Citer un jeu de données scientifiques en 5 points

1. Comprendre la nécessité de publier et de citer un jeu de données scientifiques

Les **résultats de recherche décrits dans une publication** (article, ouvrage, etc.) **sont toujours étayés par un ou plusieurs jeux de données scientifiques** (*data set*).

Du fait de sa complexité ou de son volume, **le jeu de données accompagne rarement la publication**, que ce soit un article de recherche (*original paper*), un article de synthèse confrontant des jeux de données d'origines diverses (*review paper, meta-analysis*), ou un article décrivant le jeu de données (*data paper* ; voir fiche CoopIST Rédiger et publier un data paper dans une revue scientifique).

Le jeu de données scientifiques est alors **déposé** sur internet **dans un entrepôt de données** (*data repository*) institutionnel, thématique, national ou international (voir fiche CoopIST Rendre public ses jeux de données scientifiques).

Le jeu de données publié ou déposé doit être **cité dans les publications** auxquelles il est lié. Comme pour toute publication, **la référence du jeu de données et sa citation dans le texte obéissent à des règles**.

2. Comment libeller la référence bibliographique d'un jeu de données ?

Citer un jeu de données consiste à construire sa référence bibliographique et à utiliser cette référence dans une publication. Cette référence qualifie de manière univoque le jeu de données :

- elle en identifie le (ou les) auteur(s) ;
- elle permet de rechercher et de localiser le jeu de données sur internet ;
- elle facilite l'exploitation et la réutilisation des données par d'autres équipes de recherche que celles des auteurs.

La **référence bibliographique complète** d'un jeu de données comporte les éléments suivants :

- **Auteur** (*Author*) : créateur (*Creator*) du jeu de données
- **Date de publication** (*Publication Year*) : selon les situations, date de mise en ligne du jeu de données ou date de fin d'embargo à l'issu duquel le jeu de données devient accessible
- **Titre** (*Title*) : titre du jeu de données, et éventuellement titre de la collection ou de la sous-collection dont le jeu de données fait partie
- **Edition :** niveau ou stade de traitement du jeu de données, selon une nomenclature si possible appropriée au type de données concernées

- **Version :** numéro croissant au fur et à mesure des modifications apportées aux données ou au processus de traitement
- **Nom de la norme,** du standard, ou du modèle de référence des données (*Feature Name*) et son identifiant sur internet ou **URI** (*Uniform Resource Identifier*) : par exemple ISO 19101-1:2014 (https://www.iso.org/obp/ui/#iso:std:iso:19101:-1:ed-1:v1:en) si cette norme est utilisée pour référencer l'information géographique relative au jeu de données
- **Type de ressource** (*Resource Type*) : base de données (*database*, voir fiche CoopIST Rendre public ses jeux de données scientifiques), jeu de données (*data set*), logiciel (*software*), image, vidéo, etc.
- **Editeur** (*Publisher*) : organisation produisant (*Producer*) ou rendant accessible (*Distributor*) le jeu de données
- **Identifiant** (*Identifier*) : code identifiant le jeu de données de façon pérenne et univoque, par exemple un DOI (*Digital Object Identifier*, identifiant numérique d'objet)
- **Localisation** (*Location*) : Adresse URL où le jeu de données est accessible

Le **format minimal d'une référence bibliographique** d'un jeu de données comporte 5 éléments (recommandation de DataCite, consortium international dont l'objectif est de faciliter l'accès aux données de la recherche et leur réutilisation) :

- **Auteur (Année de publication) : Titre. Editeur. Identifiant**
  Creator (PublicationYear): Title. Publisher. Identifier

Ce format minimal peut être complété, si besoin, par la version et le type de ressource :

- **Auteur (Année de publication) : Titre. Version. Editeur. Type de ressource. Identifiant**
  Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

La **granularité** d'un jeu de données complique son référencement et sa citation, un grain pouvant correspondre à un ou plusieurs fichiers, un fichier contenant un ou plusieurs tableaux, et un tableau contenant plusieurs données. Vous pouvez référencer le jeu de données au niveau de granularité auquel a été attribué l'identifiant par l'entrepôt. Ensuite, si vous avez besoin de citer un grain plus fin, vous indiquerez dans le texte de votre publication les informations permettant au lecteur de retrouver le sous-ensemble concerné.

La **dynamique** d'un flux de données et la **fugacité** d'une donnée compliquent également le référencement (données météorologiques par exemple). L'auteur du jeu de données définira des versions successives de ce jeu, mémorisera et affichera la date et l'heure auxquelles les données observées correspondent afin que l'utilisateur puisse y faire référence dans sa publication.

3. Exemples de formats de références bibliographiques de jeux de données

Les instructions aux auteurs (*Guide for Authors*) de revues scientifiques proposent un format de citation de jeux de données. Si ce n'est pas encore le cas, vous pouvez suivre les exemples ci-après.

Exemples issus de DataCite (https://www.datacite.org/services/cite-your-data.html) :

Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo. http://dx.doi.org/10.1594/PANGAEA.726855

Geofon operator (2009): GEFON event gfz2009kciu (NW Balkan Region). GeoForschungsZentrum Potsdam (GFZ). http://dx.doi.org/10.1594/GFZ.GEOFON.gfz2009kciu

Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. http://dx.doi.org/10.1594/WDCC/dphase_mpeps

Référence créée à partir du DOI du jeu de données :

A partir de la saisie du DOI d'un jeu de données (*Digital Object Identifier*), l'application en ligne DOI Citation Formatter beta (développée par DataCite et CrossRef) affiche la référence bibliographique du jeu de données dans un format à choisir parmi **500 formats de revues scientifiques**.

- Exemple pour la revue *Plant Biology* :
  Heneghan, C, Thompson, M, Billingsley, M, Cohen, D (2011) Data from: Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts. [online] URL: http://dx.doi.org/10.5061/dryad.585t4
- Exemple pour la revue *PLoS* :
  1. Heneghan, C, Thompson, M, Billingsley, M, Cohen, D. Data from: Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts [Internet]. Dryad Digital Repository; 2011. http://dx.doi.org/10.5061/dryad.585t4

Formats de références proposés par des entrepôts de données :

- Entrepôt biologie et écologie Dryad :
  Heneghan C, Thompson M, Billingsley M, Cohen, D (2011) Data from: Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.585t4
- Entrepôt sciences de la terre et environnementales PANGAEA :
  Mercier, Herlé (2005): Shipboard acoustic doppler current profiling during cruise 35A3CITHER3_1 (SAC ID 00272). http://dx.doi.org/10.1594/PANGAEA.319620
- Entrepôts de données gérés par le logiciel Dataverse :
  GIDA-IFAD, 2015, "LACOSREP Upper-East Ghana Reservoir Inventory", http://dx.doi.org/10.7910/DVN/MEAPCE, Harvard Dataverse, V1
- Entrepôt multidisciplinaire Zenodo :
  Federman, Sarah et al. (2015). Supporting data: The biogeographic origin of a radiation of trees in Madagascar: Implications for the assembly of a tropical forest biome. Zenodo. http://dx.doi.org/10.5281/zenodo.31503

4. Quels logiciels gèrent les références bibliographiques de jeux de données ?

Le logiciel bibliographique commercial EndNote (version X4) propose le type de référence *Data Set* avec les champs spécifiques suivants : *Investigators* (équivalent du champ *Author* des publications), *Producer, Distributor, Study Number, Original Release Date, Series Title, Version, Date of Collection, Version History, Geographic Coverage, Time Period, Unit of Observation, Data Type, Dataset(s).* Les références de jeux de données peuvent être exportées et insérées dans un article selon le format de la revue.

Le logiciel libre Zotero (version 4) n'a pas de type de référence pour décrire les jeux de données. Les références de jeux de données importées dans une bibliothèque Zotero apparaissent sous le type *Document*.

Le logiciel gratuit Mendeley (version en ligne 1.15.2) n'a pas de type de document spécifique pour gérer les références de jeux de données.

5. Rechercher des jeux de données

Début 2016, **Google Scholar** et **Microsoft Academic Search**, moteurs de recherche spécialistes de la littérature scientifique, n'avaient pas encore de recherche spécifique sur les jeux de données.

DataCite Metadata Search beta, moteur gratuit de recherche de DataCite, permet de rechercher un jeu de données à partir de ses métadonnées : mots-clés, date de publication, DOI, etc.

Find a Repository, autre application gratuite de DataCite, permet de rechercher des entrepôts par mots-clés, type de ressources, pays, et d'y accéder via le répertoire mondial **re3data.org** (Registry of Research Data Repositories).

Data Citation Index (DCI), base de données payante de la société américaine Thomson Reuters, indexe plus de 3 millions d'enregistrements issus de 300 entrepôts de données scientifiques accessibles en ligne. Les données indexées sont réparties en 3 types : entrepôts de données (*Repositories*), jeux de données (*Data Sets*), données issues d'études (*Data Studies*).

Dans Data Citation Index, une recherche peut se faire par type de document, auteur, affiliation (adresse), titre, année de publication, langue, sujet, source de financement, DOI. Chaque résultat affiché est associé à un résumé, au lien internet (*Source URL*) vers le jeu ou l'entrepôt de données référencé, et à sa référence bibliographique (*How to cite this Resource*).

- Exemple de référence mise en forme par Data Citation Index :
  Maynaud, Geraldine; Brunel, Brigitte; Mornico, Damien; Durot, Maxime; Severac, Dany; Dubois, Emeric; Navarro, Elisabeth; Cleyet-Marel, Jean C; Le Quere, Antoine (2013): GSM1112034: mRNAseq_STM2683_Cd. Gene Expression Omnibus.
  http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1112034

Chaque résultat est accompagné du **nombre de citations reçues** à partir d'autres jeux de données et des publications indexées dans les bases de données de Thomson Reuters (Data Citation Index, Web of Science Core Collection, BIOSIS Citation Index, SciELO Citation Index).

Ouvrages et sites utiles

Ball A., Duke M. 2015. How to cite datasets and link to publications. Edinburgh (UK): Digital Curation Centre (DDC), 15 p. http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Corti L., Van den Eynden V., Bishop L., Woollard M. 2014. Managing and sharing research data: a guide to good practice. Los Angeles: SAGE Publications Ltd, 222 p.

DataCite. Cite your Data. https://www.datacite.org/services/cite-your-data.html

DataCite. Format your Citation. https://www.datacite.org/services/format-your-citation.html

DataOne [Data Observation Network for Earth]. Data Citation and Attribution.
https://www.dataone.org/citing-dataone

Lawrence B., Jones C., Matthews B., Pepler S., Callaghan S. 2011. Citation and peer review of data: Moving towards formal data publication. International Journal of Digital Curation, 6(2), 4-37. http://dx.doi.org/10.2218/ijdc.v6i2.205

Martone M. (ed.). 2014. Joint Declaration of Data Citation Principles. San Diego CA Data Citation Synthesis Group: FORCE11 (The Future of Research Communication and e-Scholarship). https://www.force11.org/datacitation

Ray J. M. 2014. Research data management: practical strategies for information professionals. West Lafayette: Purdue University Press, 436 p.

Marie-Claude Deboin
Délégation à l'information scientifique et technique, Cirad
12 janvier 2016

Informations

Comment citer ce document :
Deboin, M.C.. 2016. Citer un jeu de données scientifiques en 5 points. Montpellier (FRA) : CIRAD, 5 p. http://url.cirad.fr/ist/citer-jeu-donnees