestion et la conservation des données tout au long de leurs travaux de recherche et au-delà. Cela couvre toutes les étapes : production, publication, diffusion et partage de données.

Dès lors que des données issues de recherches sont publiées pour être connues et réutilisées, leur découverte doit être facilitée :

- des entrepôts de données (*Research Data Repository* ou *Data Repository*) appelés parfois archives ouvertes (*Open archives,* même si ce terme est plutôt réservé aux entrepôts de publications) sont dédiés au dépôt, à la conservation et à la diffusion de certains types de données et de leurs métadonnées (voir fiche CoopIST Déposer des données dans un entrepôt) ;
- de nouveaux outils sur internet (bases de données pluridisciplinaires et moteurs de recherche académiques) permettent de rechercher des jeux de données sans savoir a priori dans quel entrepôt ceux-ci ont été déposés et sont accessibles.

2. Les bases et moteurs de recherche pour trouver des jeux de données

Quels bases de données ou moteurs de recherche seront pertinents pour trouver des jeux de données sans avoir de connaissance précise sur les entrepôts susceptibles de les héberger ? Cela nécessite de connaître quelques caractéristiques de ces outils pour évaluer leur intérêt.

- Le **périmètre couvert** : combien et quels types d'entrepôts de données sont moissonnés pour être indexés par la base de données ou le moteur de recherche ? Combien de jeux de données sont référencés ? Des informations précisant l'origine, les modalités d'accès et les fonctionnalités de l'outil de recherche sont un gage de transparence et de fiabilité.
- Le **mode de recherche** : un formulaire de **recherche avancée** sur les différents champs d'un jeu de données est-il proposé ? Permet-il d'interroger par nom d'auteurs, organisme d'affiliation, année de publication, mots-clés ou mots du titre des jeux de données ? Des requêtes complexes sont-elles possibles, avec des opérateurs booléens, par exemple entre champs et entre les mots d'un même champ (voir fiche CoopIST Du sujet à l'équation de recherche) ? Des **filtres** sont-ils proposés pour limiter les résultats à un éditeur de jeux de données (*Publisher*), à l'entrepôt originel qui les contient (*Provider),* à la plate-forme d'où ils ont été collectés *(Source*), à un domaine scientifique, à un bailleur (*Funder*) qui a financé le projet de recherche ayant donné lieu aux jeux de données ?
- Les fonctions **d'affichage et d'export** des références des jeux de données : les champs (c'est-à-dire les métadonnées telles que titre, auteurs, éditeur, année, résumé, etc.) décrivant les jeux de données sont-ils bien identifiés et permettent-ils que les références soient prises en charge par des logiciels bibliographiques comme EndNote ou Zotero (voir fiche CoopIST Citer un jeu de données scientifiques) ? L'éditeur de jeux de données (*Publisher*), l'hébergeur des données éditées (*Provider)* ou le site les collectant (*Source*)*,* et le lien d'accès aux métadonnées du jeu des données et au(x) fichier(s) des données sont-ils identifiables ? Un lien internet est-il fourni vers la publication associée au jeu de données quand elle existe (URL, DOI – voir fiche CoopIST Identifier et rechercher une publication ou un jeu de données par son DOI) ?
- D'autres possibilités comme la **création d'un compte personnel** autorisant une connexion par authentification pour une gestion personnalisée (sauvegarde des résultats d'une recherche, export) sont appréciés par les utilisateurs plus aguerris.

Une sélection de bases de données et de moteurs académiques est proposée ci-après pour aider l'utilisateur à faire ses premiers pas dans la recherche bibliographique de jeux de données. Ces outils manquent encore de fiabilité car ils n'ont pas la maturité des bases de données bibliographiques de publications en termes de représentativité, de qualité et de pertinence.

3. Data Citation Index, la base de données payante de Clarivate Analytics

Périmètre - Lancée fin 2012, Data Citation Index (DCI) est une base de données payante de l'entreprise américaine Clarivate Analytics, qui référence une sélection d'entrepôts de données (*Repository*), de jeux de données (*Data set*) et de données issues d'études (*Data study*) accessibles en ligne. DCI analyse aussi les références de publications citant les ensembles de données référencés.

DCI est composée d'une section Sciences de la terre et de la vie (DCI-S) et d'une section Social Sciences & Humanities (DCI-SSH), couvrant toutes deux la période de 1900 à aujourd'hui.

DCI affichait le 25 août 2020 dans sa Master Data Repository List 427 entrepôts indexés, et annonçait en juin 2020 contenir 9 567 435 *datasets* et 1 231 647 *data studies*.

Parmi les entrepôts indexés : entrepôts pluridisciplinaires de données (Dryad, Figshare…), entrepôts thématiques de données (GenBank en génétique, Pangaea en sciences de la terre et environnementales…), entrepôts publics internationaux de publications et de données (Zotero…),

entrepôts institutionnels (Harvard Dataverse, CIFOR et ICRISAT Dataverses…), entrepôts de données d'éditeur commercial (Mendeley Data d'Elsevier).

Mode de recherche - Data Citation Index est accessible à partir de la plate-forme Web of Science (http://webofknowledge.com/) :

- dans la barre de menus sélectionner dans *All Databases* la base *Data Citation Index* ;
- dans la page qui s'affiche, cliquer sur *More Settings* pour cocher les bases souhaitées : Science (DCI-S) et/ou Social Sciences & Humanities (DCI-SSH).

La recherche peut être simple (*Basic Search*) ou avancée (*Advanced Search*). Elle peut se faire par type de document, auteur, affiliation, titre, année de publication, langue, sujet : *Topic, Web of Science Category, Subject Area,* descripteurs *(Subject Descriptors),* source de financement *(Funding Text),* DOI. Chaque référence fournit un résumé et le lien internet *(Source URL)* vers le jeu ou l'entrepôt de données référencé.

Pour en savoir plus sur DCI : https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/.

4. Dataset Search, le moteur gratuit de Google pour rechercher de données

Périmètre - Lancé par Google LLC (USA) en 2018 dans une version beta puis en 2020 dans sa version finale toujours gratuite, Google Dataset Search (https://datasetsearch.research.google.com/) est présenté comme un complément de Google Scholar (voir fiche CoopIST Utiliser des moteurs de recherche académiques gratuits).

Dataset Search indexe les pages web dont les données (gouvernementales ou académiques) sont décrites selon le standard schema.org, avec les différents champs ou rubriques introduits par des balises textes (https://schema.org/Dataset). Le 25 août 2020, Dataset Search déclarait sur son blog indexer 31 millions de *datasets*. Parmi les sites et entrepôts de données indexés issus de 4 600 sites internet : entrepôts institutionnels (Dataverse du Cirad), entrepôts thématiques (GBIF sur la biodiversité), sites de données administratives (data.gouv.fr), bases de données publiques internationales (OpenAIRE explore), réseaux sociaux (ResearchGate).

Mode de recherche - Le formulaire de recherche est simple, par mots. Des préfixes peuvent être utilisés pour rechercher un mot dans le titre (*intitle:*), dans le nom de site web (*site:*) ou dans une adresse de page web (*inurl:*).

Les résultats affichés peuvent être filtrés par période de mise à jour (depuis 1 mois, depuis 1 an…), par format de téléchargement du jeu de données (tableau, document, image, texte, archive…), par droit d'usage des données (commercial, non commercial), par thème, par accès gratuit. Le nom de la source avec un lien vers le jeu de données est fourni. Les descriptifs des jeux de données peuvent être enregistrés (via l'icône *Ajouter aux favoris*) dans son espace personnel après connexion et partagées sur les réseaux sociaux. L'icône *A propos* permet d'afficher quelques informations sur le moteur et l'icône *Commentaires* permet de poster une question ou un commentaire.

L'aide en ligne accessible après connexion à son compte Google s'adresse surtout aux développeurs de site web souhaitant que les jeux de données publiés sur leur site soient correctement balisés afin d'être indexés par Dataset Search.

5. Dimensions, le moteur de recherche académique gratuit de Digital Science

Périmètre - Lancé en 2018 par la société commerciale Digital Science (UK), le moteur de recherche académique Dimensions (https://app.dimensions.ai/discover/data_set) permet de mener une

recherche dans 1 556 3184 *datasets* (25 août 2020). Ils sont issus d'une centaine d'entrepôts, parmi lesquels des entrepôts pluridisciplinaires (Dryad, Figshare contenant lui-même des données associées aux publications d'éditeurs scientifiques comme PLoS), des entrepôts publics internationaux (Zenodo), des entrepôts thématiques (Pangea), l'entrepôt de données Mendeley data de l'éditeur commercial Elsevier. (Voir aussi la fiche CoopIST Utiliser des moteurs de recherche académiques gratuits).

Mode de recherche - Comme pour les publications, la recherche peut porter sur l'ensemble des données du jeu de données (*Full data*), ou être limitée aux titre et résumé (*Title and abstract*), ou aux seuls mots-clés (*Keywords Search*) ou au seul résumé (*Abstract Search*) du jeu de données.

Les résultats peuvent être filtrés par année (*Publication year*), par auteurs (*Researchers*), par thématique (*Field of research*), par titre de la revue publiant les jeux de données (*Source*), et par entrepôt (*Repository*).

Après création d'un compte et connexion, le bouton *Save/Export* est actif et permet une sauvegarde des résultats (*Save as favorite*) dans Dimensions ou un export (*Export results*) des références de jeux de données obtenus.

L'aide en ligne (https://support.dimensions.ai/support/home) sous la forme d'une foire aux questions (FAQ) est bien documentée avec un moteur de recherche spécifique.

6. OpenAIRE explore, la plate-forme européenne d'accès aux publications et datasets

Périmètre - Lancé en 2008, OpenAIRE (https://explore.openaire.eu/search/find/datasets) est à la fois un projet européen de soutien à la science ouverte (*Open Science*) et une infrastructure technique « OpenAIRE explore » indexant les références des publications et des données de recherche d'entrepôts ou de sites web de fournisseurs d'information.

Au 25 août 2020, OpenAIRE explore contenait 11 864 933 références de données de recherche (*Research data*) issues de plus d'une centaine de fournisseurs d'information (*Content Providers*, exemple : agence Datacite d'enregistrement de DOI de jeux de données déposés dans des entrepôts de toutes sortes) ou d'entrepôts — pluridisciplinaires (Figshare…), thématiques (Omics Discovery Index, Pangaea…), publics internationaux (Zenodo…), institutionnels (Dataverse du Cirad…), d'éditeur commercial (Mendeley Data d'Elsevier).

Mode de recherche - Une recherche simple par mots propose des filtres par agence de financement, projet, date, mode d'accès, type de données, (*dataset,* image, essai clinique, son…), fournisseur de contenus…

La recherche avancée se fait sur différents champs d'une référence de jeux de données : titre, auteur, éditeur (*Publisher*), projet (*Project*), date de publication, fournisseur de contenus (*Collected from Content Provider*), source de contenus (*Hosting Content Provider*), organisme (*Organization*), agence de financement (*Funder*), thématique (*Subject*), langue…

Les résultats peuvent être triés par pertinence ou par ordre croissant ou décroissant de date.

Une aide en ligne très générique de type FAQ est accessible à : https://www.openaire.eu/faqs.

7. BASE, Bielefeld Academic Search Engine, de l'Université allemande

Périmètre - Créé en 2004 par l'Université allemande Bielefeld, BASE (https://www.base-search.net/Search/Advanced) se définit comme l'un des moteurs de recherche académique mondiaux les plus volumineux, moissonnant les entrepôts de publications et de données de recherche. Au 25 août 2020, BASE indexait 8 302 858 *datasets* issus de 8 334 sources d'information dont près de 200

entrepôts de données internationaux, nationaux, institutionnels, publics ou privés, parmi lesquels DataCite Metadata Store, Zenodo, Pangaea, Figshare…

Mode de recherche - Le formulaire de recherche avancée (*Advanced search*) permet de saisir des critères de recherche sur tout ou partie d'un document (titre, auteur, thématiques, éditeur…) et de filtrer les résultats par type de documents comme les jeux de données (*Dataset)* et par licence d'utilisation (*Terms of Re-use/Licences*).

Les résultats peuvent être affichés de façon sommaire ou en détail avec les différents champs indexés comme auteurs (*Author*), résumé (*Abstract*), éditeur (*Publisher*), année de publication (*Year of Publication*), langue, type de document (*Document Type*), termes de réutilisation (*Terms of Reuse*), fournisseur de contenu (*Content Provider*), etc.

Pour sauvegarder une recherche, l'utilisateur doit avoir au préalable créé son compte dans BASE et s'y être connecté. Sa recherche et tout ou partie de ses résultats peuvent alors être enregistrés, partagés par messagerie ou exportés en différents formats bibliographiques.

BASE fournit des informations détaillées sur les sources indexées (*Content Sources*), ainsi qu'une aide en ligne (Help) et une foire aux questions (FAQ) précieuses pour la recherche.

8. Data Mendeley, la base gratuite d'Elsevier pour les données de recherche

Périmètre - Lancé en 2015 par l'éditeur néerlandais Elsevier, Data Mendeley (https://data.mendeley.com/) se définit comme un entrepôt ouvert de données de recherche où les chercheurs et les institutions peuvent charger et diffuser leurs données, associées ou non à des publications. Au 25 août 2020, il affichait 24,9 millions de *datasets* issus de plus de 1 700 entrepôts spécialisés et pluridisciplinaires, privés et publics, comme ScienceDirect d'Elsevier, Zenodo, Dryad, Figshare, Harvard Dataverse…

Mode de recherche - La recherche peut être simple par mots-clés, ou avancée par saisie d'une requête selon une formulation propre au moteur de recherche et portant sur un ou plusieurs champs (*Author, Title, Institution,* DOI, *Keywords, Subject Area*…) pouvant être combinés avec des opérateurs booléens (AND, OR, NOT).

Les résultats qui s'affichent peuvent être filtrés par types de données (*tabular data*, *dataset*…), types d'entrepôts (*article repositories*, *data repositories*), et sources de données.

Une aide en ligne (Advanced search help) sous le formulaire avancé et une foire aux questions (FAQ) sont accessibles en ligne.

Marie-Claude Deboin
Délégation à l'information scientifique et technique, Cirad
28 août 2020

Informations

Comment citer ce document :
Deboin, M.C.. 2020. Trouver des jeux de données via des bases pluridisciplinaires et des moteurs de recherche. Montpellier (FRA) : CIRAD, 5 p. https://doi.org/10.18167/coopist/0071