Rédiger et publier un *data paper* dans une revue scientifique en 5 points

1. Qu'est-ce qu'un *data paper* ?

Le *data paper* est une publication qui décrit un jeu de données scientifiques brutes (*data, dataset*), notamment à l'aide d'informations précises, appelées métadonnées (*metadata*). Les données décrites doivent être accessibles, soit sous forme de fichiers annexés, soit plus généralement par un lien pérenne (URL, DOI) vers « l'entrepôt de données » en ligne (*data repository,* ou *repository of research data*) où elles sont déposées et correctement formatées. Les métadonnées détaillent pourquoi, par qui et comment ces données ont été collectées, qui en est propriétaire, sous quel format elles sont stockées, etc.

Le *data paper* est publié sous la forme d'un article examiné par les pairs dans une revue scientifique classique publiant différentes formes d'articles dont des *data papers* ou dans un *data journal*, c'est-à-dire une revue contenant exclusivement des *data papers.*

Le *data paper* informe la communauté scientifique de la disponibilité de ces jeux de données et de leur potentiel pour des utilisations futures. Contrairement à un article de recherche classique, le *data paper* décrit uniquement des données scientifiques et les circonstances et méthodes de leur collecte. Il ne rend pas compte des hypothèses ni des conclusions issues de l'analyse de ces données. Néanmoins, il présente les analyses techniques et statistiques validant la qualité des données.

Le *data paper* montre l'originalité et la portée du jeu de données qu'il décrit. Les revues qui publient des *data papers* s'intéressent particulièrement à la portée des données soumises, c'est-à-dire à leur potentiel de réutilisation par d'autres scientifiques. Il s'agit là de l'argument majeur pour convaincre le rédacteur en chef d'accepter votre *data paper.*

2. Pourquoi publier un *data paper* ?

- Le *datapaper* a pour objectif d'informer la communauté scientifique de l'existence et de la disponibilité d'un jeu de données qui est déposé dans un entrepôt de données et auquel cet entrepôt a attribué un identifiant pérenne (*Digital Objet identifier* (DOI).
- Il valorise les données en exposant leur potentiel pour des utilisations et projets futurs.

- Il facilite la réutilisation des données en mettant en évidence la qualité des données et des procédures, ainsi que la rigueur scientifique de l'étude.
- Il apporte de la visibilité aux données, les rend plus facilement repérables et citables par d'autres études.
- Le *data paper* est une publication citable, au même titre que tout article scientifique publié, qui met en valeur son (ses) auteur(s) en tant que créateur(s) de données. Il permet la traçabilité des citations et des réutilisations.

3. Comment structurer un *data paper ?*

La structure du *data paper* est particulière. Elle varie néanmoins selon les revues scientifiques, entre une structure simple et une structure plus élaborée et détaillée. Lisez attentivement les instructions aux auteurs de la revue ciblée qui, parfois, proposent aussi des modèles de fichiers, de présentation ou d'organisation (*templates, tool kit…*).

La caractéristique du *data paper* est qu'il est lié au jeu de données brutes qu'il décrit. Dans ce but, le jeu de données est déposé dans un entrepôt de données, au préalable ou au moment de la soumission de l'article. L'identifiant pérenne du jeu de données est indiqué dans le *data paper*. L'accessibilité des données est vérifiée par les pairs lors de la révision du manuscrit (*peer-reviewing*).

La plupart des revues préconisent des entrepôts de données qu'elles jugent de confiance, en termes de pérennité notamment. Ces entrepôts sont généralement listés dans les instructions aux auteurs, en fonction du type de données (génétique, biodiversité, écologie, géoscience, sciences humaines et sociales, etc.). Certaines revues disposent aussi de leur propre entrepôt de données.

Le *data paper* comprend deux parties :

- l'ensemble des fichiers des données (*data files)* accessibles directement ou via un entrepôt de données. Ces données sont en libre accès ou en restriction d'accès temporaire ;
- la partie descriptive, c'est le *data paper* proprement dit. Cette partie explique le contexte d'obtention des données, les présente et en démontre la fiabilité.

La partie descriptive comprend en général les éléments suivants :

- page de titre avec les noms et affiliations des auteurs,
- résumé, parfois des mots-clés,
- introduction présentant l'arrière-plan de l'étude (contexte et enjeu généraux et spécifiques), les questions de recherche à l'origine de la collecte des données, et la plus-value de cette collecte (originalité, importance et potentiel d'utilisation en recherche),
- description suffisante des matériels et méthodes pour permettre de reproduire l'étude : protocole expérimental, méthode d'échantillonnage, descripteurs physiques, procédures de contrôle qualité…
- description suffisante des données pour permettre de les réutiliser : structure, format, disponibilité, explication de données aberrantes…
- information et discussion justifiant la fiabilité et la rigueur des données, si besoin accompagnées de figures et tableaux : validation de la procédure de collecte de données, analyses statistiques de l'erreur expérimentale, évaluation d'échantillons biologiques...

- si besoin, conseils pour la réutilisation des données,
- remerciements, contributions des auteurs, mention d'éventuels conflits d'intérêt,
- liste des références bibliographiques,
- figures, tableaux, annexes, relatifs à la méthodologie, à la qualité des données, ou proposant une synthèse des données.

4. Exemples de structure de *data papers* en sciences du vivant

Les *data papers* sont une forme nouvelle de publication qui se développe dans un objectif d'ouverture des données, c'est à dire pour qu'elles soient accessibles librement et gratuitement. Cela explique qu'en 2014, les revues scientifiques publiant des *data papers* sont encore peu nombreuses.

Nous proposons ci-dessous quatre exemples décrivant le corps de *data papers* publiés dans des revues dans les thèmes de l'écologie, la biodiversité, la génomique, et les sciences de la vie.

4.1. ***Ecology* et *Ecological Archives*** (Ecological Society of America, ESA)

Ecology (http://www.esajournals.org/loi/ecol), créée en 1920, est une revue sur abonnement mais qui autorise un libre accès aux articles après une période d'embargo de 2 ans. Elle publie différents types d'articles, dont des *data papers* depuis l'année 2000 (rubrique *Data papers*), qui sont en libre accès total.

Plus précisément, *Ecology* publie le résumé et donne le lien vers la version complète du *data paper* (contenant les données et métadonnées) qui est publiée dans *Ecological Archives* (http://esapubs.org/archive/default.htm). *Ecological Archives* a été créé par l'ESA pour publier les *data papers* et tout matériel complémentaire aux articles publiés dans ses revues.

Dans ce cas, le *data paper* donne un accès direct à chaque jeu de données, stocké dans *Ecological Archives* et identifié par un lien URL. Les instructions aux auteurs sont disponibles à : http://esapubs.org/archive/instruct_d.htm.

Le corps du *data paper* comprend les sections suivantes :

- *Introduction* : contexte, question de recherche et objectifs de l'étude,
- section *Metadata* qui rassemble les éléments suivants :
  - *Data set descriptors, Data structural descriptors* et *Data set status and accessibility :* description du jeu de données, du format des fichiers, du statut, de la qualité et de l'accessibilité aux données (libre, restrictions d'accès, coût),
  - *Research origin descriptors* : contexte de l'étude, méthodes et procédures.

En 2014, le coût de publication d'un *data paper* dans *Ecology/Ecological Archives* est de 250 $. Si la taille des fichiers dépasse 10 MB (formats *plain text, .txt, .csv)*, un coût supplémentaire est dû.

4.2. ***Biodiversity Data Journal*** (Pensoft Publishers)

Biodiversity Data Journal (http://biodiversitydatajournal.com/), créée en 2013, est l'une des 14 revues en libre accès total publiées par Pensoft. Toutes ces revues publient différentes formes d'articles dont des *data papers.*

Biodiversity Data Journal publie des *data papers* décrivant des données dans le domaine de la biodiversité, incluant écologie et environnement. Les petits jeux de données peuvent être accessibles directement sur le site de la revue sous forme de fichiers supplémentaires. Les jeux volumineux sont déposés dans un entrepôt de données (liste des entrepôts recommandés fournie dans les instructions aux auteurs).

Les instructions aux auteurs sont disponibles à :
http://biodiversitydatajournal.com/about#Datapublication.

Le corps du *data paper* comprend les sections suivantes :

- Introduction : contexte, question de recherche et objectifs de l'étude,
- Metadata, renseignées selon le *GBIF Metadata Profile elements* :
  - Taxonomic Coverage, Spatial Coverage, Temporal Coverage,
  - Project Description: *title of the project, personnel involved, funding sources, Study area description, and design description*,
  - Methods*: method step, Sampling, quality control*,
  - Dataset Descriptions: *Object name, character encoding, format name and version, distribution/online/URL, publication date, language, and intellectual rights*.

Pour faciliter la rédaction du *data paper* et la publication des données, l'éditeur propose aux auteurs de se connecter à la plateforme logicielle *Integrated Publishing Toolkit* (IPT, http://www.gbif.org/ipt) qui gère trois types de données : les données primaires, les checklists et les métadonnées. Les données sont publiées sous format *Darwin Core Archives* et *Ecological Modeling Language*.

A la date de rédaction de cette fiche (septembre 2014), cette revue ne demande pas de frais de publication aux auteurs.

4.3. ***Genomics Data*** (Elsevier)

Genomics Data (http://www.journals.elsevier.com/genomics-data/) est une revue en libre accès total créée en 2013. Elle publie différentes types d'articles dont des *data papers* (rubrique *Data in Brief*) décrivant des données génétiques avec un lien vers l'entrepôt où sont déposées les données. La liste des entrepôts recommandés est disponible à : http://www.elsevier.com/about/content-innovation/database-linking#supported-data-repositories.

Les instructions aux auteurs et le modèle (*template)* de *data paper* sont disponibles à :
http://www.elsevier.com/journals/genomics-data/2213-5960/guide-for-authors#2001.

Le corps du *data paper* comprend les sections suivantes :

- tableau de *Spécifications* des données,
- section *Experimental design, Materials and Methods,*
- *Discussion* courte mettant en valeur la portée du jeu de données.

En 2014, le coût de publication d'un *data paper* dans *Genomics Data* est de 500 $.

4.4. ***Scientific Data*** (Nature Publishing Group)

Scientific Data (http://www.nature.com/sdata/) est une revue en libre accès total créée en 2014. Elle publie uniquement des *data papers* (nommés *Data Descriptors* par la revue) décrivant des jeux de

données en sciences de la vie, environnement et biomédecine, avec un lien vers l'entrepôt où ces données sont librement accessibles. La liste des entrepôts recommandés est disponible à : http://www.nature.com/sdata/data-policies/repositories.

Les instructions aux auteurs et les modèles (*templates)* de *data paper* sont disponibles à : http://www.nature.com/sdata/for-authors/submission-guidelines.

Le corps du *data paper* comprend les sections suivantes:

- *Background & Summary* : contexte scientifique, question de recherche, objectifs de l'étude, valeur et potentiel des données pour leur réutilisation,
- *Methods :* méthodes, procédures et analyses. Le protocole expérimental peut être présenté sous forme de figure et les variables dans un tableau,
- *Data Records* : description des données et de l'entrepôt où elles sont déposées, format des fichiers. La présentation sous forme de tableaux est recommandée,
- *Metadata Records* : métadonnées présentées dans des tableaux (modèle fourni) : processus expérimental, méthode d'échantillonnage, description du site, conditions d'expériences, protocoles suivis, données obtenues…
- *Technical Validation* : explication de la rigueur scientifique de l'étude et de la qualité technique des données,
- *Usage Notes* : facultatives, indications facilitant la réutilisation des données par d'autres scientifiques.

En 2014, le coût de publication dans *Scientific Data* est de 750 €. Dans le cas où les données seraient déposées dans les entrepôts généralistes DRYAD (http://datadryad.org/) ou FigShare (http://figshare.com/), ce coût de publication inclut la possibilité de stocker 10 GB de données dans DRYAD ou 5 GB dans FigShare.

5. Liens utiles : exemples et guides

5.1. Exemples de revues publiant des *data papers*

BMC Research Notes (BioMed Central) : http://www.biomedcentral.com/bmcresnotes/authors/instructions/datanote

Earth System Science Data (Copernicus Publication) : http://earth-system-science-data.net/

Ecological Research (Springer) : http://www.springer.com/life+sciences/ecology/journal/11284

Ecology et *Ecological Archives* (Ecological Society of America) : http://www.esajournals.org/loi/ecol

F1000Research (F1000) : http://f1000research.com/

Genomics Data (Elsevier) : http://www.journals.elsevier.com/genomics-data/

Geoscience Data Journal (Wiley) : http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060

Geoscientific Model Development (Copernicus Publication) : http://www.geoscientific-model-development.net/

GigaScience (BioMed Central) : http://www.gigasciencejournal.com/about

International Journal of Robotics Research (SAGE Publications): http://ijr.sagepub.com/

Journal of Chemical & Engineering Data (ACS Publications) : http://pubs.acs.org/journal/jceaax

Journal of Open Archaeology Data (Ubiquity Press) : http://openarchaeologydata.metajnl.com/

Journal of Open Psychology Data (Ubiquity Press) : http://openpsychologydata.metajnl.com/

Journal of Open Health Data (Ubiquity Press) : http://openhealthdata.metajnl.com/

Journal of Open Research Software (Ubiquity Press) : http://openresearchsoftware.metajnl.com/

Journal of Physical and Chemical Reference Data (AIP Publishing) *:* http://scitation.aip.org/content/aip/journal/jpcrd

Scientific Data (Nature Publishing Group) : http://www.nature.com/sdata/

Les 14 revues de *Pensoft Publishers* : http://www.pensoft.net/about.php

5.2. Guides pour la description des jeux de données et des métadonnées

Le *data paper* doit décrire les métadonnées (*metadata*), de telle façon que les données puissent être réutilisables par tous. Or les auteurs sont souvent démunis face à la mise en forme de ces métadonnées, indispensables pour publier l'article et aussi pour déposer les données dans un entrepôt. Ces guides aident à la rédaction du *data paper* en expliquant les normes à respecter pour les métadonnées.

Dans le domaine de la biodiversité :

Penev L, Mietchen D, Chavan V, Hagedorn G, Remsen D, Smith V, Shotton D, 2011. Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Pensoft Publishers. http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

Dans le domaine des informations géospatiales :
http://www.ncddc.noaa.gov/metadata-standards/

Dans le domaine de l'écologie :
http://www.dcc.ac.uk/resources/metadata-standards/eml-ecological-metadata-language

Dans le domaine des sciences humaines et sociales :

Inter-university Consortium for Political and Social Research (ICPSR), 2012. Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.). Ann Arbor, MI. ISBN 978-0-89138-800-5. http://www.icpsr.umich.edu/files/deposit/dataprep.pdf

Information

Comment citer ce document :

Dedieu L. 2014. Rédiger et publier un data paper dans une revue scientifique en 5 points. Montpellier (FRA) : CIRAD, 7 p.

http://url.cirad.fr/ist/data-paper