Entrepôts de données de recherche : mesurer l'impact de l'Open Science à l'aune de la consultation des jeux de données déposés

Violaine Rebouillat

Entrepôts de données de recherche : mesurer l'impact de l'Open Science à l'aune de la consultation des jeux de données déposés

Violaine Rebouillat, EA 7339 Dicen-IdF, Cnam

Introduction

Les décennies 2000 et 2010 ont vu se développer un nombre croissant de e-infrastructures de recherche, rendant plus aisés le partage et l'accès aux données scientifiques. Cette tendance s'est vue renforcée par l'essor de politiques d'ouverture des données (Chartron, 2018), lesquelles ont donné lieu à une multiplication de réservoirs de données – aussi appelés « entrepôts de données ». Si ces dispositifs créent des conditions favorables à la réutilisation des données, ils ne garantissent toutefois que partiellement leur utilisation réelle. Des études montrent en effet que le partage des données relève de mécanismes sociotechniques complexes, permettant peu de prédire par qui, quand, comment, pourquoi et si les données seront réutilisées (Borgman, 2015 ; Mosconi, 2019). Dans ce contexte, quantifier et qualifier l'utilisation des données rendues publiques constitue un élément essentiel pour évaluer l'impact des politiques d'ouverture des données. Plusieurs approches sont possibles pour tenter de mesurer la réutilisation de ces données :

- première approche : explorer sous forme d'études de cas quelles données sont réutilisées et pour quels usages (Pasquetto, 2018) ;
- deuxième approche : suivre les citations d'un jeu de données dans la littérature scientifique (Missier, 2016 ; Piwowar et Vision, 2013) ;
- troisième approche : mesurer le nombre de consultations dont a fait l'objet un jeu de données dans un entrepôt.

C'est cette troisième approche que se propose d'explorer le présent article.

Nous questionnerons l'utilisation des données déposées dans les entrepôts de la recherche scientifique. L'objectif est d'apporter des éléments de réponse à la question : dans quelle mesure ces données sont-elles consultées et téléchargées ?

Nous commencerons par analyser la façon dont les entrepôts sont définis au sein du mouvement d'ouverture des données scientifiques. Cette réflexion nous permettra de mieux comprendre quel rôle ils sont censés jouer dans la réutilisation des données. Nous présenterons alors les premiers résultats d'une enquête réalisée auprès de 20 entrepôts, au cours de laquelle nous avons étudié les relevés statistiques de consultation des jeux de données. Cette enquête nous a notamment permis d'estimer la proportion des consultations et téléchargements en fonction du nombre de données disponibles dans l'entrepôt et d'étudier l'évolution de cette proportion au fil du temps (entre 2015 et 2020).

1. Les entrepôts de données scientifiques : définition et fonctions

Issu du domaine de l'informatique[1], le terme d'entrepôt de données (*data repository*) a été repris par les politiques d'ouverture des données de la recherche (c'est notamment le cas de la politique instaurée par la Commission européenne dans le cadre de ses programmes de financement à la recherche[2]). Le terme

1 https://fr.wikipedia.org/wiki/D%C3%A9p%C3%B4t_(informatique)

2 Voir par exemple les préconisations du programme H2020 en matière de libre accès (p.6) : Commission européenne (2017). *H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Version 3.2. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

d'entrepôt est aujourd'hui utilisé pour désigner des infrastructures spécifiquement dédiées au stockage et à la communication de données scientifiques. Le répertoire d'entrepôts re3data[3] en propose la définition suivante : « Un entrepôt de données de recherche est un sous-type d'infrastructure informationnelle pérenne, qui permet le stockage à long terme et l'accès aux données de recherche »[4] (Rücknagel et al., 2015). Cette définition présente les entrepôts comme des infrastructures permettant la découverte et l'utilisation des données. Le Réseau de la Bibliothèque Nationale de Médecine (NNLM) aux États-Unis propose, quant à lui, une définition qui met davantage l'accent sur les fonctionnalités de dépôt des données : « Un entrepôt de données peut être défini comme un espace qui contient des données, qui les met à disposition en vue de leur utilisation et qui les organise de manière logique. Un entrepôt de données peut également être défini comme un endroit approprié, spécifique à une thématique, où les chercheurs peuvent soumettre leurs données »[5].

S'il existe des entrepôts nés du contexte actuel d'*open science* (4TU.ResearchData[6], Zenodo[7], Figshare[8]...), les annuaires re3data, Cat OPIDoR[9] et FAIRsharing[10] répertorient aussi sous ce terme des infrastructures antérieures à ce mouvement ou qui, du moins, ne s'en revendiquent pas explicitement. Le principe d'ouverture des données prôné par le mouvement de l'*open science* s'appuie en effet sur une tradition de mise à disposition des données d'ores et déjà inscrite dans certaines communautés scientifiques. C'est notamment le cas des communautés qui utilisent des dispositifs instrumentaux d'envergure pour l'acquisition des données. Le coût de ces grands équipements conduit les communautés de recherche à collaborer autour de la collecte des données et à mettre en commun les données ainsi générées (Rebouillat, 2019). Ce modèle d'organisation s'appuie notamment sur l'infrastructure dématérialisée du web, ainsi que sur des dispositifs de stockage et d'accès à l'information à distance. On pense par exemple à l'astronomie et ses télescopes, à la génomique et à ses séquenceurs, ou encore aux sciences du climat et de l'environnement et à leurs capteurs terrestres, maritimes, aériens et spatiaux (André, 2014).

On constate donc, et telle en est la conséquence, qu'une très grande diversité d'infrastructures sont désignées sous le terme d'entrepôt. Le répertoire FAIRsharing recense aujourd'hui plus de 1300 entrepôts de données. Le répertoire re3data en dénombre quant à lui plus de 2500. On remarquera d'ailleurs que les concepteurs de FAIRsharing associent le terme d'entrepôt (*repository*) à celui de base de données (*database*) sans pour autant expliciter ce qui les différencie (Sansone et al., 2019). La question que l'on peut se poser est donc : à qui revient la responsabilité de définir si telle infrastructure est ou non un entrepôt ? Est-ce aux administrateurs des infrastructures d'en juger, ou bien à ceux qui les financent, ou encore à ceux qui maintiennent des annuaires d'entrepôts comme Cat OPIDoR et FAIRsharing ? Depuis quelques années se dessine un mouvement de certification des entrepôts[11], qui peut-être conduira à une forme de normalisation, y compris en termes de définition.

Cet article propose de considérer les entrepôts comme des dispositifs socio-techniques (Prost et Schöpfel, 2019), dont la spécificité dépend de trois variables :

---

3 https://www.re3data.org/

4 Traduction libre de : « A research data repository is a subtype of a sustainable information infrastructure which provides long-term storage and access to research data ».

5 Traduction libre de : « A data repository can be defined as a place that holds data, makes data available to use, and organizes data in a logical manner. A data repository may also be defined as an appropriate, subject-specific location where researchers can submit their data ».
Source : https://nnlm.gov/data/thesaurus/data-repository

6 https://researchdata.4tu.nl/

7 https://zenodo.org/

8 https://figshare.com/

9 https://cat.opidor.fr

10 https://fairsharing.org/

11 On pense en particulier à l'initiative CoreTrustSeal (https://www.coretrustseal.org/).

- le contenu de l'entrepôt, à savoir les données de recherche qui, en tant qu'objets « portables » (Leonelli, 2015) et dont la désignation est relative à un contexte de recherche particulier (Borgman, 2015), ont une typologie quasi infinie (Pampel et al., 2013) ;
- le périmètre de l'entrepôt (il existe des entrepôts à l'échelle institutionnelle, nationale, disciplinaire, propres à un éditeur…) ;
- les fonctionnalités de l'entrepôt, que Assante et al. (2016) regroupent en huit catégories (de la mise en forme des données à leur documentation, en passant par leur validation, leur mise à disposition, leurs modalités d'accès et de découverte, leur citation, ainsi que l'application d'une licence d'utilisation et le recours à d'éventuels frais de publication).

Partant de cette définition, nous nous sommes demandés si les données mises à disposition dans les entrepôts étaient consultées et si les variables que nous avons définies plus haut pouvaient avoir une influence sur la consultation des données. Pour y répondre, nous avons réalisé une enquête quantitative auprès d'un échantillon d'entrepôts.

2. Méthodologie

Les entrepôts que nous avons étudiés correspondent à la liste des entrepôts « recommandés » par le répertoire FAIRsharing. Porté par l'infrastructure européenne ELIXIR[12], FAIRsharing est une plateforme en ligne qui recense des ressources de types divers (entrepôts de données, standards, politiques de données…) sur le thème du partage des données scientifiques. A la date du 2 mars 2020, il répertoriait 200 entrepôts « recommandés » sur un total de 1373. Cette sélection correspond aux entrepôts ayant été approuvés par une ou plusieurs revues scientifiques et/ou par un ou plusieurs financeurs de la recherche dans le cadre de leur politique de données[13]. Elle rassemble des entrepôts issus de 40 pays et couvre l'ensemble des grands domaines disciplinaires (sciences humaines, sciences sociales, sciences naturelles et sciences de l'ingénieur, selon la répartition proposée par FAIRsharing). En choisissant comme échantillon la liste des entrepôts recommandés par FAIRsharing, nous faisons le postulat de la visibilité de ces entrepôts auprès de la communauté scientifique : nous supposons qu'un chercheur, ayant déposé des données dans un entrepôt à la demande de l'éditeur, sera plus susceptible par la suite de rechercher des données dans cet entrepôt plutôt que dans un autre.

Nous avons contacté par mail les administrateurs de ces 200 entrepôts, leur demandant s'ils pouvaient nous fournir leurs données de consultation sur la période de 2015 à 2020. Par « consultation », nous entendons le nombre de fois où la page descriptive d'un jeu de données a été consultée, mais aussi le nombre de fois où le jeu de données en question a été téléchargé. Nous avons ainsi demandé aux administrateurs de nous communiquer les relevés les plus précis dont ils étaient en possession. Nous avons souhaité rester le plus ouvert possible dans le type de données que nous demandions, ignorant quel mode de collecte était utilisé par chaque entrepôt. Cela nous a permis d'englober un large panel de répondants parmi les entrepôts sondés et de constater la diversité des méthodes de collecte et des données ainsi récoltées.

Parmi les 200 entrepôts sondés, 27 contacts se sont révélés impossibles (souvent en raison d'une adresse mail erronée ou obsolète) et 136 entrepôts n'ont pas donné réponse. Nous avons par ailleurs réceptionné 16 réponses négatives et 22 réponses positives, équivalant à un taux de participation de 13 %. Parmi les réponses négatives, plusieurs motifs ont été avancés, allant du manque de temps et de ressources (6) à

12 https://elixir-europe.org/

13 FAIRsharing définit la notion d'entrepôt « recommandé » dans les termes suivants : « a recommendation is a core-set of resources that are selected or endorsed by data policies from journals, funders or other organizations » (source : https://fairsharing.org/recommendations/). La liste des politiques de données en question est consultable sous le lien suivant : https://fairsharing.org/policies/.

l'absence de collecte de données d'utilisation (3), en passant par le souhait de garder les données d'utilisation confidentielles (5). Enfin, certains administrateurs ont décliné, car ils jugeaient ma demande inadaptée au service qu'ils proposaient (2). Parmi les 22 entrepôts ayant apporté une réponse favorable, 2 nous ont fourni des données qui se sont révélées inexploitables, ramenant ainsi à 20 le nombre d'entrepôts finalement étudiés.

Le tableau 1 présente quelques caractéristiques de ces 20 entrepôts de données. Pour chacun, les informations mentionnées sont extraites des répertoires FAIRsharing et re3data.

	Entrepôt	Date de création	Domaine	Pays	Type d'entrepôt	Se définit comme	Nb de politiques de données (éditeurs, financeurs) recommandant l'entrepôt	Nb de bases de données similaires
Groupe 1	**Antibody Registry** http://antibodyregistry.org	2009	**Biologie**	Etats-Unis	Disciplinaire ; Fournisseur de données	Registry ; Catalog	2	1
	Cancer Imaging Archive http://antibodyregistry.org	2010	**Biologie**	Etats-Unis	Disciplinaire ; Fournisseur de données		12	1
	European Genome-phenome Archive https://ega-archive.org/	2007	**Biologie**	Union européenne	Disciplinaire ; Fournisseur de données ; Fournisseur de services	Archive	34	5
	HUGO Gene Nomenclature Committee https://www.genenames.org/	1979	**Biologie**	Royaume Uni	Disciplinaire ; Fournisseur de données ; Fournisseur de services	Repository	11	15
	NeuroMorpho.org http://neuromorpho.org/	2006	**Biologie**	Etats-Unis	Disciplinaire ; Fournisseur de données ; Fournisseur de services	Inventory	14	1
	Nucleic Acid Database http://ndbserver.rutgers.edu/	1991	**Biologie**	Etats-Unis	Disciplinaire ; Fournisseur de données	Database	5	2
	Worldwide Protein Data Bank http://www.wwpdb.org/	2003	**Biologie**	Etats-Unis, Royaume Uni, Japon	Disciplinaire ; Autre ; Fournisseur de données ; Fournisseur de services	Databank ; Archive ; Repository	46	26
	British Oceanographic Data Centre http://www.wwpdb.org/	1989	**Géosciences**	Royaume Uni	Institutionnel ; Fournisseur de données	Data Centre ; Database	10	6
	Centre for Environmental Data Analysis Archive http://archive.ceda.ac.uk/	2016	**Géosciences**	Royaume Uni	Disciplinaire ; Fournisseur de données ; Fournisseur de services	Archive ; Data repository	10	8
	Coherent X-ray Imaging Data Bank http://archive.ceda.ac.uk/	2010	**Géosciences**	Suède, Etats-Unis	Disciplinaire ; Fournisseur de données	Databank ; database	19	0
	EarthChem http://earthchem.org/library	2003	**Géosciences**	Etats-Unis	Disciplinaire ; Fournisseur de données ; Fournisseur de services	Data Repository	10	3
	PetDB http://earthchem.org/petdb	2003	**Géosciences**	Etats-Unis	Disciplinaire ; Fournisseur de données ; Fournisseur de services	Database	10	3
	SEANOE https://www.seanoe.org/	2015	**Géosciences**	France	Disciplinaire ; Fournisseur de données	Publisher	13	0
	DANS-EASY https://easy.dans.knaw.nl/ui/home	2006	**Pluridisciplinaire**	Pays-Bas	Autre ; Fournisseur de données	Archiving system	2	1
	Open Science Framework http://osf.io	2011	**Pluridisciplinaire**	Etats-Unis	Autre ; Fournisseur de données	Platform	15	5
	Swedish National Data Service https://snd.gu.se/en	2008	**Pluridisciplinaire**	Suède	Autre ; Fournisseur de données ; Fournisseur de services	Research Data Catalog	1	0
1-2	**FlowRepository** http://www.flowrepository.org	2012	**Biologie**	Etats-Unis	Disciplinaire ; Fournisseur de données	Database	15	0
Groupe 2	**ArrayExpress** https://www.ebi.ac.uk/arrayexpress/	2003	**Biologie**	Union européenne, Royaume-Uni, Etats-Unis	Disciplinaire ; Fournisseur de données ; Fournisseur de services	Repository ; Database ; Archive	55	17
	Donders Repository http://data.donders.ru.nl	2016	**Biologie**	Pays-Bas	Disciplinaire ; Fournisseur de données	Repository	1	0
	Australian Ocean Data Network Portal https://portal.aodn.org.au	2011	**Géosciences**	Australie	Disciplinaire ; Institutionnel ; Autre ; Fournisseur de données ; Fournisseur de services	Portal ; Access Point	1	0
	Dryad https://datadryad.org/	2012	**Pluridisciplinaire**	Etats-Unis	Autre ; Fournisseur de services ; Fournisseur de données	Repository	35	2

Tableau 1 : Caractéristiques des 20 entrepôts étudiés

On constate qu'il s'agit principalement d'entrepôts disciplinaires, à l'exception de DANS-EASY, Dryad, Swedish National Data Service et Open Science Framework, classés par re3data dans la catégorie « autre », et du British Oceanographic Data Centre, classé comme entrepôt institutionnel, toujours selon la classification du re3data, ainsi que du Australian Ocean Data Network Portal, qui est à la fois catégorisé comme entrepôt disciplinaire, institutionnel et de type autre.
Dix des 20 entrepôts sont à la fois fournisseurs de données et fournisseurs de services[14], c'est-à-dire qu'ils mettent à disposition des données de recherche et leurs métadonnées et qu'ils récoltent à la fois les métadonnées de données de recherche auprès d'autres fournisseurs de données.
La biologie est la discipline la plus représentée (10), ce qui coïncide avec la composition globale de FAIRsharing[15] (tout comme de re3data), où les entrepôts de biologie sont majoritaires.
Tous, parmi les 20 entrepôts, proposent une interface en langue anglaise. Ils sont gérés par des instances européennes et nord-américaines, à l'exception de Australian Ocean Data Network Portal (Australie) et de Worldwide Protein Data Bank (qui associe le Japon à l'Union européenne et aux États-Unis). La majorité d'entre eux (9) sont des entrepôts maintenus aux États-Unis.

Les données de consultation que les responsables d'entrepôts nous ont transmises ont été collectées soit à l'aide de Google Analytics, soit à partir d'un script informatique développé en interne par les entrepôts. Nous avons homogénéisé ces données, l'objectif étant de les traiter sur un plan statistique. Seules celles qui pouvaient être comparées entre elles ont été retenues, à savoir :

- le nombre de données disponibles ;
- le nombre de pages consultées ;
- le nombre d'utilisateurs ;
- le nombre de téléchargements ;
- et le nombre de citations.

Nous avons ensuite analysé les entrepôts par groupe, en fonction de la nature des données que chacun nous avait communiquées. Certains nous ont en effet fourni des relevés annuels (comme, par exemple, le nombre total de téléchargements comptabilisés chaque année entre 2015 et 2020) ; d'autres nous ont communiqué des données cumulées (par exemple, l'ensemble des téléchargements dont a fait l'objet un jeu de données, depuis la mise en place du système de relevé).

Deux groupes ont ainsi été constitués :

- un premier groupe d'entrepôts, dont l'audience a pu être étudiée sur un plan chronologique : celui-ci est constitué de Antibody Registry, British Oceanographic Data Centre, FlowRepository, HUGO Gene Nomenclature Committee, NeuroMorpho.org, Nucleic Acid Database, Open Science Framework, SEANOE, Coherent X-ray Imaging Data Bank, EarthChem, PetDB, European Genome-phenome Archive, Centre for Environmental Data Analysis Archive, Cancer Imaging Archive, DANS-EASY, Worldwide Protein Data Bank et Swedish National Data Service ;
- un second groupe d'entrepôts, dont les données de consultation ont fait l'objet d'un relevé ponctuel à une date t, à savoir : ArrayExpress, Dryad, Australian Ocean Data Network Portal, Donders Repository et FlowRepository (qui fait partie des deux groupes).

La partie qui suit présentera d'abord les résultats issus de l'analyse des données du premier groupe, avant d'exposer ceux du second groupe.

14 Voir la définition donnée par Rücknagel et al. (2015, p.20).
15 Voir l'histogramme intitulé *Top 10 disciplines covered by databases* dans les statistiques de FAIRsharing (https://fairsharing.org/summary-statistics/?collection=all)

Avant de présenter ces résultats, il convient néanmoins d'en souligner les possibles limites. Les données que nous avons récoltées contiennent en effet un certain nombre de biais, qui font qu'elles ne reflètent pas exactement la consultation réelle des entrepôts. Nous ne pouvons déceler tous ces biais, car nous ne connaissons pas assez précisément le mode de collecte des données d'utilisation. Les administrateurs d'entrepôts qui nous ont envoyé leurs relevés d'utilisation n'ont en effet pas toujours explicité la méthode qu'ils employaient pour collecter ces informations. Lorsqu'ils l'ont fait, néanmoins, cela nous a permis d'identifier plusieurs biais, parmi lesquels :

- la présence d'activités pas ou peu significatives
  Les relevés de pages consultées et de téléchargements peuvent en effet comptabiliser des activités que l'on jugera non significatives pour cette étude. Il peut s'agir du passage de robots, des activités d'administration de l'entrepôt (contrôles, validations…), des activités de formation (pendant lesquelles un même jeu de données est consulté par un grand nombre de personnes), des consultations de pages annexes aux jeux de données (Google Analytics comptabilise par exemple dans ses « pages views » l'ensemble des pages d'un site web, y compris la page d'accueil, la page de recherche ou la page de contact). Certains entrepôts comme Seanoe tentent de filtrer ces activités non significatives, mais ce n'est pas le cas de tous les entrepôts que nous avons étudiés.

- les interruptions dans la collecte des relevés d'utilisation
  A l'inverse, il arrive qu'il y ait des lacunes dans les données d'utilisation. Celles-ci peuvent être le fait d'interruptions ponctuelles liées par exemple à un incident technique ou à la migration de l'entrepôt vers une nouvelle solution logicielle.

- le passage répété d'un même utilisateur
  Comme l'expliquait l'administrateur d'un des entrepôts étudiés, le recensement du nombre d'utilisateurs est parfois faussé par la méthode de calcul utilisée, celle-ci pouvant compter plusieurs fois un même utilisateur : « Le défi est que nous ne pouvons pas stocker de données qui pourraient être considérées comme personnelles, par exemple le numéro IP de l'ordinateur ou les fichiers qui intéressent la personne qui les a téléchargés. C'est pourquoi nous utilisons le concept de "visiteurs uniques" : en accédant à l'interface web des métadonnées [...] ou à l'interface webdav [...], chaque visiteur est "identifié" à l'aide de son adresse IP et des détails de son navigateur. Cette empreinte digitale est utilisée pour suivre son comportement, pour compter le nombre de pages de données qu'il consulte et le nombre de fichiers qu'il télécharge. Après une heure, l'empreinte digitale est supprimée. Ainsi, la même personne consultant l'entrepôt avec le même ordinateur le jour 1 et le jour 2 sera comptée comme deux visiteurs. »

N'ayant pu écarter ces biais, nous pouvons donc seulement en notifier la possible présence dans les résultats que nous présentons. Ces derniers ne doivent, par conséquent, être considérés qu'à titre indicatif.

3. Résultats

3.1. Une consultation en hausse ?

Nous présentons ici les résultats issus de l'analyse du premier groupe de 17 entrepôts. Ce groupe a permis d'étudier l'évolution de la consultation des données dans les entrepôts sur une période de quatre ans (de 2015 à 2019). Les relevés individuels de chaque entrepôt sont rassemblés dans l'annexe en ligne[16].

16Rebouillat, Violaine (2020). *Audience data from 17 research data repositories between 2015 and 2020*. Jeu de données. Zenodo. http://doi.org/10.5281/zenodo.4008305

En termes de données téléchargées, on observe des ordres de grandeur très différents selon les entrepôts. Un entrepôt comme Seanoe comptabilise un nombre de téléchargements annuels de l'ordre du millier, tandis que des entrepôts comme Cancer Imaging Archive ou Worldwide Protein Data Bank recensent des taux de téléchargements de l'ordre de la centaine de millions. Il faut noter que ces écarts se reflètent également dans le nombre de données disponibles. Ces relevés restent toutefois à nuancer pour un entrepôt comme Open Science Framework qui contient certes plusieurs milliers de données et recense plusieurs millions de téléchargements, mais qui n'est pas un entrepôt dédié uniquement à l'hébergement de données de recherche. Des prépublications, posters, communications et autres produits de la recherche peuvent y être déposés. Pour comparer ces entrepôts très différents entre eux, nous avons donc calculé un coefficient d'évolution : pour un téléchargement, une vue ou une donnée disponible en 2015, combien y en a-t-il eu en 2016, 2017, 2018 et 2019 ? Cette méthode nous a ainsi permis d'établir des graphiques comparatifs d'évolution.

3.1.1. Evolution du nombre de pages consultées

La figure 1 présente l'évolution du nombre de pages vues pour 8 entrepôts. Entre 2015 et 2019, ces 8 entrepôts ont tous globalement connus une augmentation du nombre de pages vues. L'évolution la plus significative est celle de Seanoe : les *landing pages* des jeux de données ont en moyenne été consultées 2,4 fois plus en 2017 qu'en 2016 et 6,9 fois plus entre 2016 et 2019. HUGO Gene Nomenclature Committee est l'entrepôt dont le nombre de pages vues est resté le plus constant, celui-ci n'ayant augmenté que de 110 % entre 2015 et 2019.
Deux groupes semblent se dessiner. Le premier rassemble Antibody Registry, HUGO Gene Nomenclature Committee, Nucleic Acid Database et Bristish Oceanographic Data Centre, dont la courbe d'évolution augmente peu entre 2015 et 2019 (le nombre de pages vues a été multiplié au maximum par 2). Le second groupe se compose de Seanoe, FlowRepository et Open Science Framework qui présentent, quant à eux, une croissance plus marquée du nombre de pages vues (le nombre de pages vues a été multiplié au minimum par 3,5 et au maximum par 6,9). De ces deux groupes on exclura NeuroMorpho, dont la courbe atypique reste difficilement interprétable à l'aide des seules données disponibles. Il est intéressant de noter que le premier groupe rassemble des entrepôts ayant vu le jour avant 2010 (entre 1979 et 2009), alors que le second groupe se compose d'entrepôts plus récents (postérieurs à 2012). On peut se demander si cela reflète une évolution générique des entrepôts. Ceux-ci connaîtraient à leurs débuts une croissance soutenue du nombre de consultations. Le lancement d'un entrepôt s'accompagne en effet souvent d'actions de communication visant à le faire connaître et contribuant probablement à sa consultation par un nombre croissant de visiteurs. Les entrepôts atteindraient ensuite une forme de palier, correspondant à la stabilisation du nombre de consultations sur le plus long terme. Pour vérifier cette hypothèse, il serait donc intéressant de connaître l'évolution du nombre de consultations des quatre entrepôts antérieurs à 2010 depuis leur création.

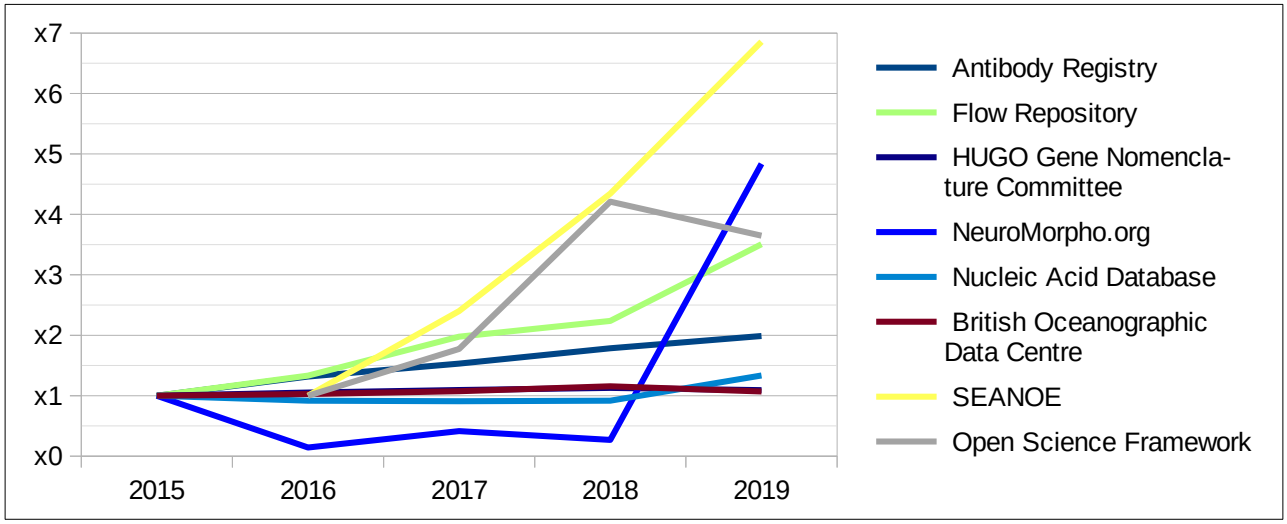

Figure 1 : Evolution de la consultation des pages descriptives des jeux de données

3.1.2. Evolution du nombre de téléchargements

La figure 2 présente l'évolution du nombre de téléchargements pour 12 entrepôts. Là encore le constat est à l'augmentation globale entre 2015 et 2019. Deux groupes semblent se dessiner. Les entrepôts Cancer Imaging Archive, Swedish National Data Service et Seanoe constituent un premier groupe pour lequel le nombre de téléchargements a connu une explosion en 5 ans (pour Cancer Imaging Archive, en 2019, les téléchargements sont 8,5 fois plus nombreux qu'en 2015). Le second groupe est composé des 9 autres entrepôts. Pour ces derniers, l'évolution du nombre de téléchargements est plus aléatoire, mais elle se traduit globalement par une multiplication du nombre de téléchargements par 1,6 en moyenne entre 2015 et 2019. FlowRepository a par exemple vu le nombre de ses téléchargements doubler, tout comme European Genome-phenome Archive.

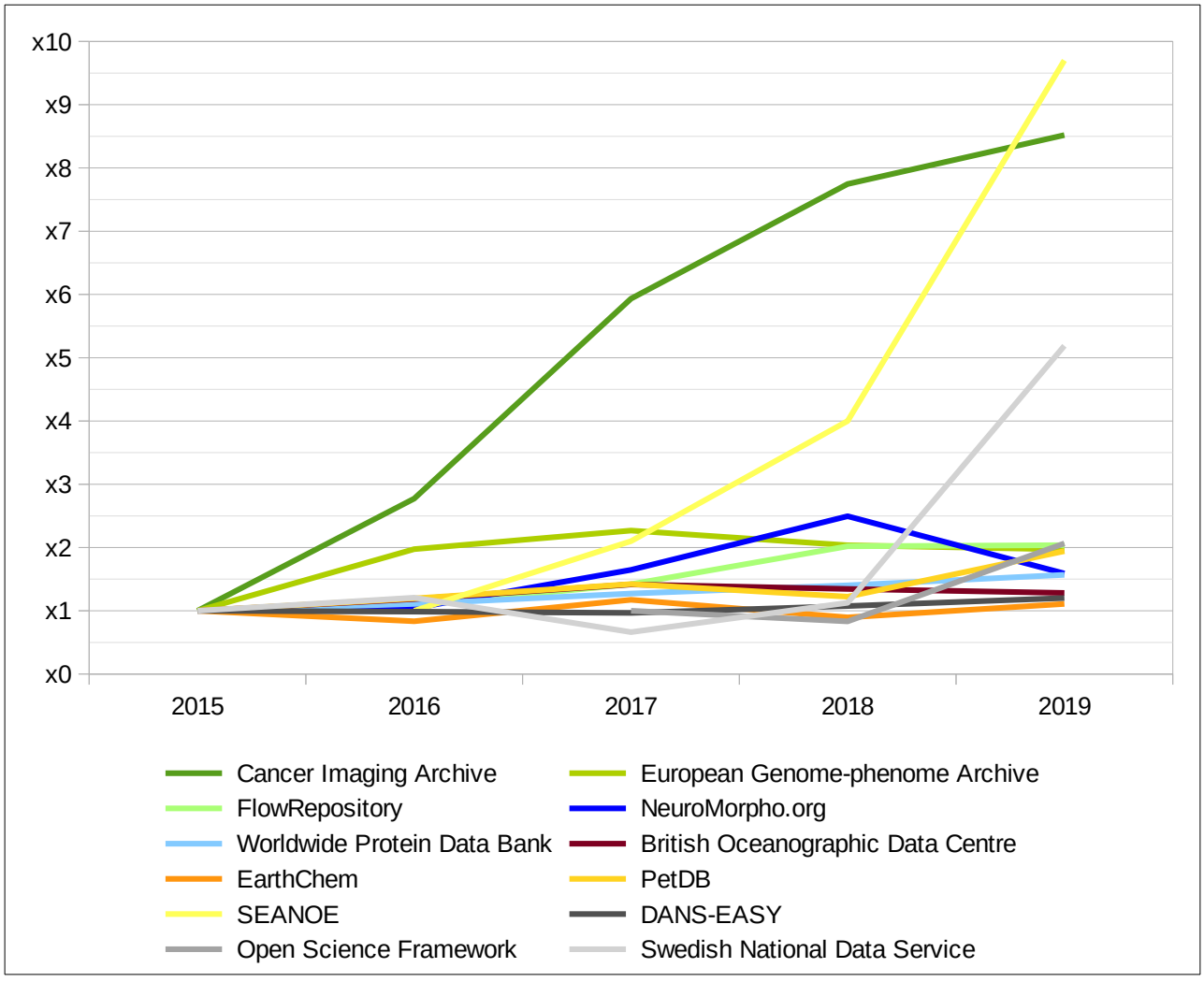

Figure 2 : Evolution des téléchargements de jeux de données

On notera que l'évolution du nombre de pages vues et celle des téléchargements ne sont pas corrélées. A l'exception de Seanoe, les entrepôts dont le nombre de pages vues croît fortement (Flow Repository et Open Science Framework) n'ont pas connu une augmentation semblable du nombre de téléchargements. Cet écart traduirait deux usages distincts, avec d'un côté des utilisateurs qui consulteraient et téléchargeraient des données et, de l'autre, des utilisateurs qui consulteraient mais ne téléchargeraient pas les données. Cette hypothèse nous amène à nous interroger sur la part des consultations attribuable aux robots, part que l'on peut supposer grandissante étant donné l'intérêt croissant porté à l'interopérabilité des plateformes.

3.1.3. Evolution du nombre de téléchargements et de pages consultées par rapport au nombre de données disponibles dans l'entrepôt

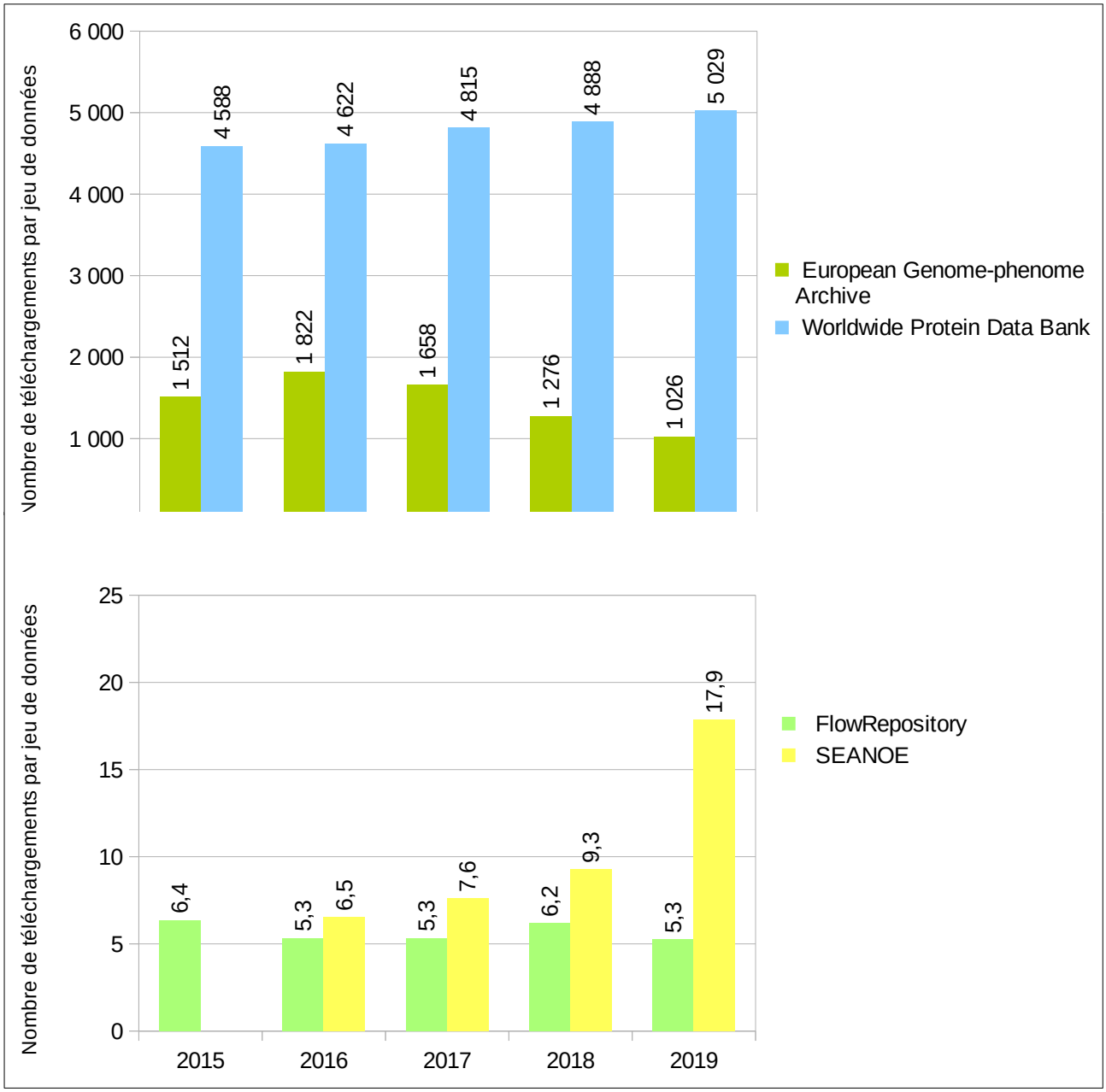

Figure 3 : Moyenne du nombre de téléchargements par jeu de données

La figure 3 montre que le nombre de vues et de téléchargements est toujours supérieur au nombre de données disponibles dans l'entrepôt, quelque que soit son ordre de grandeur. Cela signifie qu'un jeu de données est téléchargé en moyenne plusieurs fois. On observe par ailleurs une différence entre fournisseurs de données et fournisseurs de services (tableau 1). Il semblerait en effet que les données mises à disposition par les entrepôts qui sont à la fois fournisseurs de données et fournisseurs de services (ici European Genome-phenome Archive et Worldwide Protein Data Bank) soient davantage téléchargées que les données des entrepôts qui ne sont que fournisseurs de données (Seanoe et FlowRepository). Agréger des données d'autres entrepôts confèrerait-il davantage de visibilité ? Dans ce cas précis, il semblerait que l'attractivité de European Genome-phenome Archive et de Worldwide Protein Data Bank provienne non seulement de leur volumétrie mais aussi de leur rang institutionnel. European Genome-phenome Archive est en effet porté par le Laboratoire Européen de Biologie Moléculaire (EMBL)[17], qui est un organisme de recherche européen

17https://www.embl.org/

financé par les États membres. L'entrepôt fait par ailleurs partie de l'infrastructure européenne ELIXIR[18]. Worldwide Protein Data Bank est quant à lui géré par une organisation internationale, qui fédère des centres de données issus des différents continents américain, asiatique et européen. Dans les deux cas, on a donc affaire à des entrepôts qui rassemblent une communauté internationale. Le portage institutionnel par des organismes de rang mondial pourrait donc avoir une influence positive sur la consultation et le téléchargement des données qu'ils référencent.

En termes d'évolution, on constate que le nombre de téléchargements augmente de façon non proportionnelle par rapport au nombre de données disponibles.
On observe trois cas de figure.

- Le premier cas de figure concerne FlowRepository : pour cet entrepôt, le nombre de téléchargements par donnée reste relativement stable entre 2015 et 2019. Cela signifie que dépôts et téléchargements augmentent au même rythme.
- Le second cas de figure est représenté par European Genome-phenome Archive : ici le nombre de téléchargements par donnée diminue chaque année. Le nombre de téléchargements est resté constant entre 2016 et 2019 (environ 5 500 000), tandis que le nombre de données disponibles a été multiplié par 1,8 (il est passé de 3015 à 5346).
- Le troisième cas figure se rapporte à Seanoe et Worldwide Protein Data Bank : le nombre de téléchargements croît plus vite que le nombre de dépôts. Les données de ces deux entrepôts sont donc en moyenne davantage téléchargées en 2019 qu'en 2015.

Pour ce dernier cas de figure, on peut s'interroger sur l'origine du phénomène observé : est-il imputable aux seuls robots, de plus en plus nombreux à extraire des données des entrepôts ? Quelle est la part des téléchargements « manuels » et isolés ?

3.1.4. Evolution du nombre de citations

Quant au nombre de citations, seuls 4 entrepôts ont été en mesure de nous fournir cette information. Le tableau 2 montre que les jeux de données de ces entrepôts ont reçu un nombre croissant de citations au cours des cinq dernières années. Les informations dont nous disposons ne précisent toutefois pas quels jeux de données en particulier font l'objet de citations dans la littérature scientifique. Les figures 4a à 4d permettent de comparer, pour quatre entrepôts, l'évolution du nombre de citations par rapport à celle des utilisateurs ou pages vues, des téléchargements et des données disponibles. On observe une augmentation des citations similaire voire plus constante (pour Seanoe et European Genome-phenome Archive) et plus prononcée (pour EarthChem) que celle des téléchargements. Le nombre d'articles citant des données issues des entrepôts ne serait donc pas strictement corrélé au nombre de téléchargements de ces données. Deux hypothèses, probablement complémentaires, peuvent être avancées pour expliquer ce phénomène. La première hypothèse est l'expansion des pratiques de citation des jeux de données dans les publications scientifiques. Robinson-Garcia et al. (2015) montraient dans une étude en 2015 que ces pratiques restaient peu répandues mais qu'elles avaient connu un développement dans les domaines de la cristallographie et de la génomique. Cette hypothèse est d'autant plus probable qu'il existe aujourd'hui des systèmes d'identification pérenne (comme les DOI[19]), qui permettent à la fois de citer plus facilement un jeu de données et de repérer plus aisément sa mention dans un article. La seconde hypothèse est que, dès lors qu'un utilisateur a téléchargé un jeu de données, il peut l'utiliser indéfiniment, un an, deux ans ou dix ans après. Il peut en tirer une ou plusieurs publications, le combiner avec d'autres données... Le nombre d'utilisations d'un jeu de données ne transparaîtrait donc pas dans le nombre de téléchargements. Un téléchargement peut par exemple équivaloir à cinq citations dans cinq articles différents du même auteur.

18 https://elixir-europe.org/
19 Digital Object Identifiers (https://www.doi.org/)

European Genome-phenome Archive				
	Comptes utilisateurs (cumulatif)	**Téléchargements** (total annuel)	**Citations** (total annuel)	**Données disponibles** (cumulatif)
2015	7 327	2 782 511	**192**	1 840
2016	8 999	5 493 377	**272**	3 015
2017	11 158	6 310 177	**335**	3 805
2018	14 154	5 662 103	**398**	4 436
2019	16 173	5 485 556	**451**	5 346

EarthChem			
	Utilisateurs (total annuel)	**Téléchargements** (total annuel)	**Citations** (total annuel)
2015	582	1 076	**23**
2016	568	898	**29**
2017	503	1 262	**27**
2018	600	963	**42**
2019	769	1 193	**100**

PetDB			
	Utilisateurs (total annuel)	**Téléchargements** (total annuel)	**Citations** (total annuel)
2015	1 096	2 674	**45**
2016	1 325	3 196	**63**
2017	1 341	3 801	**63**
2018	1 226	3 272	**61**
2019	1 453	5 194	**77**

SEANOE				
	Vues (total annuel)	**Téléchargements** (total annuel)	**Citations** (total annuel)	**Données disponibles** (cumulatif)
2016	3 500	1 000	**22**	153
2017	8 400	2 100	**76**	276
2018	15 200	4 000	**112**	431
2019	24 000	9 700	**154**	543

Tableau 2: Total du nombre de pages vues, d'utilisateurs, de téléchargements, de citations et de données disponibles par entrepôt par an

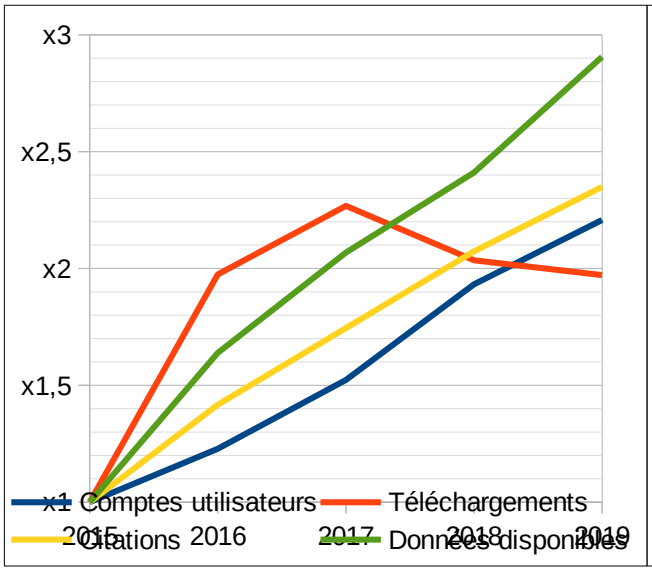

Figure 4a: European Genome-phenome Active

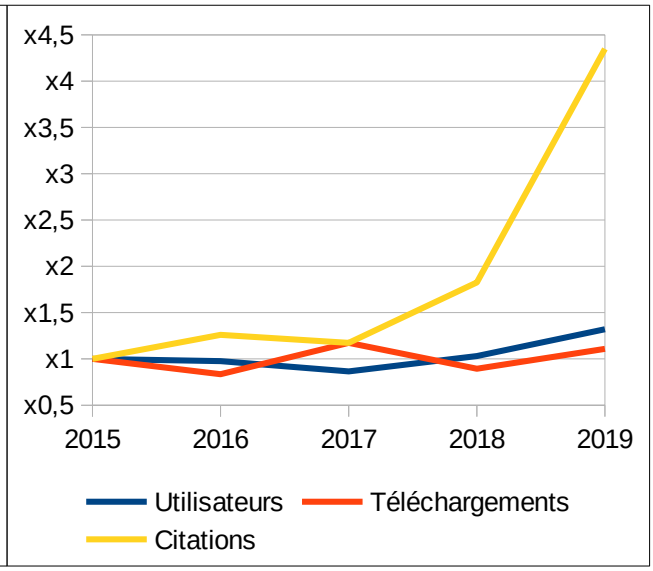

Figure 4b: EarthChem

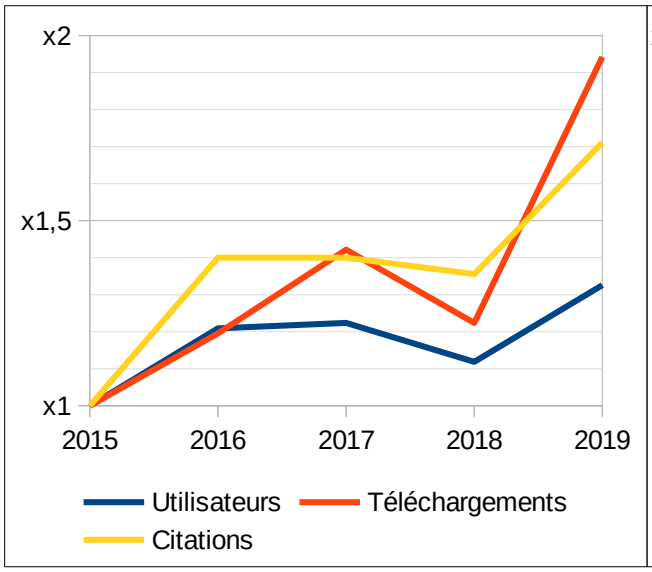

Figure 4c: PetDB

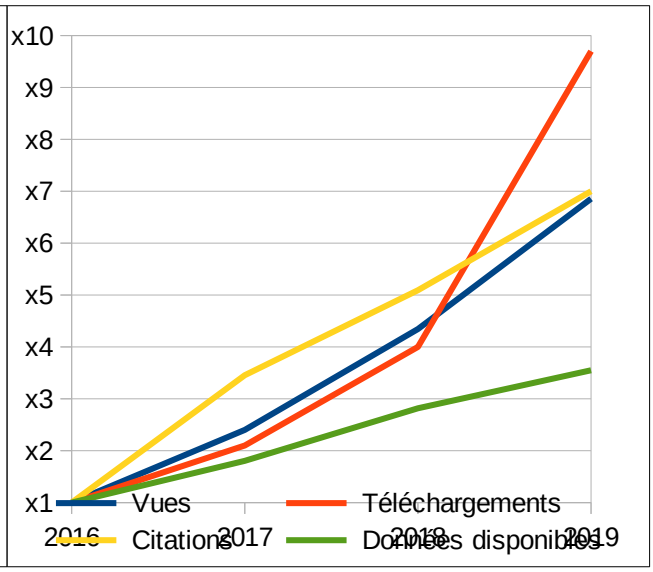

Figure 4d: SEANOE

Figures 4a à 4b: Evolution des consultations, téléchargements, citations et jeux de données disponibles par entrepôt

3.1.5. Conclusion

De manière générale, on observe donc une tendance positive, à tout le moins constante, dans l'utilisation des 17 entrepôts étudiés entre 2015 et 2019. Si l'évolution du nombre de vues, de téléchargements et de citations se traduit par une augmentation entre 2015 et 2019, il faut néanmoins rester prudent dans cette constatation, la période représentée étant probablement trop restreinte pour une analyse de ce type. Une étude sur une période de 10 à 20 ans serait probablement plus révélatrice.
Comment expliquer cette augmentation ? Trois raisons peuvent être avancées.

- La première serait la visibilité croissante des entrepôts, leur existence étant amenée à la connaissance d'un nombre de plus en plus grand d'utilisateurs potentiels. A ce paramètre s'ajouterait celui de l'utilité perçue des entrepôts et de la confiance que les utilisateurs y mettent (Yoon, 2014).
- La seconde serait le développement de nouvelles méthodes de recherche, axées sur l'analyse de données massives. L'essor de la bioinformatique en est un exemple (Rung et Brazma, 2012). Cette composante de la biologie se nourrit, parfois exclusivement, des données disponibles dans les entrepôts de sciences omiques comme ArrayExpress[20] et Gene Expression Omnibus[21].

Ce constat d'augmentation semble donc donner une légitimité aux entrepôts en tant que dispositifs informationnels. On peut néanmoins s'interroger sur ce qu'auraient révélé les statistiques d'utilisation des 179 entrepôts qui n'ont pas répondu à l'enquête.

3.2. Une consultation asymétrique ?

Si les données du premier groupe nous ont permis de suivre l'évolution des consultations, le second groupe (5 entrepôts) nous donne quant à lui, pour chaque jeu de données, un relevé individuel du nombre de téléchargements. Ce groupe est complémentaire du premier, car il permet d'étudier la part des données réellement téléchargées.

Le tableau 3 présente pour chacun des cinq entrepôts un aperçu du nombre de données disponibles et du nombre total de téléchargements. On remarque des ordres de grandeur très différents entre, d'un côté, Dryad et ArrayExpress et, de l'autre côté, FlowRepository, Donders Repository et Australian Ocean Data Network Portal. Il reste néanmoins assez difficile de comparer ces données entre elles, l'unité de mesure variant selon les choix de chaque entrepôt (Donders Repository et Australian Ocean Data Network Portal mesurent des collections de données, tandis que ArrayExpress et FlowRepository mesurent des expériences en laboratoire et Dryad des données de taille et de nature relativement variables).

	ArrayExpress	Dryad	Australian Ocean Data Network Portal	Donders Repository	FlowRepository
Date de création de l'entrepôt	2003	2012	2011	2016	2012
Période du relevé	2003 – 2018	2012 – 2020	2017 – 2020	2016 – 2020	2014 – 2020
Nombre de jeux de données disponibles	75 195	32 596	273	112	1 046
Nombre de téléchargements	6 852 219	1 663 902	26 772	208 946	27 534
Moyenne (nb de téléchargements par jeu de données)	91	51	98	1 866	26
Médiane	59	19	NC	0	5
% de jeux concernés par 80 % des téléchargements	44%	28%	19%	3%	18%

Tableau 3: Total des téléchargements et des jeux de données disponibles par entrepôt

20 https://www.ebi.ac.uk/arrayexpress/
21 https://www.ncbi.nlm.nih.gov/geo/

3.2.1. Répartition des jeux de données en fonction du nombre de téléchargements

A partir des relevés individuels de téléchargements, il a été possible de calculer la répartition des jeux de données en fonction du nombre de fois où ceux-ci ont été téléchargés. Nous reportons ici l'exemple de ArrayExpress (figure 5a) et de Donders Repository (figure 5b).

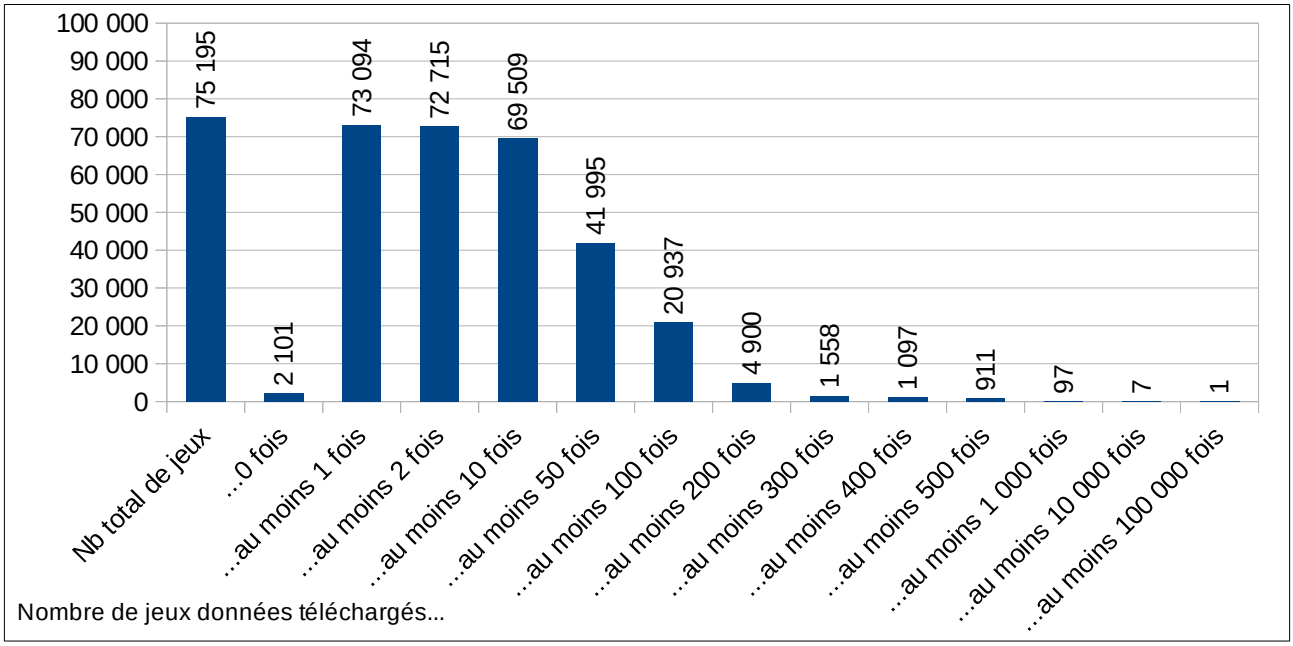

Figure 5a : ArrayExpress

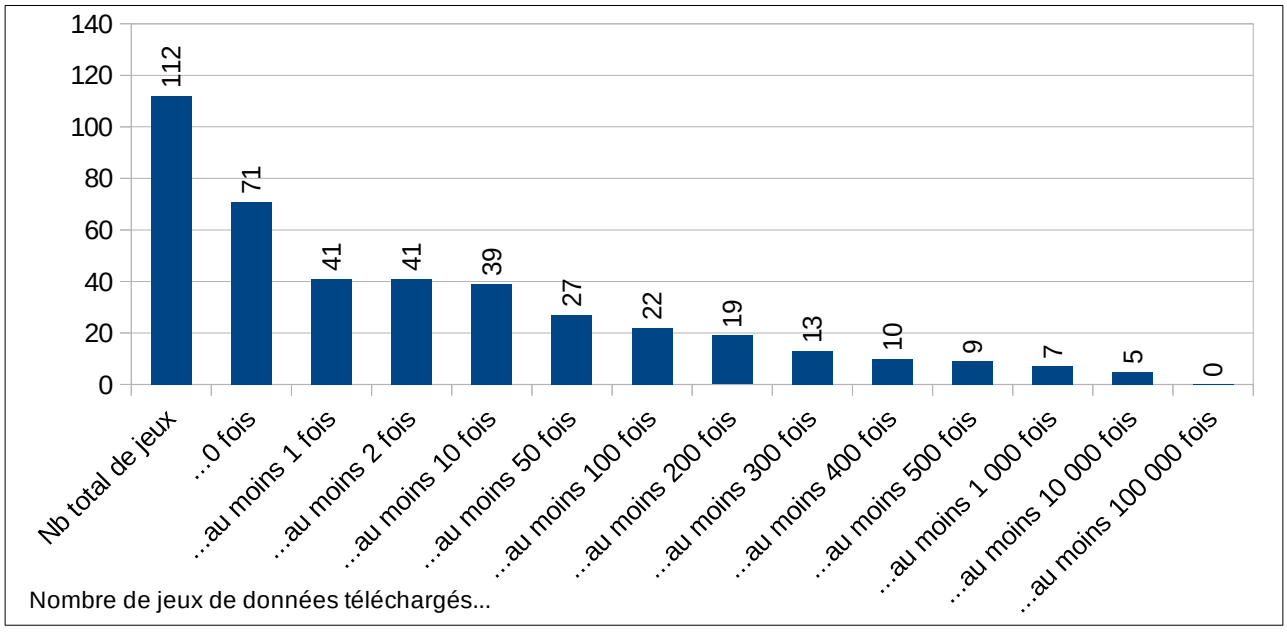

Figure 5b : Donders Repository

Figures 5a et 5b : Répartition des jeux de données en fonction du nombre de fois où ils ont été téléchargés

On constate une répartition inégale des téléchargements. Dans l'entrepôt Donders Repository, la part de données jamais téléchargées est importante : elle représente 63% du total des données. Dans ArrayExpress, cette proportion est bien plus faible : elle concerne 2 101 jeux de données sur un total de 75 195, soit 2,8 %

des données de l'entrepôt. Pour ArrayExpress, la concentration des téléchargements se situe plutôt au niveau des jeux de données téléchargés entre 10 et 200 fois : ceux-ci sont au nombre de 64 609, ce qui représente 85,9 % des données de l'entrepôt.
On peut également s'intéresser à la part que représentent les données les plus téléchargées. Le jeu de données le plus téléchargé de l'entrepôt ArrayExpress s'intitule E-GEOD-63525. En 2018 (date du dernier relevé), il avait fait l'objet de 203 175 téléchargements. Parmi les records de téléchargements, on trouve également six autres jeux de données qui ont, quant à eux, été téléchargés entre 14 000 et 85 000 fois. Ces sept jeux de données, qui représentent 0,01 % du total, concentrent donc 40 683 téléchargements, soit 6 % du total des téléchargements. Pour l'entrepôt Donders Repository, le jeu de données le plus téléchargé s'intitule « Stimulus-induced gamma power predicts the amplitude of the subsequent visual evoked response »[22]. Il a fait l'objet de 71 593 téléchargements. Après lui, 4 autres jeux de données ont été téléchargés entre 15 000 et 54 000 fois. Ce qui signifie que 5 jeux de données (4,5 % du total) concentrent 201 035 des 208 946 téléchargements (soit 93 % de l'ensemble des téléchargements comptabilisés par l'entrepôt).
L'inégale répartition des téléchargements n'est pas spécifique à ArrayExpress et Donders Repository ; elle vaut aussi pour Dryad et FlowRepository[23]. La médiane du nombre de téléchargements par jeu de données (tableau 3) en est un indicateur : dans chacun des quatre cas, elle est inférieure à la moyenne des téléchargements. Ce qui signifie que la moitié des jeux de données concentre moins de la moitié du total des téléchargements et donc que la seconde moitié des jeux de données reçoit la majeure partie des téléchargements.

3.2.2. Concentration des téléchargements et loi de Pareto

Dans la figure 6, le nombre total de données et le nombre total de téléchargements ont été ramenés à 100, de façon à pouvoir comparer les cinq entrepôts entre eux. On constate que la moitié des téléchargements (50%) concerne moins de 20 % des données. Pour Dryad, FlowRepository, Donders Repository et Australian Ocean Data Network Portal, il s'agit même de moins de 10 % des données. Ce qui signifie qu'une minorité de données reçoit la majorité des téléchargements. Cette courbe exponentielle rappelle la loi de Pareto et son application en bibliométrie. Comme le résument Larivière et Sugimoto (2018), plus ou moins 20 % des publications scientifiques reçoivent 80 % des citations. « Les taux de citation sont asymétriques à l'extrême, on l'a vu : quelques documents seulement reçoivent la grande majorité des citations. [...] L'un des moyens bien connus pour étudier la concentration est la règle du 80-20 (ou loi de Pareto) qui, lorsqu'elle est appliquée aux citations, montre que 80% des citations devraient provenir de 20% des articles. Cela reflète assez bien les tendances observées, même si le chiffre exact varie selon la discipline : 80% des citations renvoient à 12% des documents en arts et en humanités, à 16% en sciences sociales, à 24% en sciences naturelles et à 26% en sciences médicales. Une partie de ces différences entre disciplines s'explique par les articles non cités, qui représentent une proportion considérable des publications en arts et en humanités, et une petite proportion seulement dans les autres domaines ».

22 https://doi.org/10.34973/17xm-e258

23 Quant au cinquième entrepôt, Australian Ocean Data Network Portal, nous n'avons pas été en mesure de calculer la médiane de ses téléchargements.

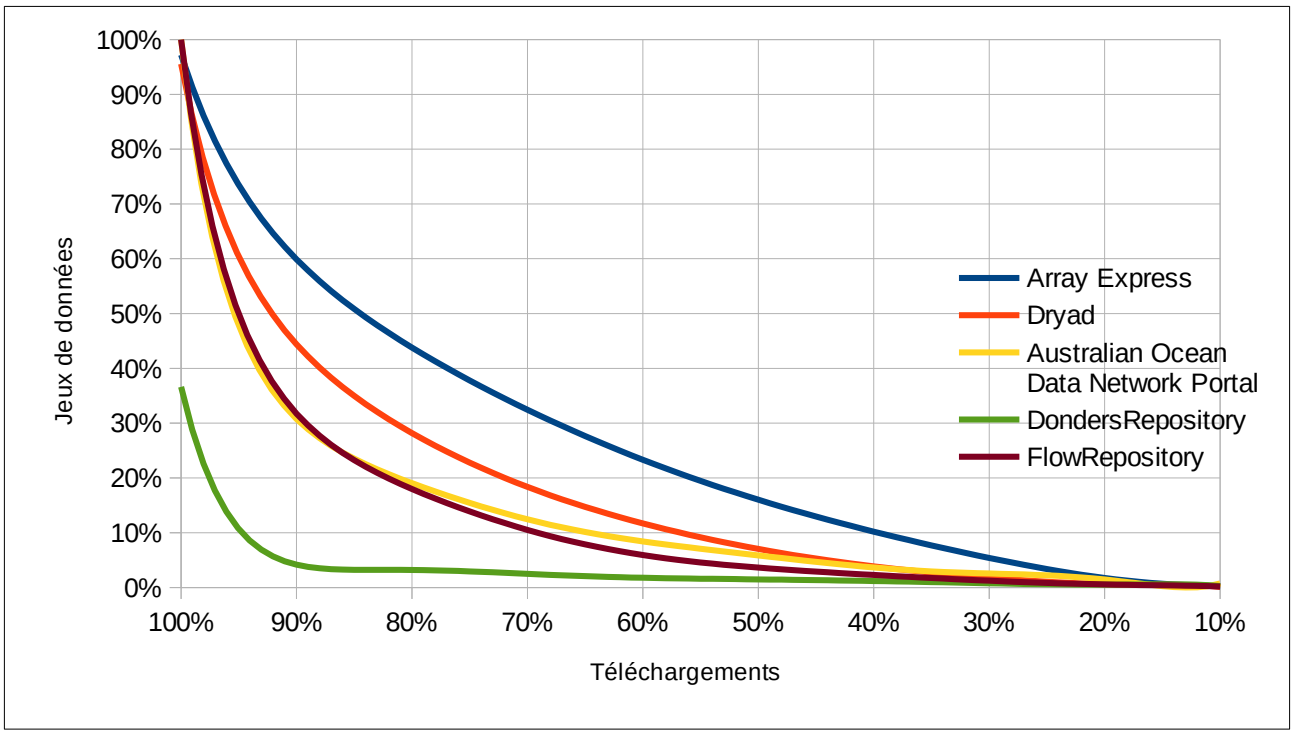

Figure 6 : Répartition des jeux de données en fonction du nombre de fois où ils ont été téléchargés (en %)

Si l'on regarde pour chacun des cinq entrepôts quel pourcentage de données reçoit 80 % des téléchargements, on obtient les résultats suivants :

- pour ArrayExpress, 44 % des données font l'objet de 80 % des téléchargements ;
- pour Dryad, 28 % ;
- pour Australian Ocean Data Network Portal, 19 % ;
- pour FlowRepository, 18 % ;
- pour Donders Repository, 3,2 %.

La règle du 80-20 s'applique donc particulièrement bien pour AODN Portal et FlowRepository. Au-delà de ce constat, on peut se demander si ce sont ces jeux de données, les plus téléchargés, qui font l'objet d'une réutilisation à des fins scientifiques. Peut-être s'agit-il au contraire des jeux de données qui sont téléchargés seulement trois ou dix fois ? Car l'une des limites des relevés de téléchargements est qu'ils ne permettent pas de savoir quelle utilisation est faite ensuite des données téléchargées.

Conclusion

Le présent article a eu pour but d'étudier les taux de consultation et de téléchargement des données disponibles dans 20 entrepôts et d'en analyser l'évolution et la répartition sur la période de 2015 à 2020. L'article s'est attaché, en introduction, à souligner la diversité des entrepôts en termes de périmètre, de fonctionnalités et de contenu. Ce préambule a permis de montrer qu'une étude comparative des données de consultation des entrepôts ne pouvait faire sens que de manière relative. Les résultats de l'étude que nous avons menée sont donc à considérer avec prudence. Ils nous ont permis d'esquisser des tendances générales.
La première tendance est l'augmentation générale du nombre de consultations, de téléchargements, de citations et de données disponibles dans les entrepôts. Des disparités ont néanmoins pu être observées. Parmi lesquelles :

- Une croissance du nombre de pages vues plus marquée dans les entrepôts postérieurs à 2010 que dans les entrepôts plus anciens ;

- Un nombre de téléchargements plus important dans les entrepôts qui ne sont pas seulement fournisseurs de données mais qui mettent également à disposition des données issues d'autres entrepôts.

La seconde tendance observée est la proportion relativement faible de données faisant l'objet du plus grand nombre de téléchargements. Nous avons vu que, de manière générale, les téléchargements se concentraient sur une petite part des données disponibles dans l'entrepôt, de l'ordre de 20 %.

Toutefois notre panel s'est révélé trop petit pour identifier plus avant d'éventuelles corrélations. A l'origine de cette étude, nous souhaitions en effet étudier l'influence potentielle du contexte politique et disciplinaire sur la consultation des données des entrepôts. Par exemple : les recommandations des éditeurs et des financeurs de la recherche en faveur de tel entrepôt ont-elles une influence positive sur le taux de consultation des jeux de données de celui-ci ? Ou encore : une communauté disciplinaire habituée à déposer des jeux de données dans un ou plusieurs entrepôts est-elle d'autant plus encline à utiliser ce ou ces entrepôts pour y rechercher des données ? Ces lacunes appellent donc à un approfondissement de la présente étude, sous un angle quantitatif comme qualitatif.

L'un des apports de cette étude est de nous avoir permis d'appréhender quelles données les administrateurs collectent pour mesurer l'utilisation de leur entrepôt. A savoir : la plupart de ceux que nous avons sondés comptabilisent le nombre de pages consultées, mais aussi le nombre de téléchargements et, pour certains d'entre eux, le nombre de citations dont font l'objet les jeux de données. Ces résultats concordent avec ceux qu'avaient obtenus Kratz et Strasser lors d'une enquête en 2015 (Kratz et Strasser, 2015). Si la nature des relevés ne semble donc pas avoir sensiblement évolué en quatre ans, un désir de changement semble néanmoins avoir émergé. Les échanges que nous avons eus avec les administrateurs des entrepôts témoignent en effet d'un souhait d'amélioration des mesures, afin de pouvoir mieux suivre la consultation et la réutilisation réelle des jeux de données. Deux participants à l'étude ont évoqué le projet Make Data Count[25], créé en 2017. Celui-ci vise à inciter les entrepôts et les éditeurs à adopter des pratiques normalisées pour la citation des données. Il est né du constat que : 1) il n'existe pas de normes communautaires pour les statistiques d'utilisation des données ; 2) il n'existe pas d'outil *open source* permettant aux entrepôts de données de collecter ces statistiques d'utilisation ; 3) il n'existe pas de lieu centralisé pour stocker, indexer et accéder à ces statistiques. La question de la réutilisation des données semble donc faire partie intégrante des préoccupations des entrepôts que nous avons étudiés (ne serait-ce que parce que ceux-ci ont mis en place un système de relevé des consultations, téléchargements et/ou citations). Cette réflexion suit néanmoins davantage une logique de démonstration de l'efficacité de l'*open science* qu'une logique d'incitation à l'utilisation des jeux de données en ligne. Autrement dit, elle s'adresse davantage aux déposants, qu'il convient de convaincre des bénéfices du dépôt de leurs données en ligne, qu'aux utilisateurs secondaires de ces données.

Or les enjeux qui se posent pour les entrepôts de données et le mouvement de l'*open science* ne doivent pas être confondus avec celui des sciences de l'information et de la communication, qui est de mener une réflexion sur les pratiques de réutilisation des données dans un contexte d'ouverture de la science. Au-delà de la définition d'indicateurs de téléchargements et de citations, l'enjeu est de mesurer quantitativement et qualitativement ces pratiques de réutilisation des données ouvertes.

Le présent article ouvre donc de nouvelles pistes de recherche. Il nous invite, premièrement, à nous interroger sur l'identité de ceux qui consultent et/ou téléchargent les données des entrepôts, ainsi que sur leurs intentions. Une évaluation qualitative de la réutilisation des données serait donc nécessaire. Une seconde piste de recherche consisterait à qualifier le petit nombre de jeux de données les plus téléchargés d'un entrepôt. Quelles sont ces données ? Comment sont-elles réutilisées ? Leur proportion tend-elle à augmenter au fil des ans ? Telles sont les questions auxquelles il serait intéressant de répondre.

25 https://makedatacount.org/

Remerciements

Nous tenons à remercier les administrateurs des entrepôts, qui ont accepté de participer au sondage et qui ont bien voulu prendre le temps de compiler les données dont nous avions besoin.

Bibliographie

André (F.), *Déluge des données de la recherche ?*, dans Calderan (L.), Laurent (P.), Lowinger (H.) et Millet (J.), *Big data : nouvelles partitions de l'information*, Louvain-la-Neuve, De Boeck, 2015, p. 77-95.

Assante (M.), Candela (L.), Castelli (D.) et Tani (A.), *Are Scientific Data Repositories Coping with Research Data Publishing?*, dans *Data Science Journal*, 15:6, 2016, p. 1-24. http://dx.doi.org/10.5334/dsj-2016-006

Borgman (C. L.), *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge (MA), The MIT Press, 2015, 416 p.

Chartron (G.), *L'Open science au prisme de la Commission européenne*, dans *Éducation et sociétés*, 41(1), 2018, p.177-193. https://www.cairn.info/revue-education-et-societes-2018-1-page-177.htm

Kratz (J. E.) et Strasser (C.), *Comment: Making data count*, dans *Scientific data*, 2, 150039, 2015. https://doi.org/10.1038/sdata.2015.39

Larivière (V.) et Sugimoto (C. R.), *Mesurer la science*, Montréal, *Les presses de l'Université de Montréal*, 2018, 176 p.

Leonelli (S.), *What Counts as Scientific Data? A Relational Framework*, dans *Philosophy of Science*, 82(5), 2015, p.810-821. https://www.jstor.org/stable/10.1086/684083

Missier (P.), *Data Trajectories: Tracking Reuse of Published Data for Transitive Credit Attribution*, dans *International Journal of Digital Curation*, 11(1), 2016. https://doi.org/10.2218/ijdc.v11i1.425

Mosconi (G.), Li (Q.), Randall (D.), Karasti (H.), Tolmie (P.), Barutzky (J.), Korn (M.), Pipek (V.), *Three Gaps in Opening Science*, dans *Computer Supported Cooperative Work*, 28, 2019, p.749-789. https://doi.org/10.1007/s10606-019-09354-z

Pampel (H.), Vierkant (P.), Scholze (F.), Bertelmann (R.), Kindling (M.) et al., *Making Research Data Repositories Visible: The re3data.org Registry*, dans *PloS ONE*, 8(11):e78080, 2013. https://doi.org/10.1371/journal.pone.0078080

Pasquetto (I.), *From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses.* PhD Thesis, University of California, 2018. https://escholarship.org/uc/item/1s1814cj

Piwowar (H. A.) et Vision (T. J.), *Data Reuse and the Open Data Citation Advantage*, dans *PeerJ*, 1:e175, 2013. https://doi.org/10.7717/peerj.175

Prost (H.) et Schöpfel (J.), *Les entrepôts de données en sciences de l'information et de la communication (SIC). Une étude empirique*, dans *Études de communication*, 52(1), 2019, p.71-98. https://www.cairn.info/revue-etudes-de-communication-2019-1-page-71.htm

Rebouillat (V.), *Ouverture des données de la recherche. De la vision politique aux pratiques des chercheurs*, Thèse de doctorat en sciences de l'information et de la communication, Cnam, Paris, 2019. https://tel.archives-ouvertes.fr/tel-02447653

Robinson-García (N.), Jiménez-Contreras (E.) et Torres-Salinas (D.), *Analyzing data citation practices using the data citation index*, dans *Journal of the Association for Information Science and Technology*, 67(12), 2016, p.2964-2975. https://doi.org/10.1002/asi.23529

Rücknagel (J.), Vierkant (P.), Ulrich (R.), Kloska (G.), Schnepf (E.), Fichtmüller (D.), Reuter (E.), Semrau (A.), Kindling (M.), Pampel (H.), Witt (M.), Fritze (F.), van de Sandt (S.), Klump (J.), Goebelbecker (H. J.), Skarupianski (M.), Bertelmann (R.), Schirmbacher (P.), Scholze (F.), Kramer (C.), Fuchs (C.), Spier (S.) et Kirchhoff (A.) (2015), *Metadata Schema for the Description of Research Data Repositories*, sur le site *re3data.org*. http://doi.org/10.2312/re3.008

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