Améliorer l'exposition des données de la recherche : la publication de data papers

Nathalie Reymonet

AMELIORER L'EXPOSITION DES DONNEES DE LA RECHERCHE : LA PUBLICATION DE *DATA PAPERS*

Nathalie REYMONET

Université Paris Diderot, Direction d'appui à la recherche

Janvier 2017

MOTS-CLE

données de la recherche ; communication scientifique ; open science ; open data ; data paper

RESUME

Les données de la recherche sont l'objet de l'intérêt des financeurs de la recherche publique, qui incitent les chercheurs à partager ces données, afin de répondre à des enjeux financiers comme de circulation des savoirs. Parmi les différentes modalités de la communication scientifique, la publication d'un « data paper » est une démarche relativement nouvelle. Le « data paper », ou article sur des données, décrit des données scientifiques et propose un lien vers un entrepôt de données qui les stocke. La description est en particulier très précise sur les points techniques et la méthodologie de production des données. Cette démarche va dans le sens de l'exposition des données, de leur accessibilité, leur interopérabilité et leur réutilisabilité, répondant ainsi aux recommandations des communautés d'intérêt de la recherche académique. Ce texte présente la structure et le contenu d'un « data paper » ainsi que des exemples de revues qui publient de tels articles.

LES DONNEES DE LA RECHERCHE

Les données de la recherche sont un ensemble d'informations factuelles enregistrées sur des supports, produites ou collectées, selon divers procédés au cours d'un processus de recherche (Cartier, 2015). Elles peuvent être des données d'observation (relevés systématiques, réponses à des enquêtes), expérimentales (produites selon un protocole qui permet de les reproduire), de simulation (produites selon un modèle), compilées (résultant du traitement de données brutes), de référence (certifiées et servant d'appui à une communauté d'intérêt), etc.

L'intérêt des financeurs de la recherche publique pour les données scientifiques va croissant, car celles-ci représentent d'importants enjeux financiers et scientifiques : elles sont coûteuses à produire et représentent un gisement de connaissances actuelles comme futures. Les différents bailleurs de fonds publics français comme étrangers prennent de ce fait des mesures incitatives visant à faciliter le partage de ces données :

- en France, le *Plan d'action* de l'ANR et son *Appel à projets générique 2016* encourage les chercheurs à tirer parti, lorsque cela est possible, des infrastructures et des grandes bases de données existantes, ainsi qu'à promouvoir leurs résultats en accès ouvert ;

- la Commission européenne avec Horizon 2020 et son *Open research data pilot*, demande aux chercheurs de décrire les données dans un plan de gestion des données[1] (*Data Management Plan*) et de rendre accessibles ces données et leurs métadonnées dans un entrepôt ;
- à l'international, l'US National Science Foundation, les Research Councils anglais ou la Netherlands Organisation for Scientific Research exigent depuis plusieurs années déjà des plans de gestion des données pour les recherches qu'ils financent.

Les données de la recherche nécessitent en effet d'être intelligibles pour être réutilisables dans de futures recherches : les chercheurs sont incités à les décrire et à les documenter, notamment à l'aide de métadonnées. Les métadonnées sont un ensemble de données structurées décrivant des ressources physiques ou numériques, et rendant possibles le partage de l'information ainsi que l'interopérabilité des ressources électroniques. Les métadonnées décrivent des données au sein d'entrepôts de données. Ceux-ci sont le plus souvent des serveurs ouverts de grande capacité, généralistes ou thématiques, maintenus par des consortiums d'intérêt (chercheurs, institutions, financeurs, éditeurs). Les métadonnées peuvent être écrites selon plusieurs standards, dont certains sont généralistes et d'autres disciplinaires : balises HTML <meta>, Dublin Core, RDF (Resource description framework), TEI (Text encoding initiative), etc. Elles peuvent être exprimées dans le même format de codage que celui des données qu'elles accompagnent dans un entrepôt.

Les chercheurs sont également incités à exposer leurs données dans des entrepôts ouverts. Lors de la diffusion de données dans un entrepôt, il est nécessaire de vérifier les possibles restrictions de partage/diffusion des données sensibles. Un délai d'embargo ou des restrictions d'utilisation peuvent en effet être fixés pour protéger les données en fonction de divers impératifs : sensibilité des données à caractère personnel ou stratégique, propriété intellectuelle, priorité d'exploitation, etc. En cas de partage des données, il est conseillé d'utiliser un entrepôt public, tel que « Zenodo », l'entrepôt de la Commission européenne, où, *a minima*, un entrepôt dont le propriétaire présente des garanties de pérennité d'accès. A cet égard, la récente loi « pour une République numérique »[2] indique dans son article 30 : « Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'État (...) ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur (…), leur réutilisation est libre ». Le législateur a ainsi souhaité protéger les données produites par la recherche publique d'une captation par des intérêts privés, et fournir à la communauté scientifique les moyens de la pérennité d'accès aux données. Les éditeurs privés ont en effet créé ou racheté des entrepôts de données, comme *SPedia* de Springer, ou *Mendeley Data* de Elsevier.

REDIGER UN *DATA PAPER*

En complément de la description des données par *a)* les métadonnées ainsi que *b)* dans un plan de gestion, il peut être judicieux de rédiger un *data paper* : un *data paper* est un article dans une revue à comité de lecture, décrivant les données d'un projet de recherche. Ce type d'article décrit des données liées à une publication (*underlying research data*) ou indépendantes d'une publication.

L'intérêt du *data paper* est d'exposer des données, en répondant à plusieurs besoins :

[1] Plan de gestion de données : il s'agit d'un document formel précisant la manière dont les données seront produites, traitées, décrites, partagées ou protégées et conservées au cours et à l'issue d'un projet de recherche. C'est un livrable requis par la Commission européenne dans le cadre des programmes Horizon 2020.

[2] La loi « Pour une République numérique » a été promulguée le 9 octobre 2016

- répondre aux exigences des financeurs de la recherche en termes de visibilité et d'accessibilité des résultats scientifiques, comme mentionné plus haut, mais aussi aux recommandations de la communauté scientifique internationale concernant les données ;
- présenter un accès aux données avec un lien vers l'entrepôt choisi, et les rendre intelligibles en les décrivant ;
- fournir une référence citable, car il est publié dans une revue scientifique (*peer-rewieved journal*) ;
- permettre la génération de citations par des services tels que le *Data Citation Index* de Thomson Reuters ;
- permettre la reconnaissance du travail réalisé par l'équipe de recherche qui a produit les données décrites en matérialisant ces données par un article publié.

Tout comme les articles traditionnels, les *data papers* sont structurés de la façon suivante : DOI[3], auteur(s) et affiliation(s), titre, résumé, mots-clés, texte, références bibliographiques.

La particularité du texte d'un *data paper* porte sur la description fine à la fois de la méthode de production des données et des données elles-mêmes, ainsi que sur l'absence de résultat et discussion :

- contexte de la recherche et travaux antérieurs dans lesquels celle-ci s'inscrit, apport des données dans ce contexte et potentiel de réutilisation ;
- protocole de production des données : qualification du producteur des données, méthode de constitution de l'échantillon, matériel utilisé, procédures de traitement, mise en œuvre du contrôle qualité sur les données, questions éthiques soulevées par la collecte de ces données (consentement de patients), etc.
- description du jeu de données : nature ou type de données, format ainsi que version du format le cas échéant, volume de données, date de publication des données dans l'entrepôt choisi par l'auteur ou préconisé par l'éditeur, identifiant des données (attribué par l'entrepôt), lien pérenne vers l'entrepôt choisi, licence d'utilisation attribuée aux données.

L'article regroupe donc un texte spécifique à la description des données, ainsi qu'un lien vers le jeu de données décrit, dans l'entrepôt où celui-ci a été déposé.

Les *data papers* sont publiés soit dans des revues scientifiques traditionnelles, soit dans des revues spécifiques aux données : des *data journals*. Ceux-ci publient uniquement des articles sur les données et n'exposent donc pas de nouvelles analyses.

EXEMPLES DE *DATA JOURNALS*

Les exemples ci-après sont des revues à comité de lecture. Elles sont, pour la plupart, identifiées dans les grandes bases internationales de référence. Elles ne sont pas éditées par des structures soupçonnées d'être des escrocs selon la liste des « éditeurs probablement prédateurs » du bibliothécaire de l'Université du Colorado, Jeffrey Beall[4]. A l'exception de la première citée, elles sont toutes de création très récente, témoignant ainsi du nouvel intérêt porté à ce type de communication scientifique. La même première revue exceptée, toutes les revues citées en exemple sont en accès ouvert (*open access*). Enfin, les frais de publication sont relativement modérés au

[3] DOI : *digital object identifier*, ou identifiant numérique unique de ressource électronique ; ici le DOI identifie un article.
[4] *Beall list* : répertorie les sites susceptibles d'être publiés sur internet par de faux éditeurs, dans le seul but de collecter les frais de publication, sans publication effective en contrepartie https://scholarlyoa.com/publishers/

regard des montants habituellement constatés, qui peuvent atteindre environ 2 000€ à 3 000€ par article (Andro, 2014, Reymonet, 2015).

On peut trouver des listes de *data journals* dans les références suivantes :

- CANDELA et al. Data journals: A survey. *Journal of the Association for Information Science and Technology*, Volume 66, Issue 9, 2015
- AKERS, Katherine. *A Growing List of Data Journals*. Posted on May 9, 2014

Exemples de *data journals* :

Titre journal	**Éditeur**	**Référencement dans les bases**	***Beall list***	***Open access***	**Montant *Article Processing Charges***	**Date début**
Journal of Physical and Chemical Research Data	AIP	WoS	non	non	0 €	1972
Journal of Open Archaeology Data	Ubiquity Press	WoS	non	oui	100 £	2012
Genomics Data	Elsevier	WoS	non	oui	448 €	2013
Geoscience Data Journal	Wiley	WoS	non	oui	1 200 €	2014
Scientific Data	Nature	PubMed	non	oui	1 050 €	2014
Research Data Journal for the Humanities and Social Sciences	Brill	-	non	oui	0 £ jusqu'au 31 déc. 2018	2016

EVOLUTIONS RECENTES ET PERSPECTIVES

On a vu que les financeurs de la recherche exigent de voir les données décrites dans un plan de gestion. À partir d'un plan de gestion de données finalisé, il est relativement aisé de rédiger un *data paper*. En effet, la structure attendue d'un tel article peut trouver sa source dans les items renseignés lors de la préparation d'un plan de gestion. On peut d'ailleurs imaginer qu'un outil de création de plan de gestion puisse exporter les items nécessaires de façon à préfigurer un *data paper*. Par exemple, dans le domaine de la biodiversité, il existe l'outil *Integrated Publishing Toolkit* qui facilite le renseignement des métadonnées et la production automatisée d'un manuscrit de *data paper*.

Dans le programme européen Horizon 2020, le pilote *Open Research Data* est étendu à toutes les thématiques[5] à compter du programme de travail 2017 : l'accès libre devient le statut par défaut pour les données de recherche générées. On peut supposer que la production de *data papers* permettra de renforcer la dimension « dissemination » des projets financés dans le cadre européen. L'introduction de la notion *FAIR data* (*Findable, Accessible, Interoperable, Re-useable*) prolonge la démarche d'ouverture des données impulsée par la Commission européenne. Celle-ci attend en effet des bénéficiaires de ses appels à projet qu'ils rendent leurs données aisées à trouver, ouvertement accessibles, interopérables, réutilisables (en particulier en clarifiant les licences d'utilisation). La notion *FAIR data* est portée par « FORCE11 », une communauté d'académiques, bibliothécaires,

[5] Extension de l'*Open Data Research Pilot* à l'exception des *ERC Proof of concept*, *SME instrument Ph1*, *ERA-NET Cofund* qui ne produisent pas de données, *EJP Cofund.*

archivistes, éditeurs et bailleurs de fonds de la recherche organisés afin de faciliter un meilleur partage des connaissances et favoriser la création de connaissances nouvelles[6].

Enfin, en France, l'article 38 de la récente loi « pour une République numérique », précise que dès lors que les données, liées à une publication, ne sont pas protégées par un droit spécifique et qu'elles ont été rendues publiques par le chercheur, leur réutilisation est libre. L'éditeur d'un écrit scientifique ne peut de ce fait pas limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication. La description, la publication et la réutilisation des données liées à une publication sont entrées dans la loi française, le bien-fondé de leur exposition dans des *data papers* s'en trouve renforcé.

BIBLIOGRAPHIE

ANDRO Mathieu, HOLOGNE Odile, MAHÉ Annaïg. « Estimation des dépenses de publication de l'Inra dans un modèle théorique "Gold Open Access" ». *Documentaliste - Sciences de l'Information*, ADBS, 2014, 51 (4), pp.70-79

AKERS, Katherine. *A Growing List of Data Journals*. Posted on May 9, 2014 https://mlibrarydata.wordpress.com/2014/05/09/data-journals/ Pages consultées le 8 juin 2016

Australian National Data Service http://www.ands.org.au/ Pages consultées le 8 juin 2016

CANDELA Leonardo, CASTELLI Donatella, MANGHI Paolo, TANI Alice. Data journals: A survey. *Journal of the Association for Information Science and Technology*, Volume 66, Issue 9, Version of Record online: 30 JAN 2015 (2015) http://onlinelibrary.wiley.com/doi/10.1002/asi.23358/pdf

CARTIER Aurore, MOYSAN Magalie, REYMONET Nathalie. Réaliser un plan de gestion de données : guide de rédaction (V1), 09/01/2015 https://hal.archives-ouvertes.fr/hal-01138663/file/Realiser_un_DMP_V1.pdf

European commission. *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020* (Version 3.0 26 July 2016) http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

MAUREL Aka Lionel. Quel statut pour les données de la recherche après la loi numérique ? Publication du 3 novembre 2016. *S.I.Lex – Carnet de veille et de réflexion d'un juriste et bibliothécaire* https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/ Page consultée le 9 novembre 2016

République française. *LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique* https://www.legifrance.gouv.fr/affichTexte.do;jsessionid=672F89841A4CE8CD18D3D71A63899368.tpdila11v_3?cidTexte=JORFTEXT000033202746&categorieLien=id

REYMONET Nathalie. *Le coût de publication dans les revues open access à l'Université Paris Diderot*. [Rapport de recherche] Université Paris Diderot-Paris 7. 2015. <sic_01391504>

[6] FORCE11 : https://www.force11.org/group/fairgroup/fairprinciples