UNIVERSITÉ DE LILLE

FACULTE DE MÉDECINE HENRI WAREMBOURG

Année : 2020

THÈSE POUR LE DIPLOME D'ÉTAT

DE DOCTEUR EN MÉDECINE

Valorisation des jeux de données issus de la recherche médicale : la publication de *data papers*, analyse bibliométrique et scores *altmetrics*. Application aux données de la neuroscience.

Présentée et soutenue publiquement le 9 octobre 2020 à 18h
au Pôle Recherche

par Audrey PARENT

JURY

Président :
Monsieur le Professeur Emmanuel CHAZARD
Assesseurs :
Monsieur le Professeur Stefan DARMONI
Monsieur le Professeur Cristian PREDA
Monsieur le Docteur Vincent CHOURAKI
Directeur de thèse :
Madame le Docteur Amélie LANSIAUX

Avertissement

« La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs. »

Table des matières

Résumé

Contexte

Les données scientifiques sont devenues tout aussi importantes que le résultat de leur analyse. La réutilisation efficace des données rend la question de leur partage centrale mais leur publication est limitée par manque de reconnaissance spécifique. Le *data paper* semble pouvoir répondre en partie à ce problème. L'objectif de notre travail était d'étudier la publication des *data papers* dans les sciences de la santé et plus particulièrement en neuroscience.

Méthodes

Nous avons mené une étude empirique au sein des revues reconnues comme publiant des *data papers* et sélectionné les *data papers* du domaine médical publiés jusqu'au 31 octobre 2018. Les articles ont été classés selon leur spécialité médicale et leur thématique. Pour les *data papers* de neuroscience, des informations générales et spécifiques au contenu de l'article ont été relevées, et l'accès aux données décrites testé. Nous avons recherché leur typologie d'indexation au sein des trois grandes bases bibliographiques (Pubmed, Web Of Science, Scopus). Enfin, les données de bibliométrie et les scores *altmetrics* disponibles sur la base Scopus ont été recueillis.

Résultats

745 *data papers* médicaux ont été identifiés dont 163 en neuroscience. Leur publication a fortement augmenté ces dernières années mais reste hétérogène. 68,9% des données médicales décrites sont directement accessibles, essentiellement des données d'électrophysiologie et d'imagerie. Le nombre de citations annuelles

augmente. 75% des *data papers* ont été cités pour une moyenne de 2,7 citations/an/article et un délai médian de citation à 7,9 mois (IC95% 6,7 – 10,3). Une activité *altmetric* a été retrouvée pour 53% des articles sur Twitter et pour 86% sur les sites de *bookmarking* sociaux. Des différences existent selon la revue.

Conclusion

Le *data paper* reste pour le moment un phénomène récent, à petite échelle et peu homogène. En attendant une reconnaissance spécifique des citations de données, le *data paper* semble être un moyen alternatif acceptable dont le nombre continuera probablement à croître dans les années à venir.

Liste des abréviations

BMJ	British Medical Journal
CIM-10	Classification Internationale des Maladies 10ème version
CWTS	Centre for Science and Technology Studies – Leiden University
DOI	Digital Object Identifier
DS	Déviation standard
EEG	Electroencéphalographie
EMG	Electromyographie
EOG	Electro-oculographie
FAIR	Findable Accessible Interoperable Reusable
GEO	Gene Expression Omnibus
HUVEC	Human Umbilical Vein Endothelial Cell
IC 95%	Intervalle de Confiance à 95%
ICMJE	International Committee of Medical Journal Editors
IMRAD	Introduction Méthodes Résultats Discussion
INDI	International Neuroimaging Data-sharing Initiative
IRM	Imagerie par Résonance Magnétique
JAMA	Journal of the American Medical Association
MDPI	Multidisciplinary Digital Publishing Institute
MeSH	Medical Subject Headings
NEJM	The New England Journal of Medicine
NIH	National Institute of Health
NITRC	Neuroimaging Informatics Tools and Resources Clearinghouse

NSF	National Science Foundation
OCDE	Organisation de Coopération et de Développement Economique
ODE	Opportunities for Data Exchange
PDB	Protein Data Bank
PDF	Portable Document Format
PRIDE	PRoteomics IDEndifications Database
PLOS	Public Library Of Science
RIN	Research Information Network
TEP	Tomographie par Emission de Positons
VIH	Virus de l'Immunodéficience Humaine
WoS	Web of Science

Liste des figures

Liste des tableaux

Introduction

Les données scientifiques sont devenues aujourd'hui toutes aussi importantes que le résultat de leur analyse. Avec la possibilité de générer et d'analyser de grandes quantités de données, la pratique scientifique a évolué vers une science axée sur les données. La réutilisation efficace des données fait partie des attentes de cette nouvelle approche et la question du partage des données est donc centrale. La mise à disposition de ces données nécessite un investissement supplémentaire pour lequel il n'existe pas encore de reconnaissance spécifique. Le *data paper*, nouvelle forme de communication scientifique, semble pouvoir répondre en partie à ce problème et pourrait faire évoluer le paysage des publications scientifiques dans les années à venir.

La première partie de cette introduction rappelle les principales approches scientifiques afin de comprendre en quoi l'avènement des *big data* remet en question la manière « de faire » de la recherche. La problématique générale de notre travail porte sur le partage et la publication des données, éléments clés de cette nouvelle science, qui sera développée dans un second temps. Puis, nous nous intéresserons plus spécifiquement à la publication des *data papers*, nouvelle forme de communication dont le but est d'inciter au partage de données tout en étant valorisée au même titre qu'un article scientifique classique. Enfin, les objectifs de notre travail portant sur la publication des *data papers* dans les sciences de la santé et plus particulièrement en neuroscience seront présentés.

1 Les approches méthodologiques en recherche

Les chercheurs considèrent les données comme des objets de recherche de premier ordre constituant la base de leurs travaux scientifiques. Elles nécessitent un certain nombre de ressources tant matérielles, humaines et financières afin de les acquérir, de les traiter, de les stocker, de les analyser, de les interpréter et de les valoriser. Les avancées technologiques et numériques de ces dernières années permettent d'obtenir et de gérer un nombre de plus en plus important de données, on parle dans certains cas de *big data*, avec des formats de plus en plus diversifiés, dans des domaines scientifiques variés. Cette évolution transforme le monde de la recherche et relance le débat sur la place des différentes approches méthodologiques, en remettant notamment en question le rôle des tests d'hypothèses dans la pratique scientifique (1).

L'approche empirique et l'approche rationnelle sont les deux principales approches existantes. Leur différence réside notamment sur le sens de leur raisonnement. Pour comprendre, il faut distinguer d'un côté le monde des connaissances et des hypothèses (i.e « Idées »), et de l'autre, celui de la réalité que l'on peut capter et mesurer par nos sens ou à l'aide d'instruments donnant lieu à des observations (i.e « Données »). Le raisonnement de l'un est de passer des « Idées » aux « Données » et l'autre des « Données » aux « Idées ». Le point de départ et la construction du raisonnement qui s'en suit sont différents. La difficulté est qu'on ne peut pas considérer que l'un est simplement le contraire de l'autre (2). Ceci a donné lieu à de nombreux débats sur l'approche la plus adaptée pour montrer la vérité du monde qui nous entoure.

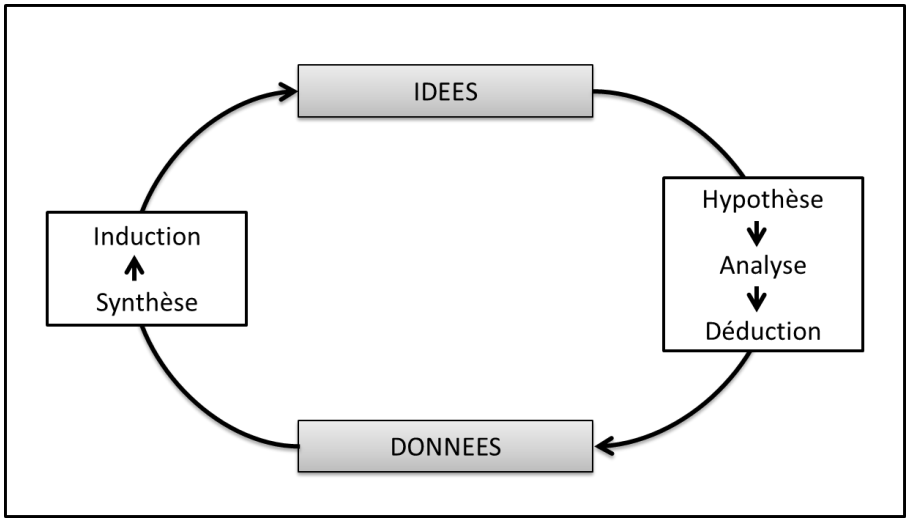


Figure 1 : Le cycle des connaissances, adapté de (2)

Le progrès scientifique peut être considéré comme un cycle itératif alliant connaissances et observations utilisant les modes de raisonnement déductif et inductif (Figure 1). La recherche de demain, à l'ère des *big data*, devrait développer cette mixité des approches autour d'un processus itératif et cesser de les opposer, l'un pouvant apporter à l'autre (1–4) ; l'objectif final étant en tous temps de combler les lacunes de la science.

1.1 L'approche empirique, le raisonnement inductif

L'empirisme est une pensée philosophique qui soutient que l'expérience et l'observation sont à l'origine de toute la connaissance humaine (5). Ainsi, nos sens sont à la source de nos connaissances. De par l'accumulation d'observations et de faits mesurables on peut en extraire des informations intéressantes afin d'établir des lois générales. L'opération mentale consistant à remonter du cas particulier au général, du concret à l'abstrait, de l'expérience à la théorie constitue le raisonnement inductif ou induction (2,6).

L'induction cherche à établir des lois générales à partir d'un ensemble de données particulier ou d'une série d'observations en générant des modèles dans le but de dire que le même résultat se produira à l'avenir (6). Toutefois, les lois ainsi établies sont sans certitude, car elles peuvent être à tout moment démenties par un contre-exemple (Exemple 1 et 2). L'induction s'oppose à la déduction qui, si elle est formulée correctement et avec des prémisses vraies, aboutit à une conclusion toujours vraie (Exemple 2).

Exemple 1 issue de (6) : Tous les corbeaux que je vois sont noirs, je n'ai jamais rencontré de corbeaux d'une autre couleur. J'en « induis » la loi générale que tous les corbeaux sont noirs. Mais il ne s'agit que d'une quasi-certitude car le premier contre-exemple (voir par exemple un corbeau blanc), mettra en cause la loi précédemment établie qui s'avèrera fausse.

Exemple 2 issue de (2) : L'herbe est mouillée (observation) j'en induis qu'il a plu (théorie). Cependant on ne peut pas affirmer avec certitude qu'il ait plu car il est tout à fait possible que le mouillage ait été fait avec un tuyau d'arrosage. Au contraire, dans un raisonnement déductif on sait que la prémisse « la pluie mouille » est vrai donc je peux en déduire avec certitude que s'il pleut (théorie) alors l'herbe sera mouillée (observation).

La notion d'induction s'appuie donc sur le fait que l'accumulation d'événements concordants, sans contre-exemple, augmente la probabilité de les voir se renouveler et la loi ainsi induite devient une quasi-certitude (6).

Bien qu'elle ne soit pas considérée comme un raisonnement fiable de nos jours, l'induction est cependant universellement utilisée dans toutes les sciences et prédominait entre le XVIIème et XIXème siècle (Figure 2) (1,7). Pour faire face aux figures de l'époque, telles que Robert Boyle et Robert Hooke, préconisant l'utilisation d'hypothèses, Francis Bacon et Isaac Newton, partisans de l'approche inductive, ont fait valoir que les scientifiques pourraient aboutir à des conclusions faussées si les conjectures et les hypothèses utilisées étaient initialement non prouvées (7). Newton écrit en 1729 que les inductions « doivent être considérées exactement ou presque comme vraies jusqu'à ce que de nouveaux phénomènes puissent les rendre plus exactes ou susceptible d'exception » (7). Ainsi, la plupart des scientifiques ne sont pas favorables à l'utilisation d'hypothèse au XVIIIème siècle mais cette vision de la science change complètement au cours du XXème siècle avec notamment la pensée de Popper sur la méthode hypothético-déductive (1,7).

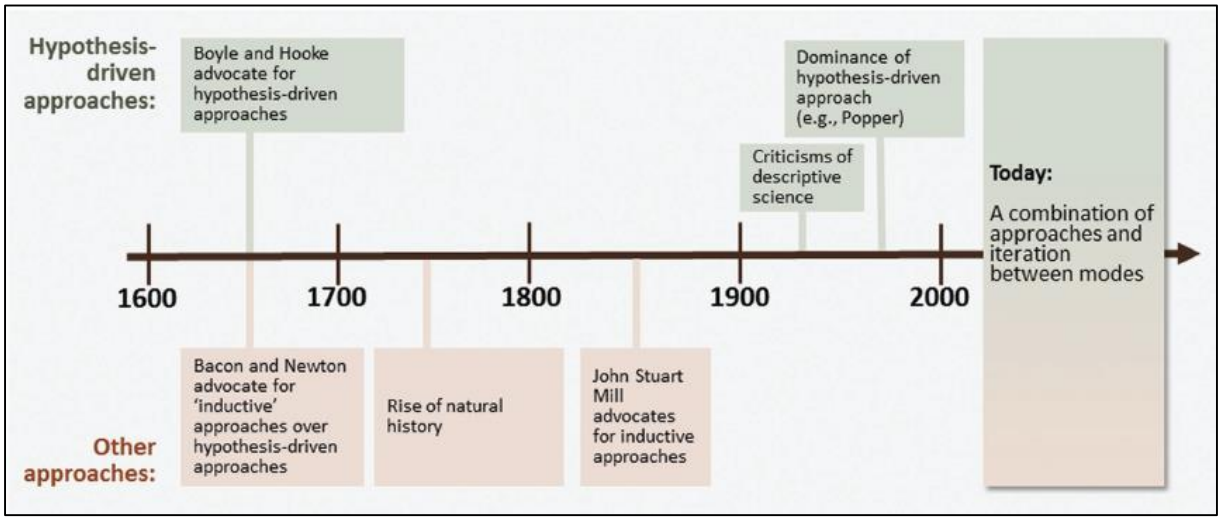


Figure 2 : Evolution de la dominance des approches scientifiques (1)

Aujourd'hui, l'approche inductive est considérée comme une approche exploratoire, pouvant être utile dans la mesure où elle alimente la recherche fondée sur les hypothèses (travaux préliminaires) sans toutefois être considérée comme participant au progrès de la science, la meilleure science reposant sur les hypothèses dont l'objectif est de les tester (8). Certains scientifiques et philosophes considèrent même cette approche de « non scientifique » en la qualifiant « d'expédition de pêche » alors qu'une partie des grandes découvertes scientifiques découlent d'observation, parfois fortuite, sans hypothèse ni paradigme initial (2,9). Plus récemment, Water a suggéré que l'objectif général de la recherche exploratoire « est de générer des découvertes significatives sur des phénomènes sans faire appel à une théorie sur ces phénomènes » (10). Ainsi, les chercheurs ne sont pas guidés directement par la théorie scientifique existante, mais plutôt par l'espoir que tout ce qu'ils découvriront sera pertinent pour des projets en cours ou futurs (8).

1.2 L'approche rationnelle, le raisonnement déductif

Le rationalisme est une pensée philosophique selon laquelle la raison est la seule source de connaissances (11). Tout ce qui existe a une explication rationnelle et peut être décrit par la raison humaine à l'aide d'un raisonnement. Ainsi, toute connaissance certaine découle de principes *a priori* et universels. Le procédé logique par lequel on va du général au particulier, du principe aux conséquences, de la cause aux effets constitue le raisonnement déductif ou déduction (12).

La déduction permet d'exploiter un modèle, une loi, une théorie pour expliquer un fait. Cette méthode de raisonnement utilise des prémisses établies non pas par une observation directe des faits mais par référence à des prémisses déjà établies et sa validité dépend du respect de sa formulation (12). Il est possible d'aboutir à des

conclusions fausses si l'une des prémisses est fausse (Exemple 1) mais si les prémisses sont vraies avec une formulation correcte alors la conclusion est toujours vraie (Exemple 2).

Exemple 1 issue de (12) : Soient les deux prémisses suivantes : les chats sont des oiseaux, les oiseaux pondent des œufs. On peut en déduire que les chats pondent des œufs. Bien que le raisonnement soit valide, l'une des deux prémisses étant fausse, la conclusion est fausse.

Exemple 2 issue de (12) : Soient les deux prémisses suivantes : les mammifères sont des animaux, les chats sont des mammifères. On peut en déduire que les chats sont des animaux. Les prémisses étant vraies et le raisonnement valide, la conclusion est donc correcte.

La méthode hypothético-déductive développée par le philosophe autrichien Karl Popper, pour faire face au « problème » du raisonnement inductif, a été extrêmement influente au cours du XXème siècle et reste encore aujourd'hui l'approche traditionnelle adoptée par la plupart des disciplines scientifiques. A la place de chercher à vérifier qu'un concept, qu'une règle, qu'une hypothèse, issu pour la plupart d'un raisonnement inductif, est vrai, par un nombre d'expériences qui sera toujours limité, Popper suggère une méthode où les concepts, les règles, les hypothèses sont soumises à la falsification. Autrement dit, Popper suggère de se concentrer à prouver que l'idée de départ est fausse. Ainsi, un seul élément de preuve à l'encontre de l'hypothèse initiale serait suffisant pour affirmer que cette hypothèse est fausse. Le

corollaire étant qu'aucun élément de preuve à l'appui de l'hypothèse ou non falsifiant serait suffisant pour affirmer que cette hypothèse soit vraie (7).

Bien que les définitions d'hypothèses varient en fonction des disciplines (13), le raisonnement hypothético-déductif permet de tester ces hypothèses en confrontant les résultats attendus aux résultats de l'expérience ou de l'observation selon un processus linéaire (Figure 3) :

- Une hypothèse est définie résultant la plupart des cas d'une induction suite à une série d'observations ou un ensemble de données ayant entrainé une série de questions du fait d'un manque de connaissance, d'un état d'ignorance
- Les résultats attendus (prédictions) sont déduits de cette hypothèse selon un raisonnement déductif du type « si H est vraie alors on doit obtenir les résultats suivants ou observer les faits suivants … »
- Une expérience est réalisée et les données issues de cette expérience sont analysées afin de tester leur compatibilité avec l'hypothèse
- Si les résultats obtenus ne sont pas conformes aux résultats attendus, alors l'implication est fausse et l'hypothèse est rejetée. Si les résultats obtenus sont conformes aux résultats attendus alors l'implication est juste et l'hypothèse peut être conservée mais elle n'est pas vérifiée

Ce processus scientifique continue aujourd'hui à être considéré comme la méthode de référence pour laquelle il est plus aisé d'attirer des fonds de recherche de la part des agences de financement ; les données produites par cette méthode étant considérées comme plus fiables et plus facilement évaluables par rapport aux données issues de recherches empiriques (8,14).

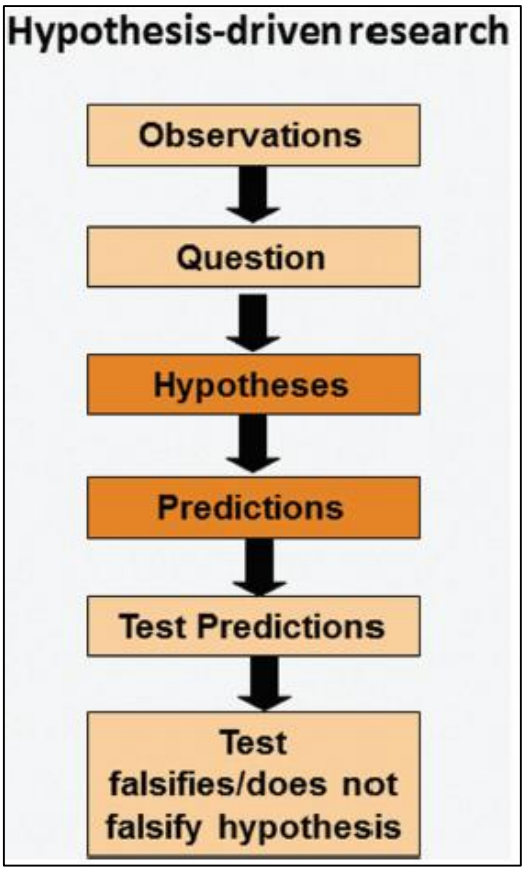


Figure 3 : Processus linéaire de la méthode scientifique (1)

Que ce soit pour élaborer des hypothèses ou pour tester la validité de ces dernières, les scientifiques utilisent des ensembles de données. Toutefois, ces ensembles de données sont souvent limités en raison de différentes contraintes notamment de temps et de coût. L'avènement de la science à forte intensité de données semble pouvoir répondre à ces problématiques en permettant l'acquisition à moindre coût d'importantes quantités de données sur des populations entières au lieu d'échantillons et permet l'exploration de nombreuses corrélations dont certaines non envisagées initialement par les scientifiques.

1.3 L’approche axée sur les données, vers un raisonnement mixte

La possibilité de collecter et d’analyser d’énormes quantités de données, pouvant être hétérogènes et provenir de différentes sources, révolutionne la manière dont la recherche scientifique est menée. Au lieu de tester des hypothèses, d’examiner des modèles hypothétiques attendus en analysant des données pertinentes provenant d’ensembles de données limités par leur portée, leur temporalité et leur taille, l’approche « axée sur les données » cherche à extraire l’information qui se dégage des jeux de données énormes, dynamiques et variés autrement dit cherche à obtenir un aperçu « né des données » (15). De plus en plus de domaines scientifiques utilisent des approches à forte intensité de données, notamment l’astronomie, la physique des hautes énergies ou encore la génomique (16–18). Les *big data* comprennent à la fois les données traditionnelles qui sont de plus en plus enregistrées et saisies numériquement, ainsi que de nouvelles données acquises par des méthodes autonomes ou manuelles (19,20). On peut attribuer l’émergence des *big data* à trois événements majeurs (21) :

- Une révolution technologique pour la génération de données
- Un développement d’outils adaptés pour l’analyse de données
- Un changement de concept dans la pratique scientifique allant vers des données ouvertes

L’accès aux *big data* et aux nouvelles pratiques de recherche conduit certains à définir l’émergence d’un nouveau paradigme de recherche qui remet en cause l’approche déductive scientifique établie (15,22). Jim Gray décrit l’évolution de la science selon quatre paradigmes (Figure 4). Il décrit le quatrième paradigme, la science à base de données, comme une méthode collaborative, en réseau, pilotée par

les données et définit la « e-Science » comme étant la synthèse des technologies de l'information et de la science permettant de relever des défis à des échelles auparavant inimaginables (16,23). Ainsi, cette approche offre plusieurs opportunités (19) :

- Générer de nouvelles connaissances plus rapidement que les approches scientifiques traditionnelles
- Collecter et analyser les données de façon non biaisée par les connaissances antérieures
- Permettre une compréhension holistique des systèmes et plus particulièrement des systèmes complexes

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalization	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

Figure 4 : Les quatre paradigmes de la science (15)

L'avènement des *big data* offre la possibilité d'une nouvelle approche de la science. Toutefois, la forme de cette approche ne semble pas encore définie. Le terme « science à forte intensité de données » est assez flou comme le souligne Sabina Leonelli, « il est difficile de caractériser de manière générale les méthodes basées sur les données, étant donné le large éventail d'activités et d'objectifs épistémiques actuellement inclus dans cette rubrique » (24,25). On retrouve néanmoins deux caractéristiques principales « l'une est l'intuition que l'induction à partir des données existantes est reconnue comme une forme importante d'inférence scientifique, qui peut guider et éclairer la recherche expérimentale ; et l'autre est le rôle central des machines, et donc du raisonnement automatisé, dans l'extraction de modèles

significatifs à partir des données » (24,25). Deux voies sont actuellement proposées pour caractériser cette science à forte intensité de données mais divergent dans l'approche méthodologique :

- Une nouvelle forme d'empirisme, dans lequel les données peuvent être utilisées, parler pour elles-mêmes sans théorie
- Une science axée sur les données qui modifie la méthode scientifique existante en mélangeant les approches d'induction et de déduction

Les partisans de la science à forte intensité de données affirment que la place de la théorie change. L'ancien rédacteur en chef du magazine *Wired*, Chris Anderson, proclame la « fin de la théorie » en affirmant que « le déluge des données rend la méthode scientifique obsolète », que « la corrélation remplace la causalité » (26). D'autres auteurs (15,27–29) rallient cette vision radicale prônant le renouveau d'une certaine forme d'empirisme dont l'objectif n'est probablement ni de comprendre le monde ni de contribuer à la pérennité des connaissances scientifiques comme le stipule Siegel « Nous ignorons généralement tout lien de causalité et nous ne nous en soucions pas nécessairement…L'objectif est davantage de prédire que de comprendre le monde…la prédiction l'emporte sur l'explication » (15,29). La forte influence de cette approche est liée notamment à quatre idées allant à l'encontre de l'approche déductive reconnue (15) :

- Les données volumineuses peuvent capturer un domaine entier et fournir une résolution complète
- Il n'y pas besoin de théorie *a priori*, de modèle ou d'hypothèse
- Grâce à l'application de l'analyse de données agnostique, les données peuvent parler d'elles-mêmes, sans parti pris ni cadrage humain, et tous les

modèles et relations au sein du *Big Data* sont intrinsèquement significatifs et véridiques

- Le sens transcende les connaissances contextuelles ou spécifiques à un domaine et peut donc être interprété par quiconque est capable de décoder une statistique ou une visualisation de données

Les exemples classiques utilisés pour illustrer cette approche proviennent principalement des secteurs bancaires, assurantiels ou encore de la vente où les quantités importantes de données clients recueillies en routine à des fins transactionnelles ou d'archivage sont analysées secondairement, via des techniques d'extraction de données automatisées, afin d'identifier des corrélations pouvant être intéressantes. Dyche, par exemple, décrit l'analyse des transactions d'achat d'une chaine commerciale où la découverte de corrélation entre certains articles a abouti à la réorganisation des rayons entrainant alors une augmentation de 16% du revenu du panier client (28). De la même manière, le système de suggestion des sites de ventes en ligne se base sur les habitudes d'achat et de consultation de pages du consommateur mais également sur celle des autres afin de proposer des articles supplémentaires susceptibles d'intéresser ce dernier lors de la finalisation de sa commande. Dans les deux cas, il n'y a aucune hypothèse selon laquelle le produit A est plus souvent acheté avec le produit Z, les données sont simplement interrogées pour découvrir quelles relations existent.

Bien que les quatre idées sur lesquelles se basent les adeptes du renouveau de l'empirisme semblent relativement puissantes et attrayantes, Kitchin montre en quoi elles sont fallacieuses (15). En effet, il existe des expériences où l'analyse de ces grandes quantités de données s'est révélée être un échec, c'est le cas notamment du projet *Google Flu Trends*. Ce projet devait permettre de détecter les épidémies de

grippe et leur évolution plus rapidement que les services d'alerte habituels en analysant les recherches faites sur Internet par les internautes (par exemple « symptôme de la grippe », « maux de tête », « fièvre ») et leur localisation. Toutefois, *Google Flu Trends* a nettement sous-estimé l'épidémie de la grippe aviaire H1N1 en 2009 et au contraire largement surestimé la grippe saisonnière de 2012-2013 aux Etats-Unis en prédisant 50% de cas de plus que le nombre de cas réellement constaté (30–32). Il est probable qu'il s'agisse de ce type de résultats qui rende sceptique certains scientifiques quant à l'adoption de cette forme d'approche la considérant comme une approche exploratoire n'entrant pas dans le cadre d'une démarche scientifique et pouvant aboutir à des résultats faussés.

L'exploration de données dans les bases de données aussi grandes soient elles, ne doit pas nécessairement être considérée comme une façon de contourner l'analyse traditionnelle des données fondée sur des hypothèses mais plutôt comme une « valeur ajoutée » (2,33). Contrairement aux nouvelles formes d'empirisme, la science axée sur les données cherche à respecter les principes de la méthode scientifique tout en la rendant plus ouverte à l'utilisation conjointe d'approche inductive et déductive (1,2,4,15,19). La science axée sur les données cherche à générer des hypothèses et des idées « nées des données » plutôt que « nées de la théorie », autrement dit, elle supprime l'hypothèse selon laquelle le scientifique dispose d'une connaissance adéquate pour générer les meilleurs hypothèses spécifiques et vérifiables mais suppose que la compréhension des scientifiques sur les systèmes, notamment complexes, est rudimentaire et utilise donc les données pour générer des hypothèses spécifiques (15,25,34).

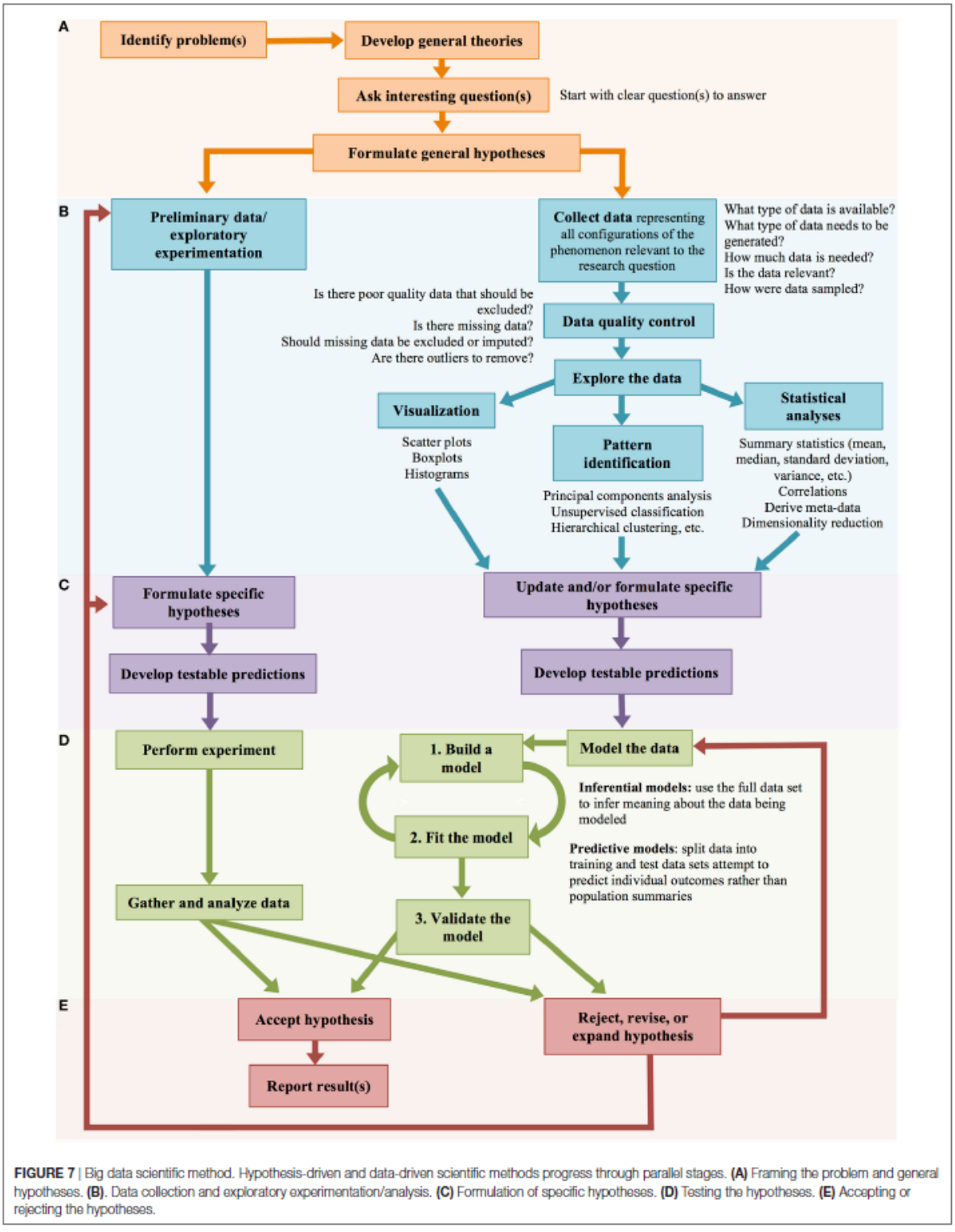


FIGURE 7 | Big data scientific method. Hypothesis-driven and data-driven scientific methods progress through parallel stages. **(A)** Framing the problem and general hypotheses. **(B)**. Data collection and exploratory experimentation/analysis. **(C)** Formulation of specific hypotheses. **(D)** Testing the hypotheses. **(E)** Accepting or rejecting the hypotheses.

Figure 5 : Démarche scientifique, parallélisme des approches (19)

O'Malley a fait valoir que la recherche axée sur les données devrait être caractérisée comme une interaction itérative entre plusieurs modes de recherche différents : la recherche basée sur les hypothèses, la recherche exploratoire incluant l'expérimentation, l'exploration de données, la modélisation ou encore la simulation, et la recherche de nouveau outils, technologies et méthodes (1,4). Il s'agit donc d'une approche ouverte, non linéaire, faisant appel à une combinaison d'approche différentes où les questions initialement posées sont précisées, revues ou donnent lieu à de nouvelles pistes de recherche au fur et à mesure de l'avancée du projet (1–3,19,34,35). Les approches inductives jouent alors un rôle important dans la génération d'hypothèses mais sans que l'explication par l'induction soit le but recherché comme dans les approches empiriques (15). En effet, il s'agit d'identifier via ces approches de nouvelles hypothèses pertinentes qui seront à même d'être étudiées de façon plus approfondie et non d'identifier toutes les relations existantes au sein d'un jeu de données et de supposer qu'elles ont un sens (15,36). Par ailleurs, la manière dont les données sont générées, réutilisées ou analysées est soigneusement pensée, régie par des hypothèses et étayée par des connaissances théoriques et pratiques quant aux techniques susceptibles de produire un matériel de recherche approprié et utile (15).

Bien que différent dans certains aspects, le processus reste partiellement similaire (Figure 5). Les principales étapes, décrites ci-dessous, peuvent pour certaines nécessiter une itération (19,37) :

- Identifier la ou les questions de recherche sur la base d'un manque de connaissance dans un domaine d'intérêt et formuler des hypothèses générales

- Générer (par expérimentation) ou collecter (données d'observation) des données représentant toutes les configurations possibles pouvant présenter un intérêt pour le problème de recherche
- Nettoyer, extraire, agréger, transformer les données acquises dans le but d'explorer et de détecter des informations potentiellement intéressantes
- Développer des hypothèses plus spécifiques concernant les relations entre les variables y compris des hypothèses concernant les relations de cause à effet
- Tester les hypothèses par modélisation des données. En cas d'hypothèse rejetée, le scientifique retourne le plus souvent aux mêmes données pour répéter la modélisation afin de tester une nouvelle hypothèse

Ainsi, on peut considérer la science axée sur les données comme une version reconfigurée de la méthode scientifique traditionnelle, offrant un nouveau moyen de construire la théorie (15).

Pour une meilleure efficience de la recherche axée sur les données, la création d'équipes de recherche interdisciplinaires devient un élément important afin de mener à bien les tâches itératives requises par les projets de recherche. Avant l'ère des *big data*, les chercheurs experts du domaine identifiaient les données dont ils avaient besoin pour effectuer leurs tests, les collectaient et les analysaient. Aujourd'hui, comme l'indique Strasser « l'analyse des données est effectuée par des chercheurs dont les antécédents disciplinaires sont différents de ceux qui les produisent […] » (24,38). La réutilisation des données de façon optimale nécessite des compétences du domaine théorique mais également des compétences techniques d'autant plus qu'il existe différentes formes d'approches et de méthodes. Il est donc difficile pour un

scientifique de devenir expert dans tous les domaines (1,24,38). Ainsi, la collaboration multidisciplinaire et le partage des données qui en découle apparaissent comme des éléments clés de cette nouvelle approche scientifique afin d'avoir une exploitation optimale des ensembles de données disponibles.

Pour finir, *The National Consortium for Data Science* (39) définit la science des données comme « l'étude systématique de données numériques utilisant des techniques scientifiques d'observation, le développement de théories, l'analyse systématique, la vérification d'hypothèses et une validation rigoureuse » (40). L'un des principaux objectifs de la science des données est d'utiliser les données pour décrire, expliquer et prévoir les phénomènes naturels et sociaux en :

- Créant des connaissances sur les propriétés de grands ensembles de données dynamiques
- Développant des méthodes pour partager, gérer et analyser les données numériques
- Optimisant les processus de données tels que la précision, la latence et le coût

La science des données vise à permettre aux scientifiques d'analyser efficacement des système plus grands et plus complexes, en complétant les processus traditionnels de génération d'hypothèses et de test d'expérimentation, dans le but d'affiner notre compréhension du monde.

2 Le partage des données de la recherche

L'exploitation et la réutilisation efficace des données fait partie des attentes de la science axée sur les données (*data driven research)*. Cette nouvelle approche scientifique met les données de la recherche au centre des attentions les considérant comme des objets de recherche de premier ordre (des « produits » de la recherche) dont la valeur peut être considérée comme tout aussi importante que celle des résultats qui en découlent. Toutefois, cela nécessite la mise en place de mécanismes de gestion, de conservation, d'accès et de partage des données de la recherche. Le partage des données de la recherche entre dans le champ de *l'Open Data*, et plus largement de *l'Open Science*, pour lequel les autorités réglementaires, les organismes de financements, les institutions et les éditeurs adoptent de nouvelles lignes directrices afin que ces données puissent être disponibles, trouvables, accessibles, interprétables, réutilisables et citables.

2.1 Données de la recherche

Définir ce qu'est une donnée n'est pas chose aisé (41,42). La définition la plus souvent retrouvée pour définir les données de la recherche est celle de l'OCDE dans leur rapport de 2007 (43) : « les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche ». Bien qu'usuelle, cette définition apparait restrictive car en pratique les données de la recherche ne se limitent pas aux « enregistrements factuels » et certains éléments produits par les chercheurs comme par exemple les carnets de laboratoires ou des analyses préliminaires sont exclus de cette définition (44,45). Par

ailleurs, Fournier (46) stipule que « les données peuvent être produites en dehors du processus de recherche : elles deviennent des données de la recherche dès qu'un chercheur les utilise et, en les utilisant, se les approprie ». Enfin, Rosemberg (44) explique en quoi les données de la recherche ne sont pas seulement des données « nécessaires à la validation des résultats » : « un chercheur produit beaucoup plus de données que celles qui sont *stricto sensu* nécessaires pour valider les résultats de la recherche. Des données qui pourraient être utilisées par d'autres chercheurs dans le cadre d'un projet de recherche inédit ». Ainsi d'autres définitions ont été proposées (47), principalement par des chercheurs d'universités britanniques ou australiennes, dont celle de la *Quennsland University of Technology* repris par Rémi Gaillard (42) « Il s'agit donc de l'enregistrement de « faits donnés », sous forme numérique, descriptive ou visuelle et sur lequel un argument, une théorie, une hypothèse ou tout autre produit de la recherche est basé. Ces données peuvent être brutes, nettoyées ou traitées, et peuvent être enregistrées sous tout format et tout support ».

Au vu de ces définitions, la typologie des données peut être large et réaliser une classification en prenant en compte toutes les dimensions d'une donnée est difficile. Généralement, les données sont classées en 5 catégories, adaptées de la classification de la *Research Information Network (RIN)*, qui dépendent de la manière dont elles sont produites (méthodes et outils utilisés) et de leur valeur supposée (42,45,48) :

- **Données d'observation** : données collectées en temps réel, indissociables d'un contexte donné. Habituellement unique et impossible à reproduire

- **Données expérimentales** : données obtenues à partir d'équipement en laboratoire, suivant une méthodologie définie. Souvent reproductible mais parfois couteuse
- **Données computationnelles ou de simulation** : données issues de simulation à partir de modèles informatiques. Souvent reproductible si le modèle est bien documenté
- **Données dérivées** : données issues du traitement, de la combinaison ou de la réorganisation de données « brutes ». Souvent reproductibles mais coûteuses
- **Données de référence** : collection ou accumulation de petits jeux de données qui ont été revus par les pairs, annotés et mis à disposition

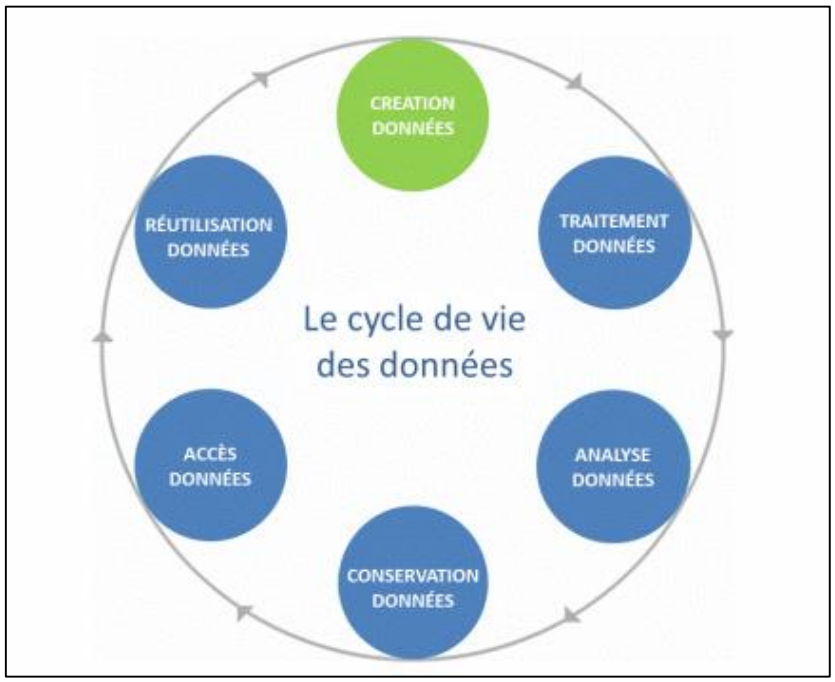


Figure 6 : Cycle de vie des données (50)

Par ailleurs, les données ne sont pas figées et leurs caractéristiques évoluent en fonction des étapes du projet de recherche, autrement dit en fonction de leur cycle de vie (49,50) (Figure 6). On distingue ainsi (45,51,52) :

- **Les données brutes** ou données sources ou données primaires : données recueillies et utilisées par les chercheurs pour leurs recherches mais qui n'ont pas encore été organisées, mises en forme ou analysées
- **Les données traitées** ou dérivées : données produites après calibration/étalonnage ou correction des données brutes (données traitées) ; données correspondant à un résumé ou à une représentation spécifique des données (données dérivées)
- **Les données analysés** ou données résultats : données produites comme résultats de recherche
- **Les données conservées** : données enregistrées et archivées en vue de les garder dans le temps de manière plus ou moins pérenne
- **Les données publiées** : données partagées avec la communauté, généralement revues par les pairs
- **Les données réutilisées** ou données secondaires : données existantes pouvant être exploitées dans un objectif différent de leur collecte initiale

Comme le mentionne Rémi Gaillard (42), « les données n'ont donc pas la même « valeur » : selon les disciplines, selon la manière dont a été planifiée la collecte, selon les objectifs pour lesquels elles ont été produites, l'intérêt que peut présenter leur conservation et leur diffusion varie ». Ainsi, toutes les données, brutes ou dérivées, peuvent être concernées par l'*open data*, il revient donc au chercheur de définir quelle donnée est unique et réutilisable, devant être conservée et potentiellement partagée.

2.2 Publication des données de la recherche

Publier permet de partager des informations, des données dont on a connaissance, avec l'ensemble d'une communauté. Cependant, cette communication n'a pas le même impact selon qu'elle est formelle (publications dans des revues) ou informelle (« mise à disposition sur le Web ») (53). L'intégration des données de la recherche au sein des publications scientifiques est illustrée par la pyramide de publication des données (Figure 7) (54).

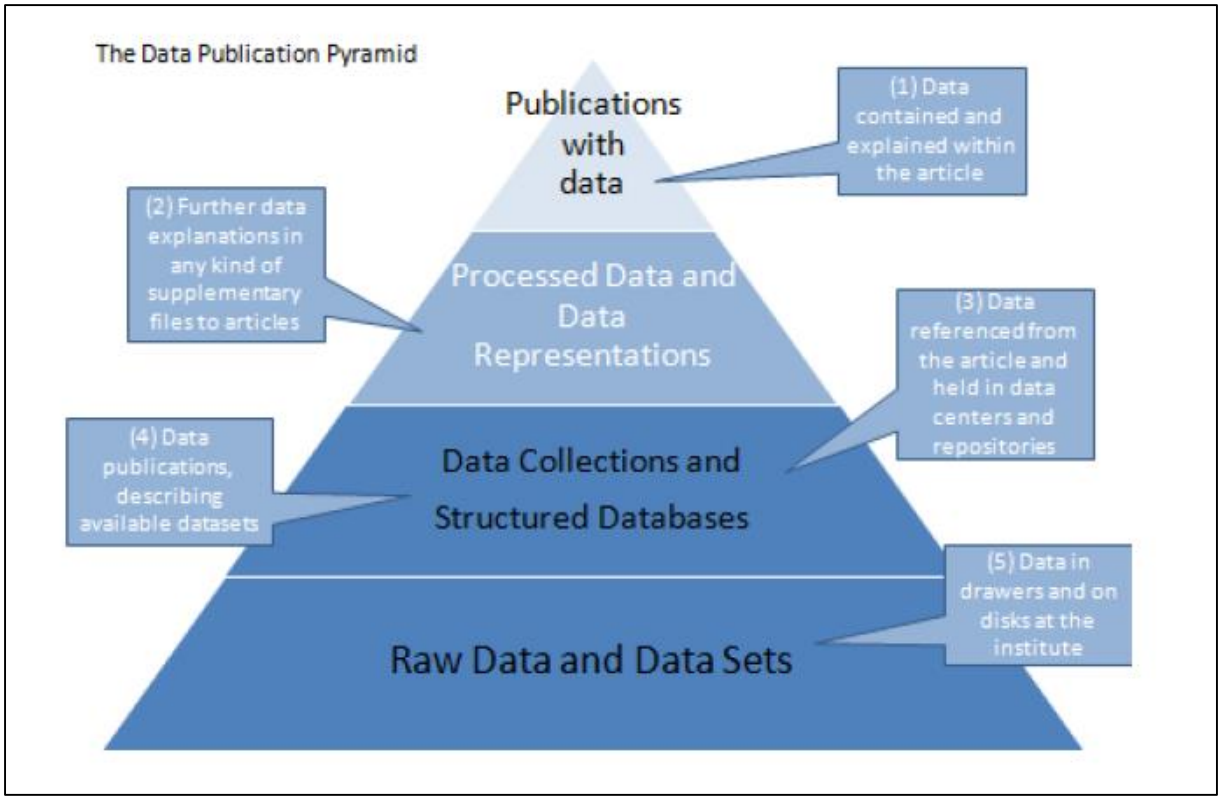


Figure 7 : Pyramide de publication des données (54)

L'objectif principal de cette pyramide est d'expliquer les différentes formes de mise à disposition des données dans le contexte d'ouverture et de partage des données. Au fur et à mesure que nous descendons dans la pyramide, le lien entre

données et publications scientifiques diminue. Les données de la recherche peuvent ainsi être :

- **Disponibles au sein de l'article**. Il s'agit du modèle de publication traditionnelle où les conclusions issues des données sont illustrées en résumant les données pertinentes. Les données sont intégrées à la publication, citables et accessibles à tous. Toutefois celles-ci sont limitées, fortement agrégées et l'accès aux données sources n'est pas garanti.
- **Disponibles dans des fichiers supplémentaires** en complément de l'article. Un grand nombre de revues autorise à joindre des fichiers supplémentaires en cas de matériel pertinent trop volumineux ou ne correspondant pas aux formats des articles (jeux de données, protocoles, fichiers multimédias, grande bibliographie, grands tableaux…). Le volume des données et le format n'est généralement plus un problème et les données restent étroitement liées à l'article. Toutefois des questions se posent quant à la conservation des fichiers et si les critères de découvrabilité et de réutilisation sont remplis. A noter que la gestion des volumes de plus en plus importants a conduit certains journaux à limiter le contenu voir même à ne plus l'accepter.
- **Disponibles au sein d'un référentiel**. L'article scientifique inclut une citation des données (numéro d'accession ou un DOI) et des liens bidirectionnels vers les données stockées au sein d'un référentiel. Les données ne sont pas limitées en volume, sont citables independemment de l'article et sont propices à être réutilisées du fait qu'elles soient faciles à trouver, normalisées et conservées de façon pérenne. Toutefois certains référentiels spécifiques n'existent que dans certains domaines et dépendent souvent d'un financement gouvernemental.

- **Non disponibles**. Ces données correspondent à l'ensemble des jeux de données produits par des études individuelles à petite échelle connues sous le nom de « données à longue traine » (55,56) qui se trouvent dans les tiroirs et disques durs des chercheurs et qui n'ont pas été publiées. De ce fait, ce sont des données introuvables et donc non réutilisables.

Ces différents niveau d'intégration sont décrits plus en détail dans le rapport de l'ODE (54) avec les avantages et inconvénients pour chaque niveau en terme de disponibilité, de repérabilité, d'interprétabilité, de réutilisation, de conservation et de citation.

Historiquement, les données brutes (ou traitées) étaient considérées comme un complément aux résultats obtenus de leur analyse et non comme un produit de recherche à part entière à partager officiellement. Les données publiées dans les articles scientifiques (données résultats) ne représentent que la « partie émergée de l'iceberg » et aujourd'hui « publier ses données » sous-entend la mise à disposition des données brutes (ou traitées). Des études montrent que les articles dont les données brutes (ou traitées) sont accessibles sur Internet sont davantage cités que les articles ne les mettant pas à disposition (57,58). Les ensembles de données pourraient même être plus cités que les articles scientifiques (59,60). Cependant, 90% des données de la recherche seraient stockées sur les disques durs locaux, et donc non disponibles pour d'autres chercheurs de façon instantanée (50,61) réduisant ainsi les chances d'accéder aux données de près de 17% par an (62). Dans le futur, l'idéal est qu'un maximum de données soit publié au sein d'un référentiel (Figure 8), permettant ainsi un accès plus facile et pérenne aux jeux de données.

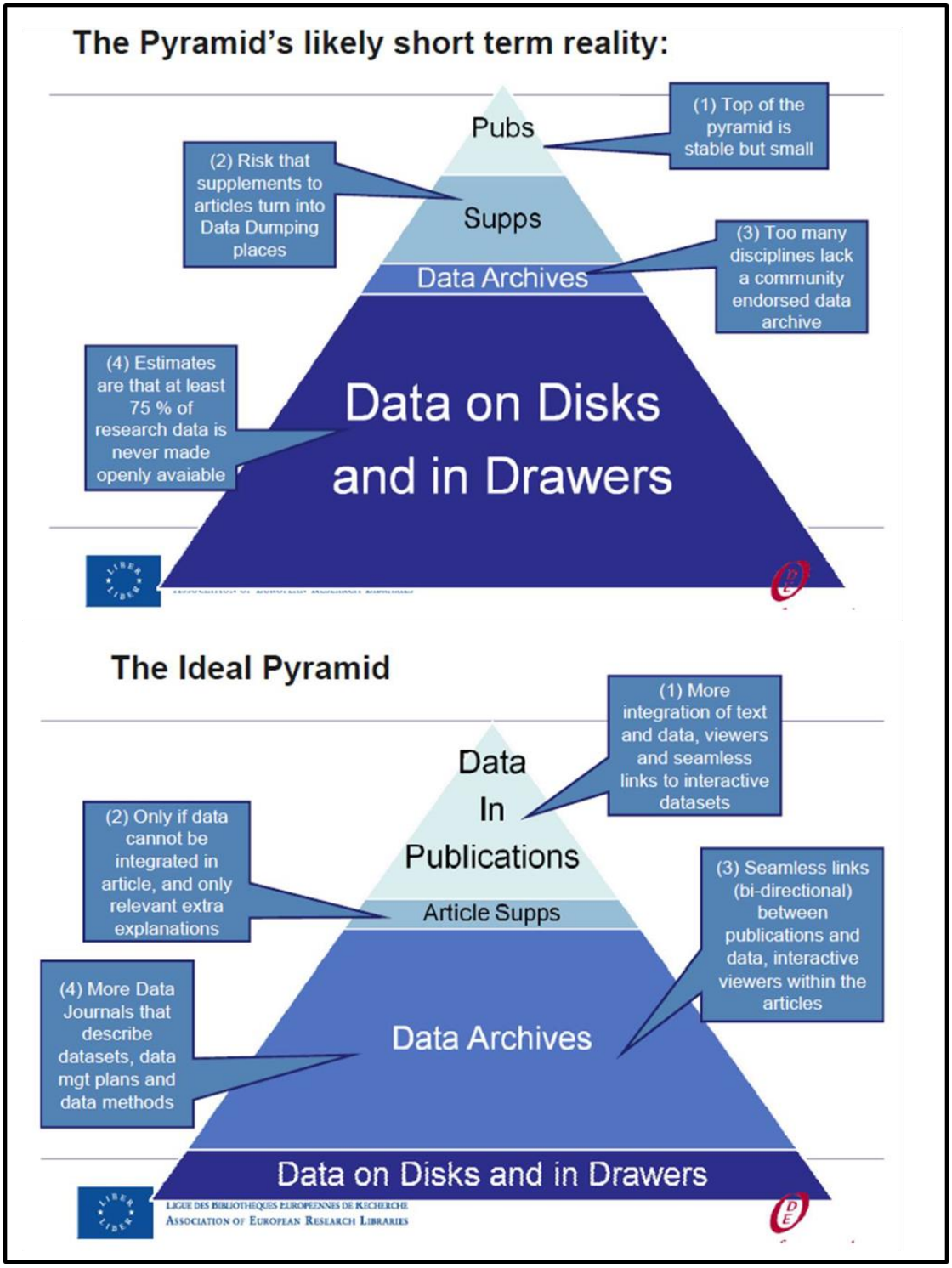


Figure 8 : Evolution attendue de la pyramide de publication des données (63)

2.3 Incitation au partage de données

Le potentiel de partage de données est important et l'intérêt reconnu (Voir chapitre 2.4 Enjeux du partage de données). Les pratiques évoluent au fur et à mesure des lignes directrices éditées en faveur de l'*Open Data* et des nouveaux modèles établis.

Dans de nombreux pays, l'évolution de la législation vise à permettre aux chercheurs de publier en libre accès à la fois les résultats de leurs recherches et leurs données (42,64). En 2003, la *Déclaration de Berlin* sur le libre accès à la connaissance vise à mettre à disposition en libre accès la littérature scientifique mondiale et l'ensemble des données et logiciels ayant permis de produire cette connaissance (65). Autrement dit, elle élargit le concept de libre accès à l'ensemble des produits de la recherche dont les « données brutes » font partie. Par la suite, l'OCDE rappelle que « des efforts coordonnés aux niveaux national et international sont nécessaires pour élargir l'accès aux données de la recherche financée par des fonds publics et contribuer à faire progresser la recherche scientifique et l'innovation » et rédige en 2004 la *Déclaration sur l'accès aux données de la recherche financées par des fonds publics* où le principe de l'accès ouvert est adopté (66). S'en suit en 2007 la rédaction des *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financées sur fonds publics*, jalon important dans l'histoire de l'ouverture des données, dont l'objectif est de fournir des recommandations aux institutions qui cherchent à définir une ligne de conduite en matière de gestion, de conservation et de mise à disposition des données (43). Plus récemment, le G8 adopte en 2013 la *Charte pour l'ouverture des données publiques* qui renforce le principe d'accès et de gratuité de leur réutilisation par tous en privilégiant les formats ouverts et non-propriétaires (67). Cette politique en faveur de l'ouverture des données est inscrit en France dans la *Loi pour une République Numérique* de 2016 (68) et depuis juillet 2018 a été lancé le *Plan*

national pour la science ouverte où « La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens, sans entrave, sans délai, sans paiement » (69).

Au-delà des possibilités données par la législation et les déclarations adoptées par les gouvernements, de plus en plus d'organismes publics qui financent la recherche appuient, voire même exigent, la rédaction de plan de partage de données en contrepartie de leur contribution afin que la publication des résultats et des données de la recherche soient disponibles le plus rapidement en *open access*. Depuis 2003, le *National Institutes of Heath* aux Etats-Unis demande un plan de partage des données ou une explication des raisons pour lesquelles le partage de données n'est pas possible pour toutes les demandes de subvention importante (70–72) et pour toutes les études d'association à l'échelle du génome soutenues ou conduites par le NIH (73). De même, plusieurs fondations dont la *National Science Foundation* exigent que des plans de partage et de gestion des données accompagnent toutes les propositions de subventions de recherche (74–76). En Europe, la *Medical Research Council* et le *Welcome Trust* au Royaume- Unis ont défini depuis plusieurs années leurs politiques quant à la mise à disposition de données (77–79). En outre, en 2014 l'Union Européenne lance *Horizon 2020* le plus grand programme européen de recherche et d'innovation dans lequel une action pilote sur le libre accès aux données de recherche est mise en place (*Open Research Data Pilot*) afin d'améliorer et de maximiser l'accès et la réutilisation des données de recherche générées par les projets qu'elle finance (80). La participation à ce programme est ouverte à tous chercheurs à condition que ces derniers diffusent les résultats qu'ils produisent et élaborent un plan de gestion des données, dans lequel ils spécifieront quelles données seront ouvertes (81).

D’autres organismes influants tel que l’Organisation Mondiale de la Santé et l’Académie nationale de médecine des Etats-Unis publient des rapports demandant un partage responsable des données issues des essais cliniques (82,83). De même, l’Agence européenne des médicaments et plusieurs entreprises de l’industrie pharmaceutique s’engagent de plus en plus quant à la mise à disposition des données (84–86).

La publication d’article, dans des journaux revus par les pairs, est le format principal et reconnu du partage de la recherche scientifique. Les éditeurs de revues scientifiques représentent donc un important « point de levier dans le processus de recherche » et peuvent influencer sur les pratiques de diffusion (42,87,88). Ils peuvent jouer un rôle important en facilitant et/ou en imposant le partage des données. Certaines revues comme *Nature* ou *PLOS* ont fait du partage de données une condition de publication (89–91). D’autres invitent à fournir une déclaration décrivant où et comment accéder aux données (92–95). *The International Committee of Medical Journal Editors* a introduit une politique standardisée de partage des données pour ses revues membres (*BMJ, Lancet, JAMA, NEJM*) et depuis le 1er juillet 2018 exige une déclaration de partage de données comme condition à la publication des essais cliniques (96). Au cours de ces dernières années, la prévalence des politiques de partage des données des éditeurs et des revues n’a cessé d’augmenter (88). Toutefois, les exigences et l’application de ces politiques varient considérablement selon les revues et le domaine de recherche (58,89,97–101). Par ailleurs, l’offre éditoriale s’est étoffée en permettant la publication d’articles spécifiques décrivant certaines parties d’un projet de recherche tels que les méthodes, les protocoles, les logiciels et les données (88). Certains journaux (*Data journals*) et articles (*Data papers*) deviennent spécialisés dans la description d’ensembles de données accessibles au

public dont l'objectif est de fournir suffisamment de détails afin qu'un chercheur puisse trouver, comprendre et réutiliser les données ; l'idée étant par ce biais de favoriser leur diffusion tout en étant reconnue et valorisée (Voir chapitre 3 Le data paper).

Enfin, les impulsions en faveur du partage de données émanent également des communautés scientifiques elles-mêmes. Comme l'indique le rapport du *Research Information Network* : « Quelques disciplines ont largement devancé les agences de financement en ayant, depuis longtemps, une culture du partage, et en développant les infrastructures et les méthodes pour qu'elles s'épanouissent. Dans d'autres disciplines, le partage des données n'est pas la norme et dans ce cas, en effet les politiques des financeurs peuvent influer de manière significative sur les attitudes et le comportement des chercheurs » (102). L'astronomie, la cristallographie ou encore la génomique font partie de ces disciplines ayant une culture du partage. Des référentiels spécifiques à ces domaines, tel que GenBank (103) ou GEO (104) pour la génomique, ont été créés pour stocker les données de recherche dans le but de faciliter leur mise à disposition auprès de la communauté scientifique. Dans le domaine des sciences de la santé, l'émergence de nouveaux référentiels, notamment en imagerie, et de nombreux projets et consortium améliorent depuis plusieurs années la gestion, le partage et l'accès aux données cliniques, biologiques, génétiques et d'imagerie (105–110).

2.4 Enjeux du partage de données

Les données de la recherche ont une valeur et une utilité importante au-delà de l'objectif pour lequel elles ont été collectées à l'origine d'où l'intérêt grandissant quant à leur partage. L'importance du partage et de la réutilisation des données de recherche est bien établie (105,108,111–113). Un accès plus large et plus complet aux publications et aux données scientifiques vise à :

- Faire progresser la science en étudiant des hypothèses supplémentaires ou nouvelles, en réalisant de nouvelles analyses (méta-analyse, application de méthode statistique, …), en développant et validant de nouvelles méthodes d'étude, techniques d'analyse ou implémentation de logiciel
- Améliorer la qualité des recherches en publiant des recherches plus fiables et reproductibles
- Limiter la fraude en permettant la vérification des résultats afin d'identifier des erreurs potentielles
- Augmenter l'efficacité de la recherche en encourageant la collaboration et en évitant la duplication des efforts (redondance de collecte) afin de diminuer les coûts et temps de recherche
- Accélérer l'innovation et la croissance économique
- Améliorer la transparence du processus scientifique

Ainsi, les enjeux du partage des données sont d'ordre scientifiques, économiques et sociétaux. Toutefois, bien que ces multiples bénéfices soient reconnus, la réalité en est tout autre.

2.5 Réalité du partage de données

Les chercheurs reconnaissent l'intérêt et l'utilité du partage de données et montrent des niveaux de motivation élevés pour partager des données et utiliser celles des autres (114–120). Parmi ces motivations, l'augmentation de l'impact et de la visibilité de leur recherche, le bénéfice public, la transparence et la réutilisation sont les principales retrouvées. Malgré ces bénéfices reconnus, près de la moitié des chercheurs déclaraient ne pas partager ses données en 2014 (117,118,120). En 2016, c'est moins de 35% (121) témoignant de l'évolution des pratiques et des perceptions de la part des chercheurs (122). La mise à dispositions des données et le mode de diffusion varient en fonction des domaines de recherche et de l'origine géographique (117,119,120,122–124). L'âge, l'expérience, le type de revue, la taille des données sont également d'autres facteurs retrouvés (122,125,126). De façon générale, le partage entre les pairs était le moyen le plus courant de rendre les données disponibles (61,114,127) mais aujourd'hui, comme le montre certains rapports (117,121,128), la disponibilité au sein de fichiers supplémentaires joints à l'article est la méthode la plus couramment utilisée. L'utilisation de fichier supplémentaire permet de rendre les données davantage disponibles que si elles n'étaient pas partagées mais n'optimise pas la possibilité de découverte, l'accessibilité ou la réutilisation contrairement au dépôt au sein d'un référentiel dont la pratique est encore sous utilisée (entre 5 et 30%) notamment dans les sciences de la santé (117,118,120,121,123,128). Cette sous-utilisation des référentiels et notamment des référentiels publics, peut être expliquée en partie par les différentes raisons liées à l'hésitation des chercheurs à partager leurs données. Les préoccupations les plus courantes sont liées au manque de connaissance sur les normes de publication et les référentiels potentiels, aux contraintes de temps et de ressource nécessaire, à la crainte d'une utilisation abusive

ou inappropriée des données, à l'incertitude sur les droits d'auteur et les licences ou encore à la crainte de ne pouvoir exploiter la totalité du jeu de données (105,114–117,120,121,129). L'ordre de priorisation varie en fonction des disciplines (123) et on retrouve parmi les chercheurs cliniciens une autre préoccupation, celle de la confidentialité des sujets.

La reconnaissance des auteurs en tant que générateur de données et la normalisation des citations de données pourraient être des moyens d'incitation au partage (130–133) en permettant l'attribution de crédits universitaires comme le stipule Bierer et al « par souci d'équité et pour encourager le partage de données, les personnes qui ont recueilli des données devraient recevoir un crédit approprié et normalisé […] » (131). Actuellement, la nécessité de reconnaitre la citation des données est établie mais les normes quant à leur publication restent encore en discussion (53,133–136). L'attribution d'un DOI aux jeux de données lors de leur publication au sein d'un référentiel permet d'améliorer leur découverte et leur accessibilité et favorise leur citation en cas de réutilisation mais ne permet pas encore de bénéficier de crédits universitaires spécifiques (88,132). Plus récemment, certains éditeurs ont proposé la création de nouveau journaux, les *data journals* dont les articles sont spécifiquement dédiés à la description de jeux de données, les *data papers*. Un des avantages de cette nouvelle forme éditoriale est qu'il représente un mode de publication reconnue des chercheurs, la rédaction d'un article scientifique, et peuvent de ce fait être valorisés au même titre qu'un article classique. Les *data journals* et *data papers* restent encore un phénomène à petite échelle mais il semble que leur popularité croit assez rapidement et pourraient devenir dans quelques années une partie beaucoup plus importante du paysage des publications et de la pratique scientifique (128).

3 Le data paper

La mise à disposition de jeux de données nécessite la rédaction de leurs métadonnées, un ensemble structuré d'informations permettant de décrire une ressource (50), afin notamment de les rendre interprétables et interopérables et ainsi faciliter leur réutilisation tel que préconisé par les principes FAIR (137). Or, ceci demande du temps et du travail supplémentaire de la part des chercheurs pour lequels il existe un manque de reconnaissance, élément expliquant en partie les obstacles liés au partage des données (138,139). De ce fait, les métadonnées sont peu voire pas produites (140). Pour tenter de pallier à cet obstacle, Chavan et Penev (140) ont promu des « documents de données sur la biodiversité » pour encourager la publication des données. Kennedy et al (141) ont quant à eux préconisé « un article original sur les données » dans la recherche en neuroscience afin de « soutenir la publication de données de haute qualité, richement réutilisables et entièrement décrites ». Les *data papers* ne seront probablement pas la solution complète aux problèmes de partage de données mais semblent être un mécanisme à fort potentiel incitatif qui pourraient transformer le paysage des publications à l'avenir.

3.1 Définition

Il est classique d'utiliser la définition fournie par Chavan et Penev (140) qui ont défini le *data paper* comme « une publication savante d'un document de métadonnées consultable décrivant un ensemble de données particulier, accessible en ligne, ou un groupe d'ensemble de données, publié conformément aux pratiques universitaires standard ». Autrement dit, le *data paper* est une publication scientifique qui a pour but de décrire de façon précise et détaillée les données, la méthodologie de collecte et le potentiel de réutilisation, soit de fournir « des informations sur le quoi, où, pourquoi,

comment et qui pour ces données » (142). Il s'agit en aucun cas de présenter des hypothèses, de rendre compte ou de discuter de résultats contrairement aux articles de recherche traditionnels mais de se focaliser sur la description des données afin de les rendre accessibles, interprétables et réutilisables. D'après la récente revue de littérature menée par Schöpfel (143), bien que l'objectif principal du *data paper* soit reconnu et défini de façon assez similaire parmi les éditeurs, il manquerait néanmoins une définition de référence du *data paper* et suggère la définition suivante : « le contenu principal est une description des ensembles de données de recherche publiées, ainsi que des informations contextuelles sur la production et l'acquisition des ensembles de données, dans le but de faciliter la recherche, la disponibilité et la réutilisation des données de recherche ; ils font partie de la gestion des données de recherche et sont réticulés aux référentiels de données ».

3.2 Concept

Avec ce nouveau type de document, il ne s'agit en aucun cas de créer un document se substituant au dépôt du jeu de donnée dans un référentiel ou à la rédaction des métadonnées de base demandées, mais plutôt de créer un document support accompagnant le dépôt (servant de « proxy »), une forme de métadonnées enrichies. Candela et al (144) ont établi un modèle assez simple pour expliquer le concept du *data paper* (Figure 9). Ce modèle est constitué de deux éléments « devant être matérialisés en objet d'information concrets et identifiables » :

- Le jeu de données, objet du document
- Le *data paper* lui-même c'est-à-dire l'artéfact produit pour décrire l'ensemble de données

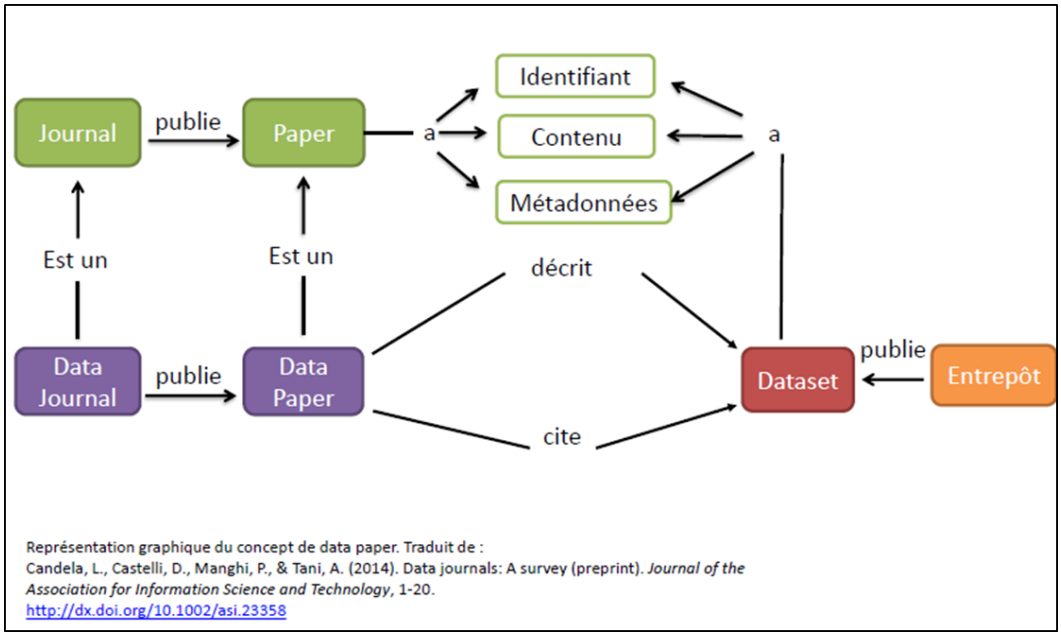


Figure 9 : Concept du *data paper* (145)

Bien que le *data paper* diffère de l'article de recherche traditionnel du fait qu'il ne s'intéresse pas à la recherche mais aux données, on retrouve néanmoins une certaine analogie (146–148). Généralement, un identifiant, un contenu (titre, auteur, résumé, sections, référence) comprenant une description avancée des données et l'accès au jeu de données (que ce soit au sein de l'article, éventuellement en fichier supplémentaire, ou par un lien vers l'entrepôt de données) est retrouvé même s'il n'existe pas encore de format standardisé (143,144,148–150). Du fait de cette similarité avec des articles de recherche traditionnels, il s'agit d'un document qui peut être plus facilement accepté, évalué par les pairs et citable (147).

Ainsi, le *data paper* a pour objectif d'accroitre la visibilité et l'accessibilité des données de la recherche et d'en améliorer leur compréhension sous une forme structurée et lisible par l'homme tout en permettant leur valorisation selon un mode reconnu par le monde de la recherche ; la publication d'un article au sein d'une revue évaluée par les pairs dont la reconnaissance via la citation et l'attribution de crédit universitaire est alors possible.

3.3 Environnement

Les *data papers* peuvent être publiés soit dans des revues scientifiques traditionnelles, dites « mixtes », qui acceptent ce nouveau type d'article soit dans des revues spécifiques, dites « pures » c'est-à-dire qui publient exclusivement ou principalement des *data papers*, les *data journals*. Bien qu'il existe de plus en plus de revues mixtes qui annonce le lancement de nouvelles rubriques consacrées aux *data paper* (151–156), la création de journaux spécifiques a augmenté au cours de ces dernières années (157). En avril 2013, *Nature* annonçait le lancement de *Scientific Data*, une revue spécifique permettant de décrire des ensembles de données de grande valeur scientifique (158). *Biomedical Data Journal* est créé en 2014 et vise à faciliter la présentation, la validation, l'utilisation et la réutilisation des ensembles de données en mettant l'accent sur la publication de jeux de données biomédicaux (159). Alors que Candela recensait en 2014 7 revues pures (144), Garcia-Garcia en identifiait 20 dans son étude de 2015 (157) et ce chiffre a été porté à 28 par Schöpfel en 2019 (143). *Geoscience Data Journal*, *Dataset Papers in Science, Journal of Open Archaeology Data* ou encore *Open Health Data* sont d'autres exemples de *data journals* ; la majorité des *data journals* identifiés concerne principalement les sciences médicales, de la vie et de la terre (143,144,157). Par ailleurs, la première étude sur les *data journals* menée par Candela (144) montrait déjà une montée en charge de ce nouveau type de publication, avec entre 2000 et 2013 23,5% (195 articles) des *data papers* publiés au cours de la dernière année. Entre la date de leur création et 2016, 60% des articles parus dans l'un des trois plus grand *data journals* (*Data in brief*, *Biodiversity Data Journal* et *Scientific Data*) l'ont été au cours de l'année 2016 (128). Cette évolution ne cesse de progresser avec un nombre de *data papers* publiés estimé

à 11500 en 2019 (143) alors qu'il était de 825 en 2013 (144) témoignant d'une popularité croissante.

Cette nouvelle forme éditoriale semble progressivement trouver sa place au sein de l'environnement scientifique. Le *Plan national français pour la science ouverte* recommande « le développement des *data papers* » (69). Alors que les discussions sur la publication et la citation d'ensembles de données sont facilement retrouvées dans la littérature (53,106,134,134,141,142,160–165), il existe peu d'études sur les approches et les tendances mises en œuvre par les éditeurs et leurs revues concernant la publication de ce nouveau genre. Les études retrouvées portent sur l'environnement de publication (éditeur, discipline, dénomination, évolution), la structuration de l'article ainsi que sur les politiques de publication des revues (disponibilité des données, évaluation par les pairs, coûts et licences) (143,144,146,148,149,157). Il en ressort une certaine hétérogénéité dans les pratiques montrant un manque de stratégie commun pour promouvoir une description efficace. Par exemple, on retrouve une dizaine de dénominations différentes pour désigner ce type de document (144). De même, il n'existe pas de consensus sur le contenu hormis la nécessité de mentionner la disponibilité des données. Par ailleurs, il n'a pas été retrouvé de retour d'expérience sur la publication de *data papers* spécifique à un domaine d'étude, tous les travaux ayant été réalisés selon une approche générale. Une des perspectives mentionnée par Schöpfel (143) serait d'obtenir « davantage d'études sur les liens spécifiques entre la gestion des données de la recherche, la publication académique et la production et diffusion des *data papers* dans un environnement et une communauté donnés (équipement, discipline, structure…) ». Des travaux supplémentaires sont à mener par tous les acteurs impliqués afin d'établir

des principes reconnus sur la contribution de ces *data papers* pour la recherche scientifique et un cadre commun pour leur publication.

Ainsi, dans ce contexte d'évolution grandissante du *data paper*, le domaine médical apparaissant comme une des disciplines prédominantes, il semble qu'il serait tout à fait pertinent d'étudier l'évolution, l'impact et l'accessibilité de cette nouvelle forme de communication scientifique au sein des sciences de la santé.

4 Objectifs

L'objectif principal de ce travail est double. Premièrement, il s'agit d'identifier les *data papers* du domaine médical et d'étudier leur évolution et les journaux qui les publient. Deuxièmement, il s'agit d'étudier plus spécifiquement les *data papers* propres à la neuroscience en décrivant :

- Leurs caractéristiques
- Leur indexation au sein des trois grandes bases de référencement (WoS, Scopus et Pubmed)
- Les équipes de recherche
- La typologie des données décrites
- L'accessibilité réelle des données décrites
- Leur impact en termes de citation en évaluant la fréquence, l'évolution et le délai de citation
- Leur impact en termes de couverture *altmetrics*

Les objectifs secondaires de l'étude sont d'étudier, en menant une analyse exploratoire :

- La corrélation entre le nombre de citations et différentes caractéristiques bibliométriques (nombre d'auteurs, nombre d'instituts, nombre de pays)
- L'influence de différents facteurs sur la fréquence, le nombre et le délai de citation ainsi que sur les scores *altmetrics*
    - Ancienneté de l'article (2010 2016 vs 2017 2018)
    - Type de journal (*Data journal* vs Revue mixte)
    - Type de données (Clinique vs Imagerie)
    - Data journal (*Scientific Data* vs *Data in brief)*

Les résultats de cette étude pourraient permettre d'apporter davantage de connaissance concernant cette nouvelle forme de communication scientifique et d'étudier ce nouveau phénomène au sein d'un domaine spécifique.

Matériels et Méthodes

1 Stratégie de recherche

1.1 Identification des articles

Recherche Pubmed

Pour ce travail nous avons dans un premier temps réalisé une recherche systématique sur la base de données PubMed. Nous avons recensé les articles identifiés comme étant des *data papers*, publiés jusqu'au 31 octobre 2018, en utilisant les termes « database » et « dataset ». La publication de ce type d'article étant relativement récente et ne connaissant pas l'importance de leur publication dans le domaine des sciences de la santé, nous n'avons pas défini de date de début de publication. Par ailleurs un filtre sur l'espèce humaine a également été appliqué.
Ainsi, l'algorithme de recherche utilisé a été le suivant :

(((Dataset[Publication Type]) OR Database[Publication Type])

AND

("1990/01/01"[Date - Publication] : "2018/10/31"[Date - Publication]))

AND

humans[MeSH Terms]

Lors de nos recherches préliminaires pour nous familiariser avec le sujet, plusieurs listes de revues identifiées par différents auteurs comme publiant des *data papers* ont été récupérées (Annexe 1). Face au nombre relativement important de ces revues, nous nous attendions à ce que ces dernières soient plus représentées dans la

recherche Pubmed que nous avons menée. En effet, les articles identifiés ont été publiés dans moins de 10% des revues mentionnées dans ces listes, mais également dans des revues non recensées. Il nous a donc semblé intéressant de mener en complémentarité une recherche manuelle à partir des revues identifiées comme publiant des *data papers*.

Recherche manuelle

Les revues ont été identifiées à partir de 8 listes établies par différents auteurs (Annexe 1). L'éligibilité des revues a été déterminée par lecture de leur titre ou en vérifiant le champ d'application en cas de libellé non explicite. Les revues ont été exclues si elles étaient en rapport avec un autre domaine que celui de la santé ou de la biologie. Les revues en rapport avec la microbiologie ont également été exclues. Les revues multidisciplinaires ont quant à elles été incluses. A noter que les revues concernant le champ du génome ont dans un second temps été exclues. En effet, les premiers journaux visités concernant la génomique ont montré une quantité importante d'articles disponibles rendant le travail peu compatible avec les moyens humains disponibles pour cette étude.

Sur les 104 revues ainsi sélectionnées, la recherche de *data papers* a été réalisée si la revue disposait de directive (politique éditoriale) dans laquelle était mentionnée de façon explicite que la revue publiait ce type d'article. La dénomination utilisée par la revue devait correspondre uniquement à l'identification de ce type de document. Autrement dit, si la revue regroupait sous le même terme à la fois des *data papers* et des documents de logiciel par exemple, la recherche n'était alors pas menée, l'objectif étant de pouvoir identifier aisément ces documents au sein de l'ensemble des articles publiés par la revue. Enfin, la revue devait donner la possibilité d'accéder

directement à l'ensemble de ses articles et de pouvoir identifier leur typologie facilement (filtre, typologie indiquée au niveau du titre, …).

Pour finir, les revues ont été classées en deux catégories selon qu'il s'agissait de revues mixtes ou de *data journals*.

1.2 Sélection des articles

De façon similaire à la recherche Pubmed, nous avons recherché les *data papers* publiés jusqu'au 31 octobre 2018 dans les revues précédemment sélectionnées, sans fixer de date initiale de publication. Pour l'ensemble des articles (recherche Pubmed et manuelle), leur éligibilité a été évaluée sur deux critères : la population et le domaine d'étude. Ainsi, étaient inclus les articles concernant la population humaine de façon unique et exclus les articles portant sur le génome ou n'appartenant pas au domaine de la santé.

Nous avons, dans un premier temps, jugé de l'éligibilité de ces papiers par lecture des titres. En cas de doute sur le domaine d'étude ou si la population n'était pas spécifiée, les articles ont été conservés afin de statuer quant à leur inclusion à l'étape suivante. Les résumés (et éventuellement les métadonnées associées), si disponibles, ont ensuite été lus et les mêmes critères ont été appliqués. Toutefois, si la population n'était pas précisée ou insuffisamment explicite alors l'article était exclu à cette étape. Par insuffisamment explicite on entend qu'il n'était pas évident de déduire par simple lecture qu'il s'agissait de population ou de matériel humain pour un individu non spécialisé du domaine d'étude. Par exemple, un article présentant des résultats sur des cellules HUVEC, si l'acronyme n'était pas explicité, alors l'article a été exclu tandis que si l'on retrouvait le terme « sur des cellules humaines HUVEC »

ou « sur des cellules HUVEC (human umbilical vein endothelial cell) » l'article était inclus.

1.3 Typage des articles

Une fois sélectionnés, les articles ont été classés en fonction de leur domaine de spécialité à partir du résumé. Pour cette classification, l'organe d'étude a tout d'abord été pris en compte. S'il n'y avait pas de précision sur l'organe ou en cas d'organe multiple, c'est la pathologie de l'étude qui a été utilisée pour choisir la spécialité. Enfin, dans le cas où ni organe ni pathologie spécifique étaient étudiés, alors les articles ont été classés dans des spécialités plus générales tel que la Santé Publique, Biologie ou Pédiatrie par exemple. A noter toutefois une exception concernant la cancérologie, où tous les articles traitant de cette pathologie ont été regroupés quel que soit l'organe d'intérêt.

Par ailleurs, les articles ont également été typés en fonction de la thématique d'étude, à partir du résumé, en trois catégories : imagerie, biologie ou autre. La catégorie « autre » a été précisée selon qu'il s'agissait d'une thématique plutôt clinique ou de santé publique pour les articles éligibles à la lecture du texte intégral.

2 Collection des données

L'ensemble des articles ne pouvant être étudié, il a été choisi de se concentrer sur le champ des neurosciences. La collecte des données s'est déroulée en trois temps afin de récupérer l'ensemble des éléments nécessaires pour répondre à nos objectifs.

Tout d'abord, des informations générales à l'article et spécifiques au contenu ont été obtenues par lecture du texte intégral : le titre, la date de publication, la

disponibilité de l'article, le nombre de pages et de références, le nombre de pays et l'origine géographique des équipes de recherche, le nombre d'auteurs et le nombre d'instituts affiliés, le nom des premiers et derniers auteurs et leur organismes de rattachements, la typologie des travaux (type d'étude, acquisition des données, population d'étude, pathologie), la typologie des données décrites et leur accessibilité. Par ailleurs, l'accessibilité des données décrites a été testée.

Puis, nous avons recherché pour chaque article s'ils étaient indexés dans chacune des trois grandes bases de recherche bibliographique que sont Pubmed, Web of Science et Scopus. Dans le cas où les articles étaient retrouvés, nous avons relevé les termes utilisés par ces bases de données pour typer ce genre d'article.

Enfin, nous avons recueilli des éléments de bibliométrie ainsi que les scores *altmetrics* en date du 19 juillet 2019 à partir de la base de données Scopus. La recherche sur la base Scopus nous a permis d'obtenir des données bibliométriques globales (H-index, nombre de citations, nombre de citations par année, domaine d'étude des citations, typologie des citations, titre et année des premières et dernières citations) mais également des données bibliométriques en excluant les auto citations des premiers et derniers auteurs (nombre de citations, titre et année des premières et dernières citations). Le nombre de citations en excluant les auto citations de l'ensemble des auteurs a également été obtenu. La date de publication de chaque citation retenue a été par la suite recherchée. Lorsque la date de publication était incomplète, une imputation au premier du mois a été faite lorsque le jour manquait et une imputation au premier janvier a été faite lorsque seule l'année était disponible. Les différents scores *altmetrics* ont été obtenus sur Scopus dont les données sont en elle-même issus de *PlumX Metrics* (Annexe 2). Ces scores permettent d'avoir des informations sur la façon dont les individus interagissent avec les éléments de

recherche dans l'environnement en ligne. La composition des scores est expliquée en Annexe 2. Nous avons considéré un score *altmetric* total en sommant l'ensemble des éléments tout en excluant les citations.

3 Analyses statistiques

Une analyse descriptive des données a été réalisée en utilisant les méthodes classiques de la statistique descriptive. La moyenne, écart-type, médiane et valeurs extrêmes ont été calculés pour les variables quantitatives, effectifs et fréquences pour les variables qualitatives.

Afin de comparer les différents groupes, un test du khi-2 ou de Fisher exact en cas de petits effectifs a été réalisé pour les variables qualitatives et un test t de Student ou de Mann-Whitney-Wilcoxon en l'absence de normalité pour les variables quantitatives. En l'absence de la normalité des variables, le coefficient de corrélation de Spearman a été calculé pour déterminer le lien entre le nombre de citations et les caractéristiques bibliométriques.

L'incidence cumulée de citation a été estimée avec la méthode de Kaplan-Meier. Le temps jusqu'à citation a été défini comme le délai entre la date de publication du *data paper* et la date de publication de la première citation. Les *data papers* sans citation au 19 juillet 2019 ont été censurés à cette date. Pour les citations dont la date de publication était antérieure à celle du *data paper*, la date de publication a été remplacée par la date de publication du *data paper.* Le délai médian de citation et son intervalle de confiance à 95% a été estimé, si le paramètre était calculable. Les taux de citation à 3, 6, 12 et 18 mois ont été estimés avec les intervalles de confiance à 95%. Le test de comparaison du Log rank a été utilisé pour comparer les délais de citations entre les différents groupes.

Nous avons considéré un seuil de significativité de 5% pour toutes les analyses. L'analyse statistique a été réalisée avec le logiciel R (version 3.3.3).

Résultats

1 Stratégie de recherche

Sur les 228 articles identifiés par la recherche Pubmed, 127 ont été inclus après lecture du titre. La recherche manuelle a quant à elle permis de recenser 1370 articles au sein des 85 revues sélectionnées. Sur les 193 revues initiales, issues des 8 listes de travail (Annexe 1), 85 ont été exclues liées au domaine d'étude, 1 n'était pas un journal, 3 correspondaient à des doublons, 11 ne permettaient pas d'identifier le type d'article souhaité et enfin 8 revues concernant le génome ont été exclues secondairement (Figure 10). Au final, 1413 articles uniques ont été retenus à partir du titre. L'application des critères d'éligibilité à la lecture des résumés a entrainé l'exclusion de 637 articles. Par ailleurs, le résumé n'était pas disponible pour 31 articles. Un total de 745 articles a donc été inclus (Figure 11).

2 Data papers en sciences médicales

2.1 Revues

Les 85 revues identifiées comme couvrant la thématique biomédicale ont été publiées par 14 éditeurs différents. La plupart de ces revues sont mixtes (90,6%). Seules 8 revues sont des *data journals* publiés par six des quatorze éditeurs (*Elsevier, Hindawi, MDPI, Nature, Procon ltd, Ubiquity Press*). Les éditeurs et de ce fait les revues, proviennent pour la plupart du Royaume-Uni. La Bulgarie, l'Allemagne et la Suisse sont également des pays de provenance retrouvés ainsi que l'Egypte et les Pays Bas. Les éditeurs provenant d'Allemagne (*Springer, Willey*) ne publient pas de *data journals* couvrant les sciences de la santé (Tableau 1).

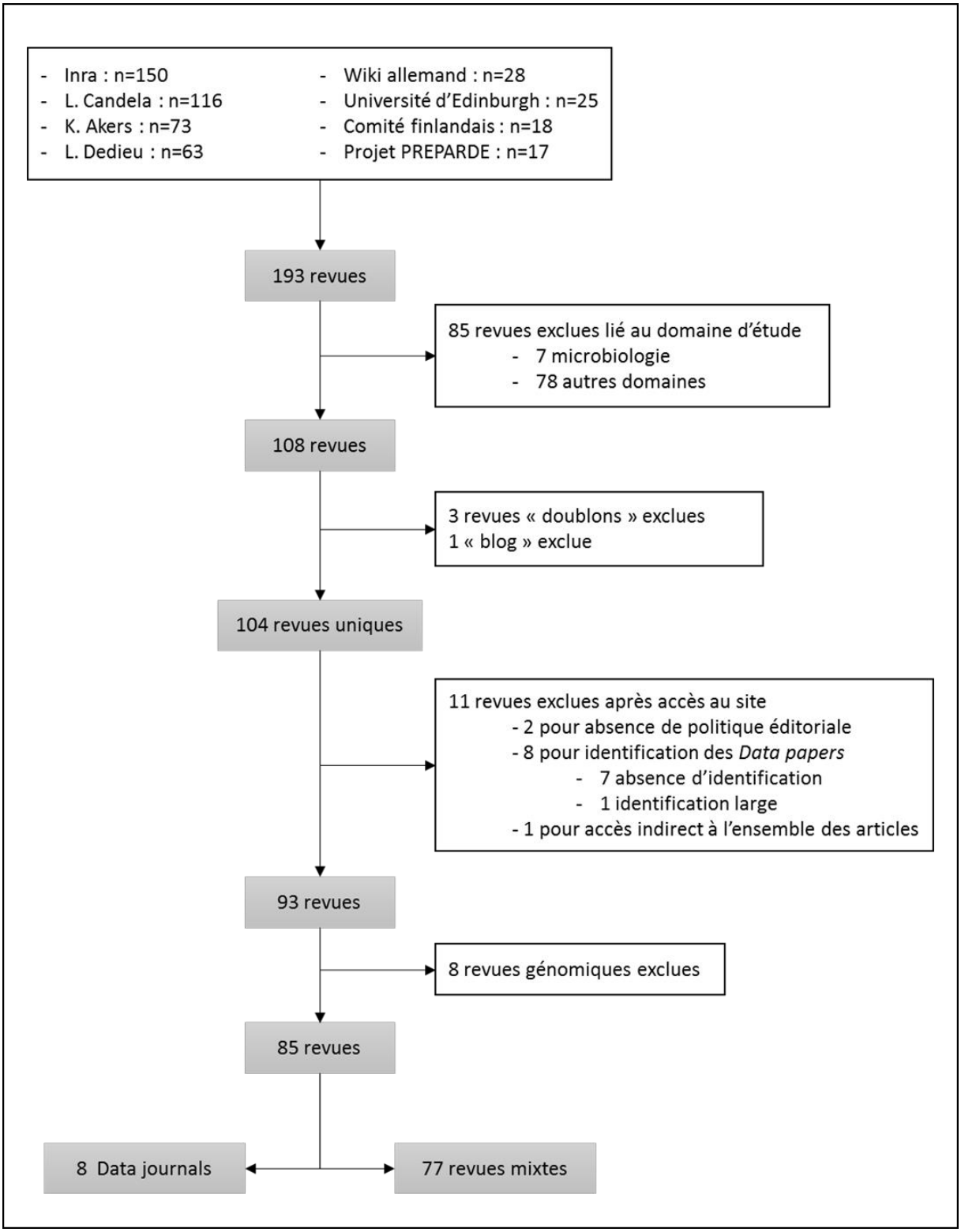


Figure 10 : Diagramme de flux des revues

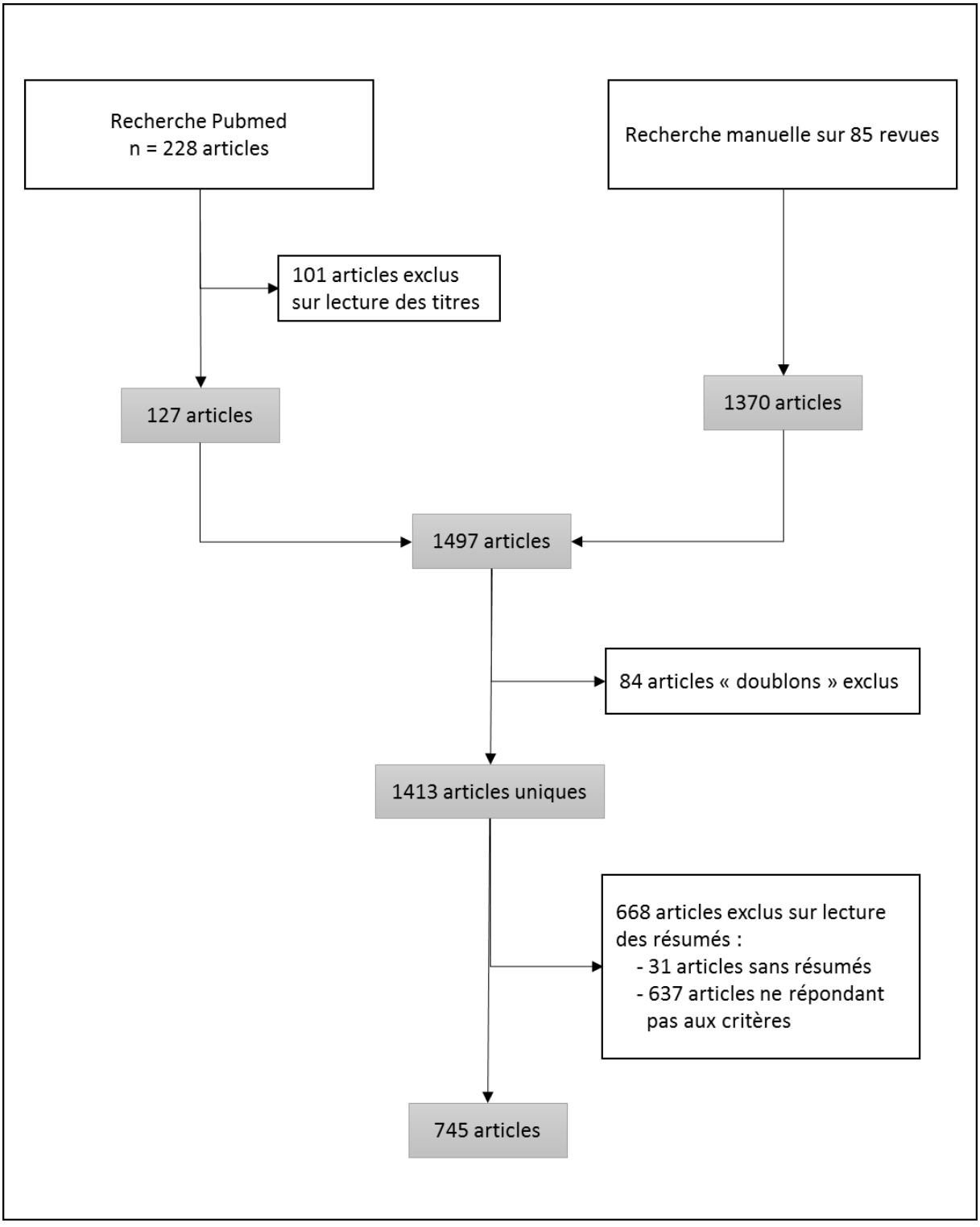


Figure 11 : Diagramme de flux des articles

Tableau 1 : Origine géographique des éditeurs

Origine	N éditeur	N journaux	N data journal	N article
Allemagne	2	4	0	43
Bulgarie	2	3	1	6
Egypte	1	1	1	2
Pays-Bas	1	1	1	424
Royaume-Uni	6	74	4	247
Suisse	2	2	1	3
Total	**14**	**85**	**8**	**725**

58 de ces revues (68,24%) sont indexées dans WoS et 73 (85,88%) dans Scopus. Toutefois, la grande majorité des *data journals* ne sont pas indexés dans ces bases de données (1 dans WoS et 2 dans Scopus) (Figure 12).

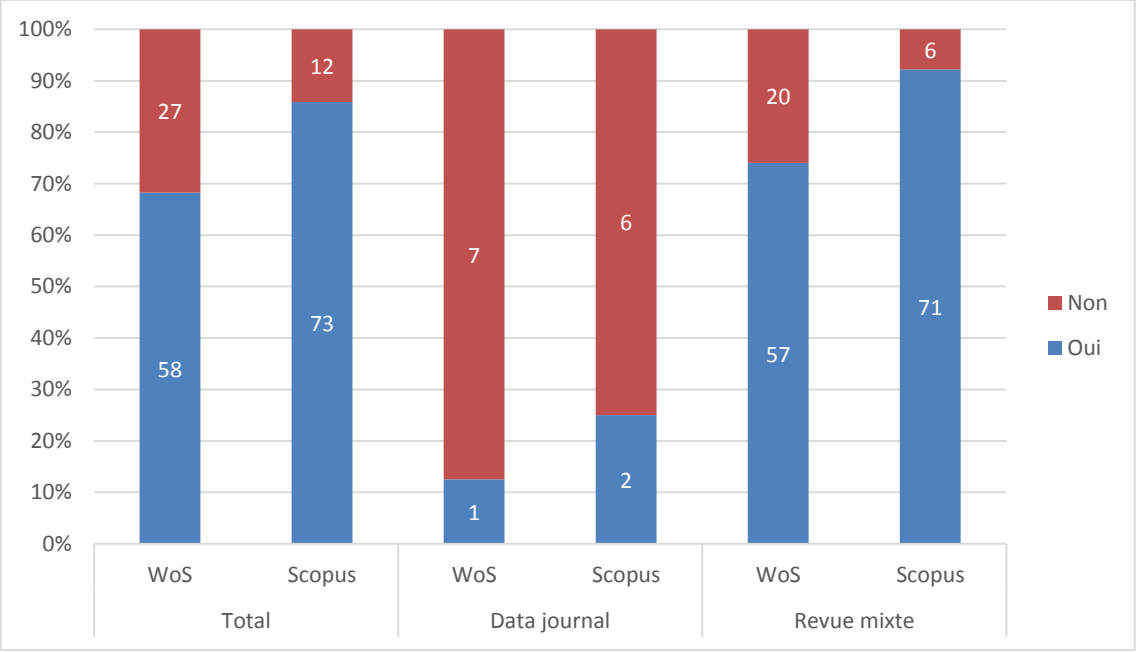
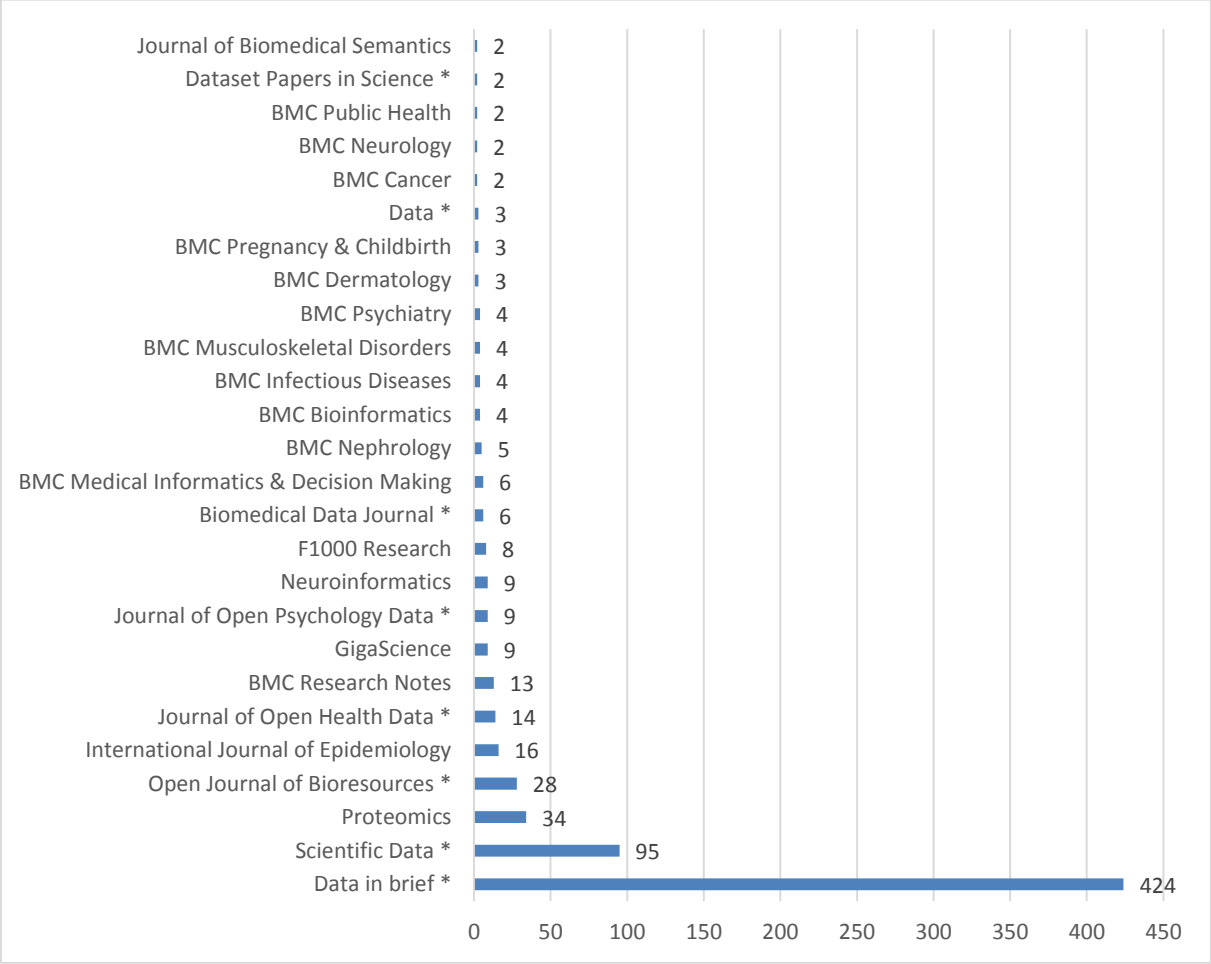


Figure 12 : Indexation des revues dans les bases bibliographiques

Parmi ces 85 revues, où la recherche manuelle a été menée, 725 *data papers* ont été retrouvés correspondant à nos critères de recherche. Il existe une grande différence de publication entre les revues allant d'aucune publication à 424. 52,94% des revues n'ont pas publié d'article de données correspondant à nos critères et

16,47% en ont publié un seul ; toutes correspondaient à des revues mixtes. Parmi les 40 revues ayant publié au moins un *data paper*, 25% en ont publié plus de 8 mais le nombre médian d'articles publiés est de 3 ce qui reste assez faible. Tous les *data journals* ont publié des articles dont 2 avec un nombre d'articles inférieur ou égal à la médiane. 5 revues mixtes font partie des 10 plus gros publieurs (Figure 13).



** Data journal*
Les revues avec moins de 2 publications sont indiquées en Annexe 3

<u>Figure 13</u> : Nombre de *data papers* par journal

Data in Brief d'*Elsevier* est de loin le journal avec la plus grande volumétrie suivie de *Scientific Data* un journal du groupe *Nature*. A eux deux, ces *data journals*

multidisciplinaires, représentent plus de 70% des *data papers* sélectionnés pour notre étude.

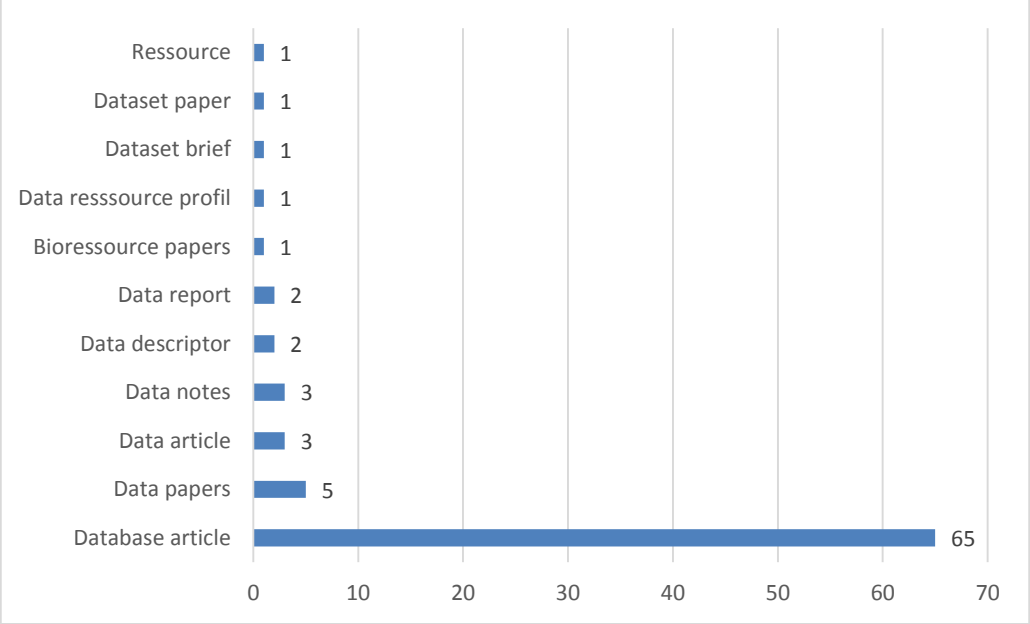


Figure 14 : Nombre de journaux par dénomination de l'article

La recherche des articles au sein des revues a montré une certaine hétérogénéité quant à l'appellation de ce type de document (Figure 14). Les 8 *data journals* utilisent 5 dénominations différentes et ce nombre s'élève à 8 pour les revues mixtes. Par ailleurs, chez un même éditeur, quel que soit le type de journal, il est aussi retrouvé des nominations différentes. Par exemple, pour les revues mixtes de *Biomed Central* on retrouve l'appellation « data notes » ou « database article », pour celle de *Springer* « data article » ou « database article », et pour celle de *Willey* « data report » ou « dataset brief ». *Ubiquity Press* utilise un nom différent entre leur revue mixte « data article » et leur *data journal* « data papers » ou « bioressource papers ». Le groupe *Nature* utilise le terme « data descriptor » ou « ressource » en fonction de la typologie du journal.

88 *data papers* inclus dans notre étude et retrouvés dans notre recherche Pubmed initiale, ont été publiés au sein de 15 revues dont 3 faisant partie des 85 revues sélectionnées pour la recherche manuelle (*Scientific Data, Gigascience, Proteomic*) et 1 appartenant au listing initial mais exclu par la suite (*Database - The Journal of Biological Databases and Curation*). 11 revues ont publié des *data papers* sans que ces derniers ne soient recensés dans les listes des différents organismes ayant travaillé sur la thématique (Annexe 4). Il n'a pas été mené de recherche supplémentaire au sein de ces revues et seuls les articles disponibles via Pubmed ont été pris en compte par la suite. A noter que pour la quasi-totalité de ces revues il n'a été retrouvé qu'un seul *data paper* indexé dans Pubmed correspondant à nos critères de recherche. Ainsi, la recherche Pubmed a permis d'inclure 20 articles supplémentaires.

2.2 Articles

Un total de 745 articles a été inclus à la suite de notre revue de la littérature. Les articles ont été publiés au sein de 52 revues différentes, 8 *data journals* et 44 revues mixtes. Bien que le nombre de *data journals* soit bien inférieur au nombre de revues mixtes, ces nouveaux journaux ont publié 78% des articles (Figure 15). 8,6% des *data papers* ont été publiés dans une des 8 revues non référencées dans Scopus, principalement des *data journals*. Pour les 44 journaux référencés dans Scopus, le CiteScore 2017 médian est de 2,5 (min 0,55 – max 8,86). A noter qu'il existe une différence entre le CiteScore des deux *data journals* référencés, *Data in Brief* et *Scientific Data* avec un indice respectif de 0,70 et 6,08.

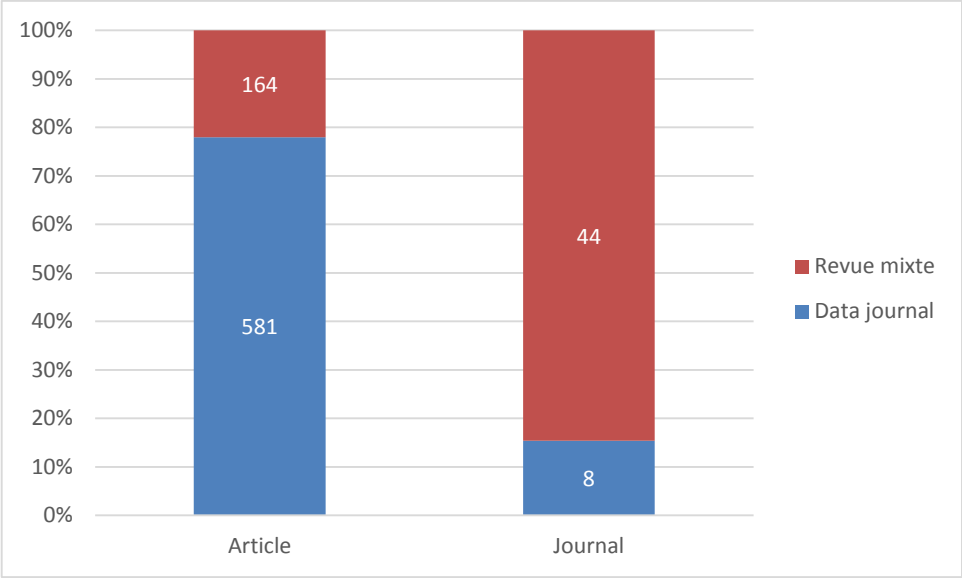


Figure 15 : Répartition des articles et revues selon la typologie de la revue

Le nombre de *data papers* publiés par année a fortement augmenté jusqu'en 2016 et est globalement stable depuis (Figure 16). A noter toutefois que l'année 2018 ne comprend que les 10 premiers mois de l'année ce qui laisse supposer un nombre d'articles potentiellement en augmentation. 71,4% des documents ont été publiés entre 2016 et 2018 témoignant d'une pratique relativement nouvelle. De même, le nombre de revues publiant au moins un *data paper* a été en croissance jusqu'en 2014 pour se stabiliser ces dernières années (Figure 17). Pour l'année 2018, 18 journaux différents ont publié un total de 182 *data papers*. L'apparition des *data journals* est récente avec un nombre qui a doublé entre 2013 et 2014 et il n'a pas été observé par la suite de nouvelle création de journal. Au contraire, certains cessent de publier comme *Dataset Papers in Science* en 2017. La publication d'un article de données au sein d'une revue mixte progresse mais reste limitée bien que le nombre de revues faisant leur promotion soit en augmentation.

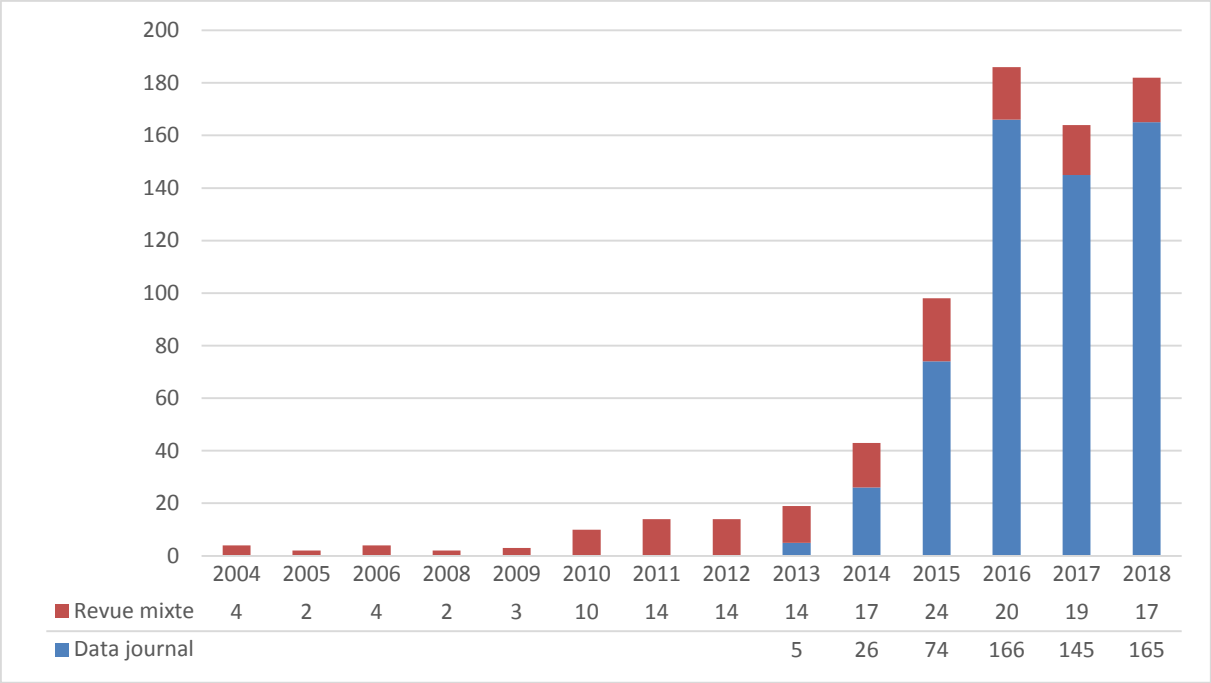


Figure 16 : Evolution de la publication des *data papers*

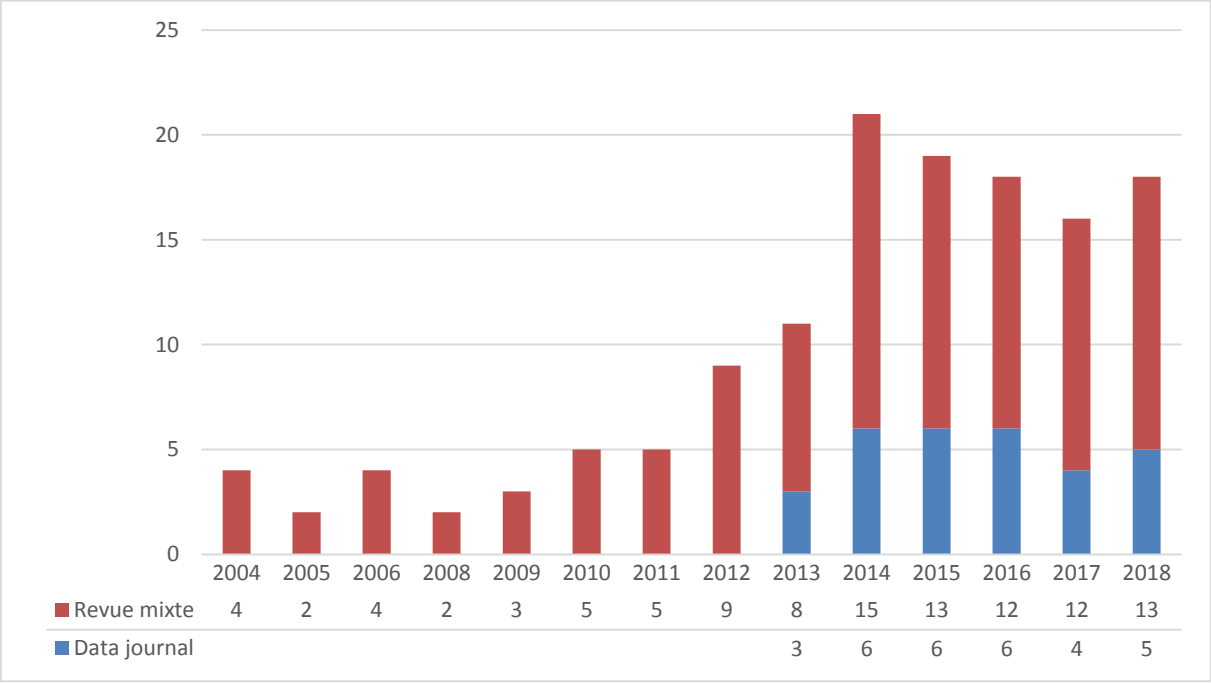


Figure 17 : Evolution du nombre de revues ayant publié un *data paper*

En ce qui concerne la thématique des articles, 12% porte sur de l'imagerie, 19% sur des données de biologie et 59% sur un autre sujet (clinique ou santé publique). En analysant la spécialité médicale des articles, il apparait que la neurologie (21,9%), la cardiologie (13,6%) et la cancérologie (10,9%) soit les trois spécialités les plus représentées comme utilisant ce nouveau type de publication (Figure 18). Il est assez intéressant de noter que la plupart des spécialités médicales a déjà publié au moins une fois un *data paper* même si la neurologie reste de loin le plus gros publieur.

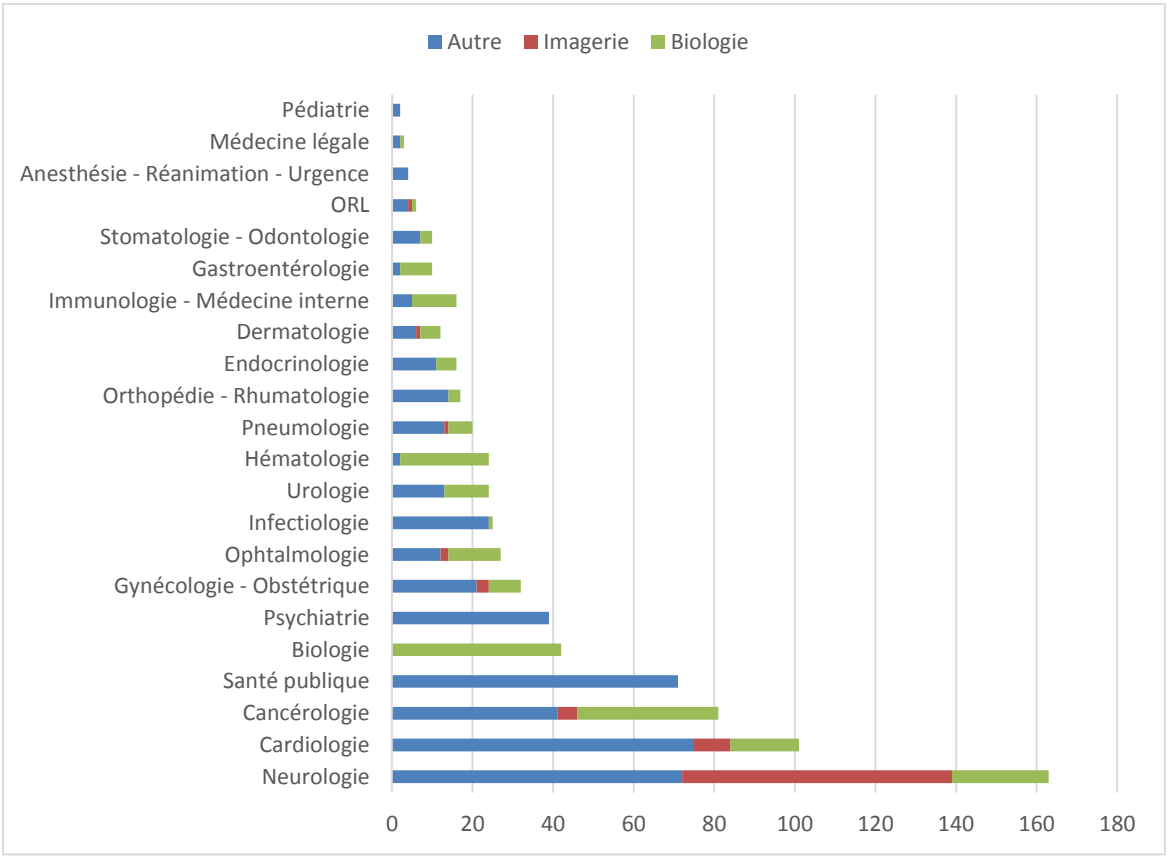


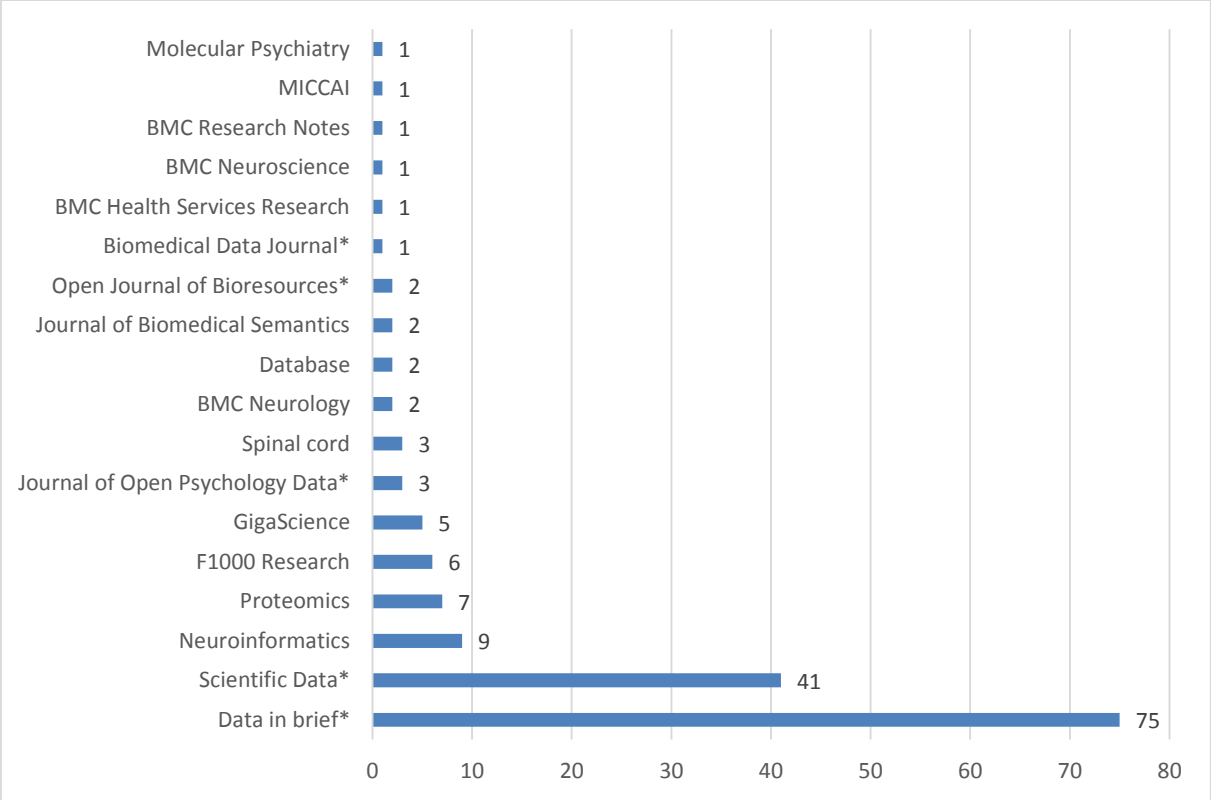
Figure 18 : Répartition des *data papers* selon la spécialité médicale

Afin d'étudier plus en détail les données décrites, leur accessibilité et leur impact nous nous sommes concentrés sur les *data papers* de neurologie. Il s'agissait de la spécialité la plus représentée avec l'ensemble des thématiques publiées dans des revues mixtes (25%) et des *data journals* (75%).

3 Data papers en neuroscience

3.1 Caractéristiques générales

L'analyse plus approfondie des *data papers* porte sur les 163 articles de neurologie, tous écrits en anglais. Pour 15 d'entre eux (9,2%), le texte intégral n'était pas disponible en *open access* au sein de la revue mais 6 l'étaient toutefois à partir de PubMed Central. Ces articles ont été publiés par 18 revues différentes dont 5 *data journals* pour lesquels le nombre d'articles publié s'élève à 122 (74,8%) (Figure 19).



** Data journal*
MICCAI : Medical Image Computing and Computer-Assisted Intervention

Figure 19 : Nombre de *data papers* en neuroscience par journal

L'évolution de leur publication au cours du temps suit la même tendance que celle de l'ensemble des *data papers* médicaux avec une grande majorité des articles publiés au cours de ces 3 dernières années ; 70,5% des articles l'ont été entre 2016 et 2018 (Figure 20). Les premiers *data papers* sont apparus en 2010 au sein de revues traditionnelles mais c'est entre 2013 et 2016 qu'on a vu leur nombre quasiment doubler chaque année avec une publication de plus en plus importante au sein des *data journals* nouvellement créés.

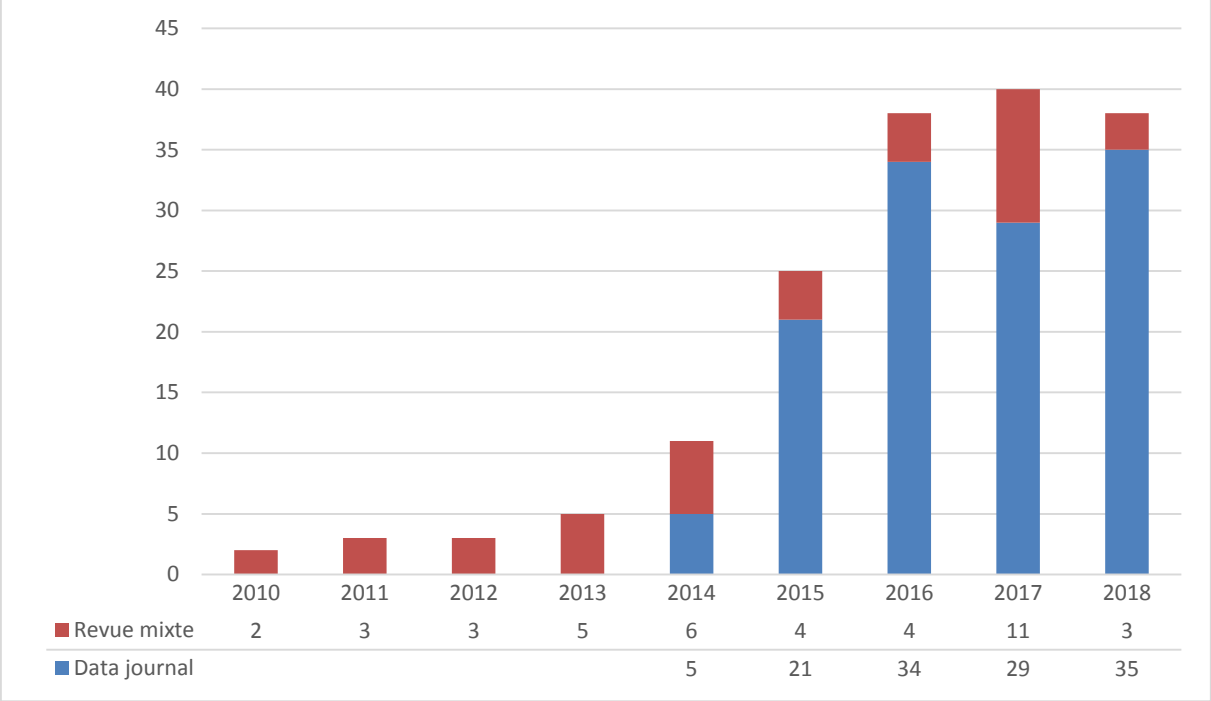


Figure 20 : Evolution de la publication des *data papers* en neuroscience

La majorité de ces articles (96,3%) est référencée au sein de l'une des trois grandes bases de données bibliographique : Pubmed (95,7%), Web of Science (90,2%), Scopus (95,7%). 6 articles ne sont référencés dans aucune de ces trois bases, 2 dans une seule base, 8 dans deux bases et 147 sont retrouvés dans les trois. Tous les articles retrouvés dans Web of Science le sont également dans les deux autres bases. La typologie du document établi par ces bases de référencement est

variable (Figure 21). Parmi les 156 articles référencés dans Pubmed, 89 n'ont pas de typologie établie, 44 ont une typologie multiple associant principalement les termes « dataset » et « research support » (41 articles) et 23 sont typés de façon unique (6 « dataset » et 15 « research support »). Pour Web of Science, les 147 articles référencés sont tous considérés comme des « articles » hormis 2 typés comme « review ». 101 articles (68,7%) sont également typés par la base comme « data paper ». Il est à noter que 15 articles publiés dans un *data journal* (2 dans *Data in brief* et 13 dans *Scientific Data*) n'ont pas été typés par la base comme *data paper*. Enfin, la base Scopus identifie les 156 articles qu'elle référence de façon unique en les typant principalement comme « articles » (94,2%). Aucun terme plus spécifique à la description de jeux de donnée n'est employé.

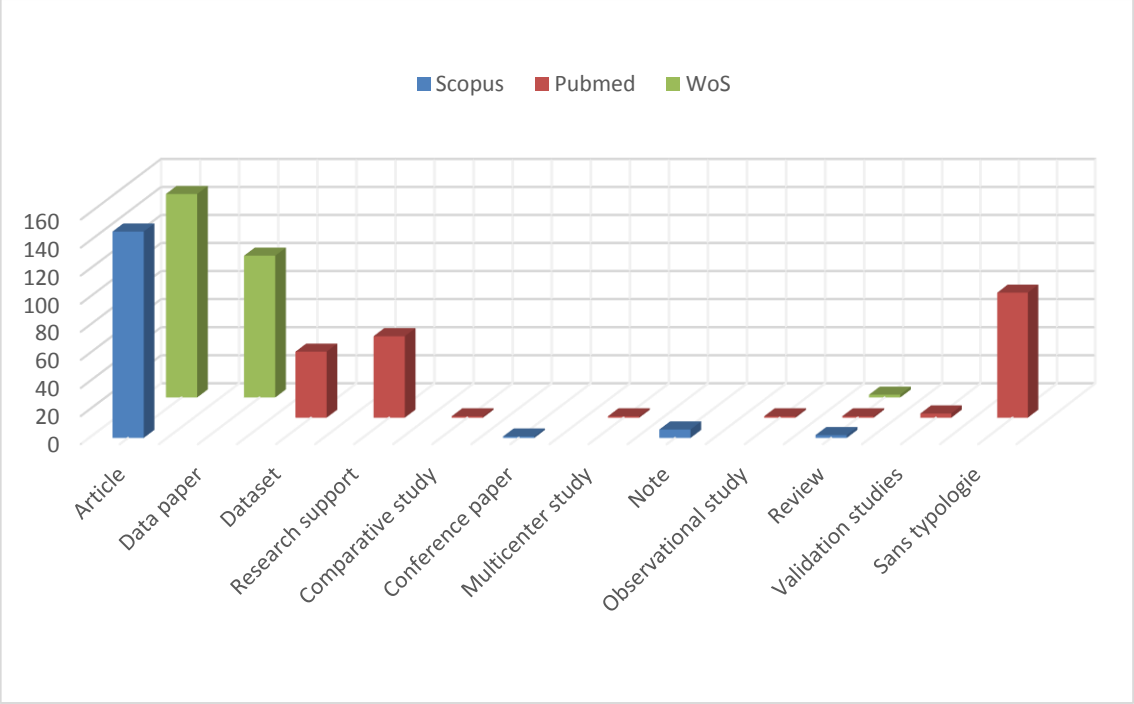


Figure 21 : Typologie du *data paper* dans les bases bibliographiques

En terme de structuration de l'article, il ne semble pas exister de standard avec des rubriques différentes en fonction des revues, certaines communes avec les articles classiques et d'autres plus spécifiques. La majorité des articles présente néanmoins une section spécifique quant à l'accessibilité des données décrites. En termes de taille, il s'agit d'article de taille standard avec 9,7 (+/- 5,9) pages en moyenne. 25% ont moins de 7 pages et 25% en ont plus de 11 avec un minimum de 3 pages et un maximum de 43 pages. Enfin, 75% des articles ont moins de 31 références bibliographiques avec un nombre de référence médian s'élevant à 14 (min 1 – max 135).

3.2 Equipes de recherche

Le nombre médian d'auteurs par article est de 6 (min 1 ; Q1 4 ; Q3 10 ; max 101) pour un nombre d'affiliation moyen de 5,1 (+/-5,7). Toutefois, la moitié des articles ont moins de 3 affiliations.

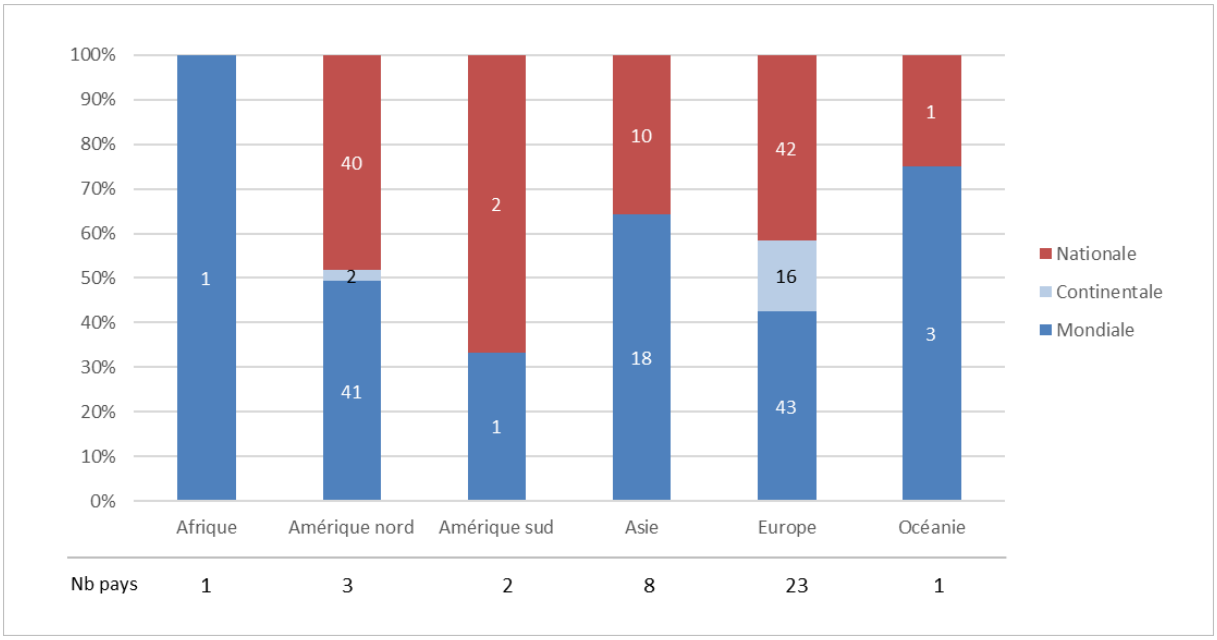


Figure 22 : Origine géographique des équipes de recherche

L'origine géographique des auteurs a été établie par le pays de rattachement de leur affiliation. Pour un chercheur rattaché à plusieurs unités provenant de nations différentes, chacune de ces nations ont été prises en compte. Ainsi, pour 58,3% des articles, l'équipe de recherche provenait d'un même pays et pour 41,7% il s'agissait d'une équipe internationale dont 11% issue d'un même continent et 30,7% de continents différents (jusqu'à 3 continents pour 7 articles). Parmi les 50 articles dont l'équipe à une origine mondiale, seul 1 article ne dispose pas de chercheurs provenant d'Europe ou d'Amérique du Nord. Bien que la majorité des équipes soient constituées de membres ayant une origine européenne ou nord-américaine, des chercheurs de tous les continents sont tout de même présents que ce soit au sein d'équipes nationales ou internationales (Figure 22).

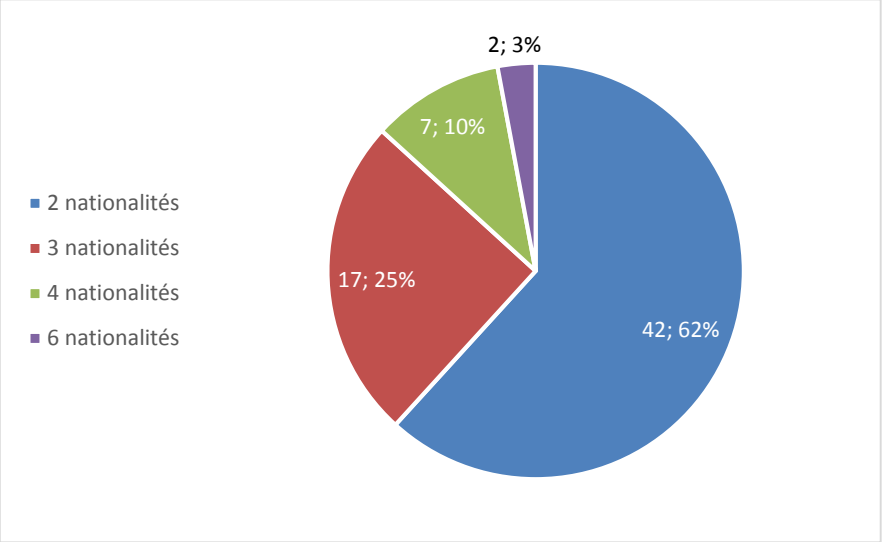
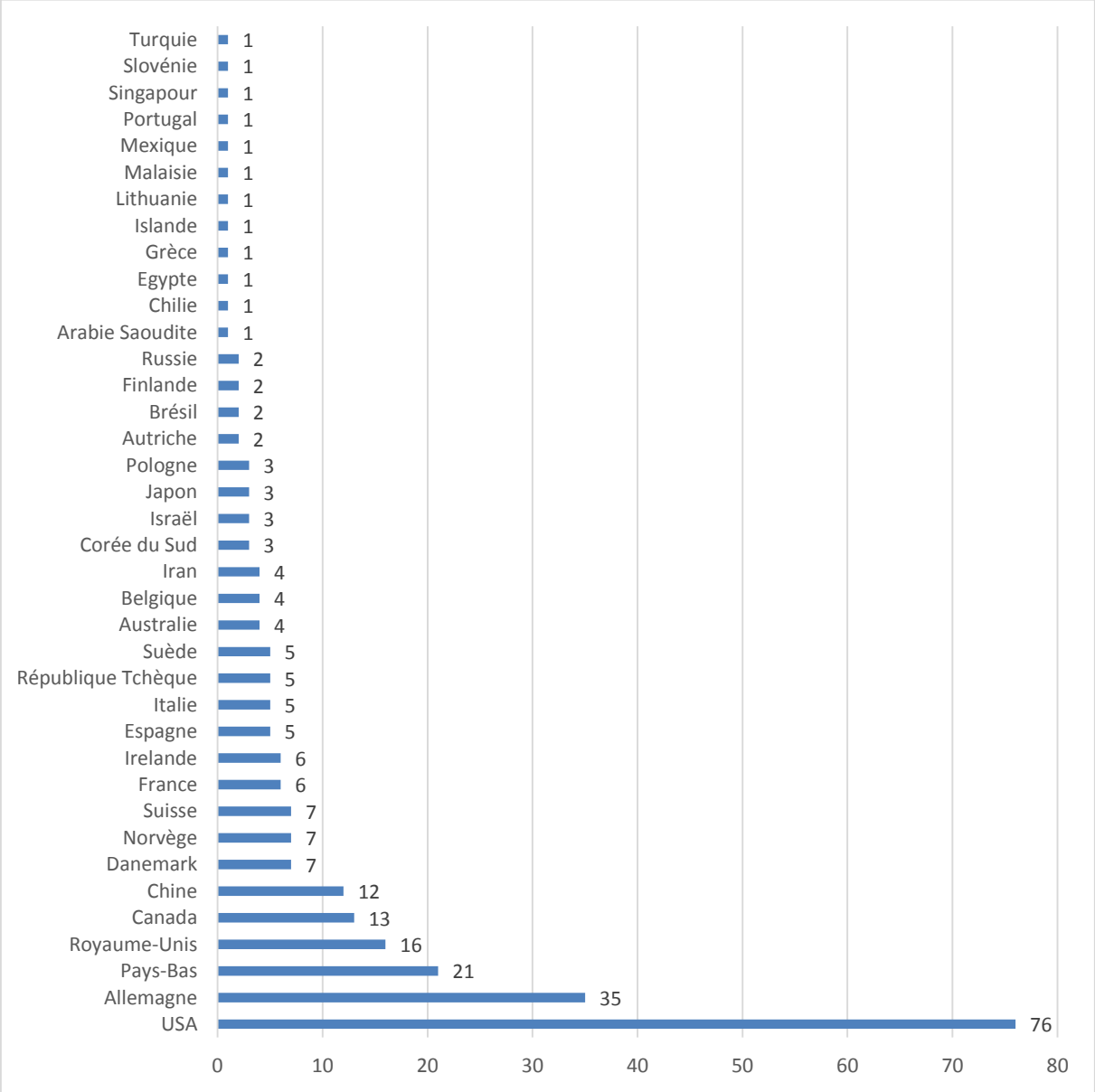


Figure 23 : Nombre d'articles avec des équipes internationales

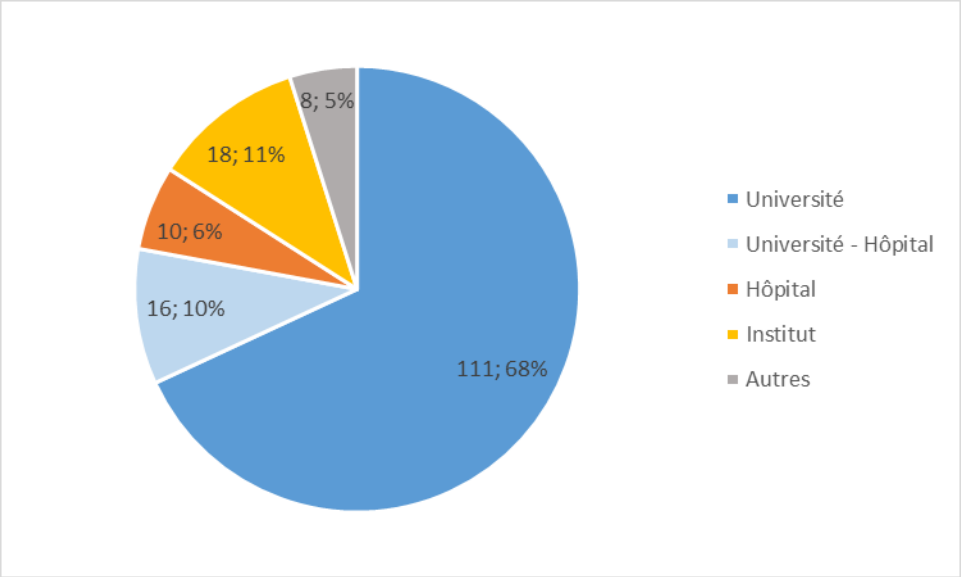
Un total de 38 nationalités a été retrouvé. Sur les 68 articles dont l'équipe est internationale, plus de la moitié (61,8%) ont des chercheurs provenant de 2 pays et un quart provenant de 3 pays (Figure 23).

La nationalité la plus représentée est les Etats-Unis avec 46,6% des articles (Figure 24). En Europe, la distribution est plus partagée avec comme principale nation l'Allemagne (21,5%), les Pays-Bas (12,9%) et le Royaume Unis (9,8%). A noter la Chine apparaissant comme la sixième nation avec une présence dans 7,4% des articles.



<u>Figure 24</u> : Nationalité des équipes de recherche

Les 163 articles ont été écrits par 152 auteurs différents (1er auteur) pour lesquels le H-index médian est de 8 (Q1 4,5 – Q3 17) avec un maximal retrouvé à 86. 2 auteurs ont écrit 3 de ces articles et 7 auteurs en ont écrit 2. 38,7% des articles ont leurs premiers auteurs affiliés à plusieurs unités de recherche. Si on considère uniquement leur première affiliation, 22% des projets ont été réalisés au sein d'un service non universitaire (Figure 25).



Autres : Laboratoire, Fondation, Ministère, Collège, Entreprise, Organisme de recherche

Figure 25 : Affiliation du premier auteur

Les chercheurs responsables du projet (dernier auteur) sont au nombre de 148, 10 ayant participé à plusieurs projets dont 1 à cinq projets. Ces derniers ont globalement une expérience plus marquée avec un H-index médian à 25 (min 1 ; Q1 17 ; Q3 39 ; max 119).

3.3 Contenu de l'article

L'ensemble des informations concernant les données décrites et leur accessibilité ont été obtenues par lecture intégrale de l'article. Pour 9 des articles, le texte intégral n'a pas pu être obtenu. Il n'y a donc pas d'informations détaillées sur les données de ces articles. Cette partie de notre étude a ainsi été menée sur 154 articles.

3.3.1 Description des données

Comme le montre la Figure 26, plus de la moitié (51,3%) des articles portent sur de l'imagerie et plus du quart (28,6%) sur une thématique clinique.

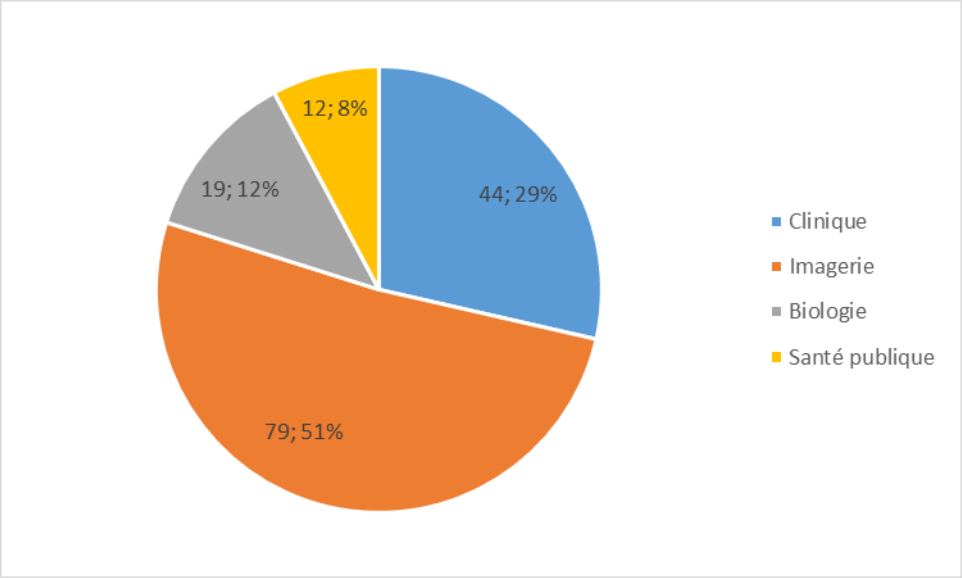


Figure 26 : Thématique des *data papers*

Pour les articles typés comme appartenant au champ de la santé publique, 6 portent sur la description de référentiel de données, 3 sur la création d'ontologie et 3 sur des études spécifiques tel que le coût de l'AVC, l'évaluation d'un *serious game* ou encore un algorithme de codage de la CIM-10. Pour les 19 articles de biologie, 12 d'entre eux étudient plus spécifiquement le protéome ou le métabolome.

Bien que le *data paper* soit défini comme un article décrivant un jeu de donnée, il a été retrouvé 7 articles pour lesquels l'objet du document était la description d'un outil et non un jeu de donnée (Tableau 2). Tous les autres articles décrivent la mise à disposition de données que ce soit dans le cadre d'une étude spécifique (142 articles) ou de projets plus globaux centralisant les données de recherches, les référentiels (5 articles). Pour les 142 études, 25 mettent à disposition des données répétées et 3 sont des données rétrospectives. Par ailleurs, 3 études correspondent à des essais contrôlés randomisés, 3 à des études interventionnelles et 5 à des cohortes ; pour le reste il s'agit d'études observationnelles.

Tableau 2 : Typologie des projets dont sont issus les données

Projet	N	%
ETUDE	**142**	**92,21%**
Longitudinale	23	
Transversale	114	
Transversale + Longitudinale	2	
Rétrospective	3	
OUTIL	**7**	**4,55%**
Logiciel	1	
Ontologie	3	
Questionnaire	3	
REFERENTIEL	**5**	**3,25%**

On retrouve principalement des données concernant des adultes mais des données sur des enfants sont également présentes (Figure 27). Pour 10 articles la population n'est pas précisée mais il semble qu'il s'agisse de population adulte. Parmi les 137 articles dont la population porte sur des individus, plus de la moitié (58,4%) porte sur des populations indemnes de pathologie (Figure 28). En terme d'effectif, on retrouve une médiane globale à 40 avec de grande disparité ; les articles typé santé

publique regroupant de grandes quantités d'individus alors que d'autres études ne portent que sur un seul individu (Tableau 3).

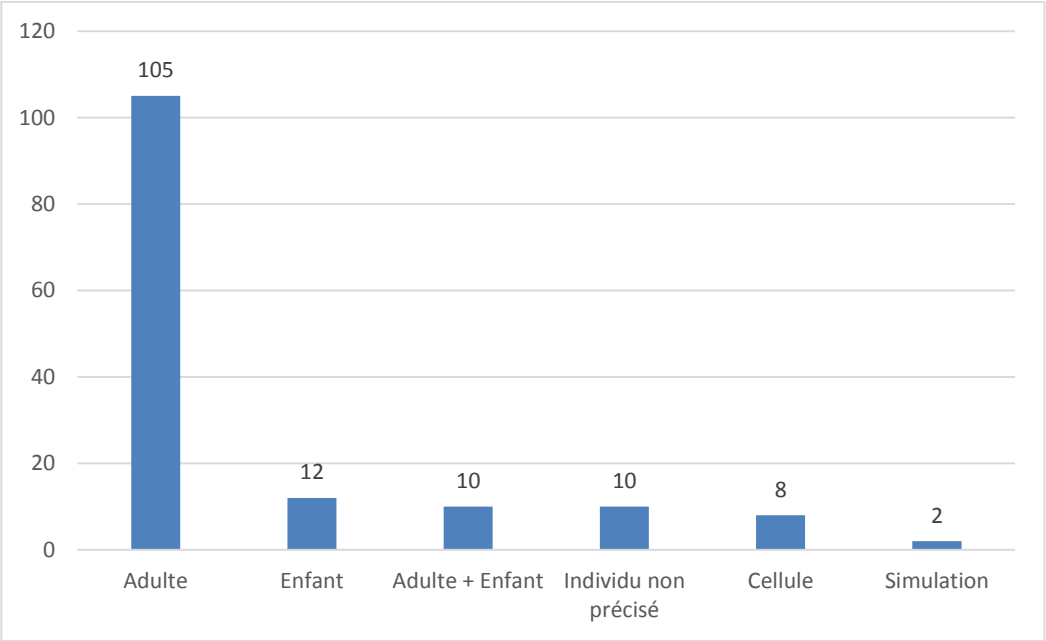


Figure 27 : Population des études

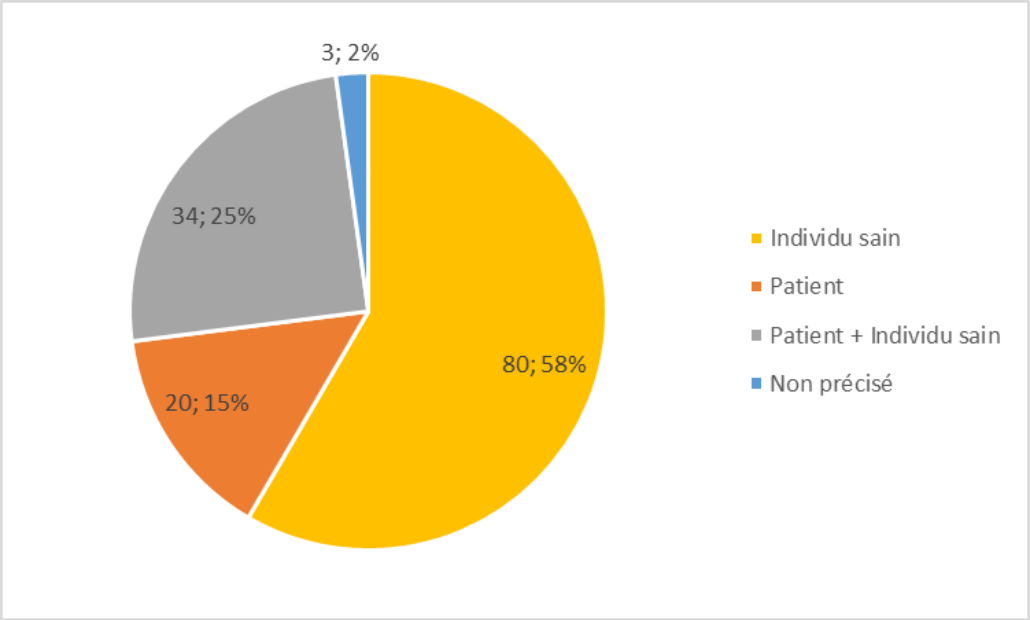


Figure 28 : Etat pathologique des individus

Tableau 3 : Effectifs des populations selon la thématique de l'article

	Moyenne	DS	Médiane	Q1	Q3	Minimum	Maximum	N
Total	10450	105337	40	18,2	188	1	1216547	134
Clinique	3131	15189	38	20	183	5	96958	41
Imagerie	175	435	37	15	137	1	2790	73
Biologie	139	216	40	26	136,5	1	730	12
Santé publique	157188	428131	1417,5	1095	14444	38	1216547	8

DS : Déviation standard

La population d'étude a été obtenue de façon indirecte autrement dit il n'y a pas eu de contact entre les chercheurs et les sujets pour 24 des articles. Dans certains cas, il s'agit déjà d'une réutilisation de données. Pour les 5 référentiels, il s'agit de projets de grande envergure dont l'objectif est de regrouper les données issues de plusieurs projets de recherche. La provenance des populations d'études pour ces 24 articles est décrite dans la Figure 29.

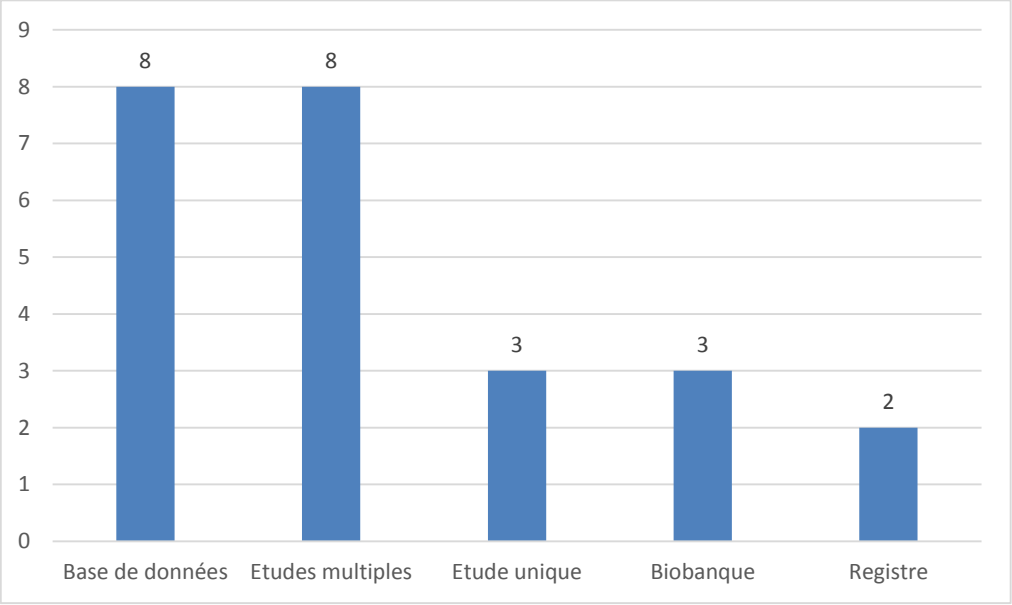
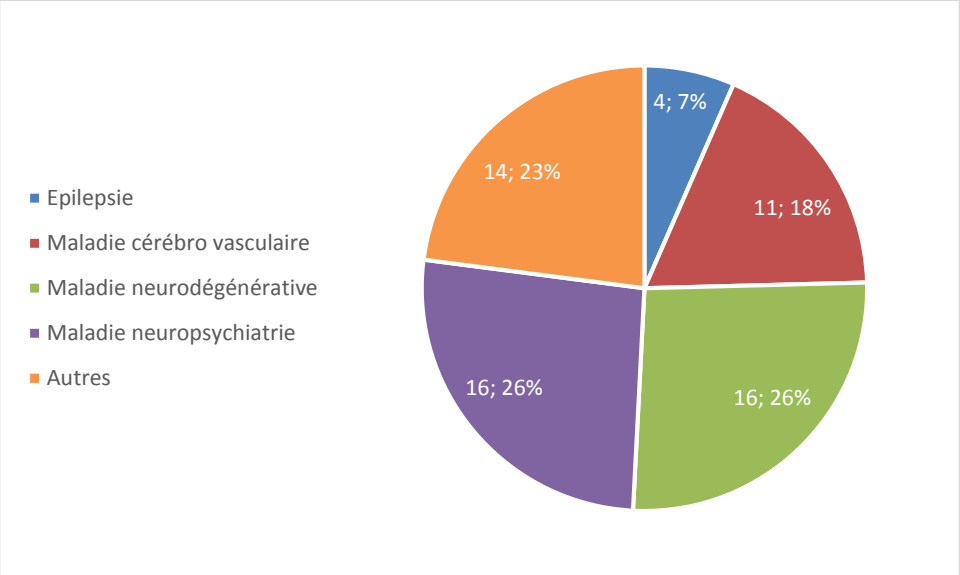


Figure 29 : Provenance des populations des projets à inclusion indirecte

Seul 39,6% des articles étudient une pathologie spécifique. Parmi ceux-ci on retrouve principalement les maladies neuropsychiatriques (26,2%) dont les troubles du spectre autistique et les troubles du développement font partie (16,4%), et les maladies neurodégénérative (26,2%) (Figure 30).



Autres : prématurité, grossesse, atteinte lysosomal, VIH, acouphène, sommeil

Figure 30 : Maladies étudiées

Les données décrites dans les articles varient en fonction de la thématique de ces derniers (Tableau 4). Dans 31% des cas, des données complémentaires sont également disponibles. Pour ce qui est des articles cliniques, plus de la moitié (52,3%) correspondent à des données d'électrophysiologie dont la plupart sont des données d'EEG pouvant être associées à d'autres données (dans 10 cas) notamment des données EOG et/ou EMG (Figure 31). Dans certains cas, les données mises à disposition correspondent aux tracés et dans d'autres aux paramètres de mesures.

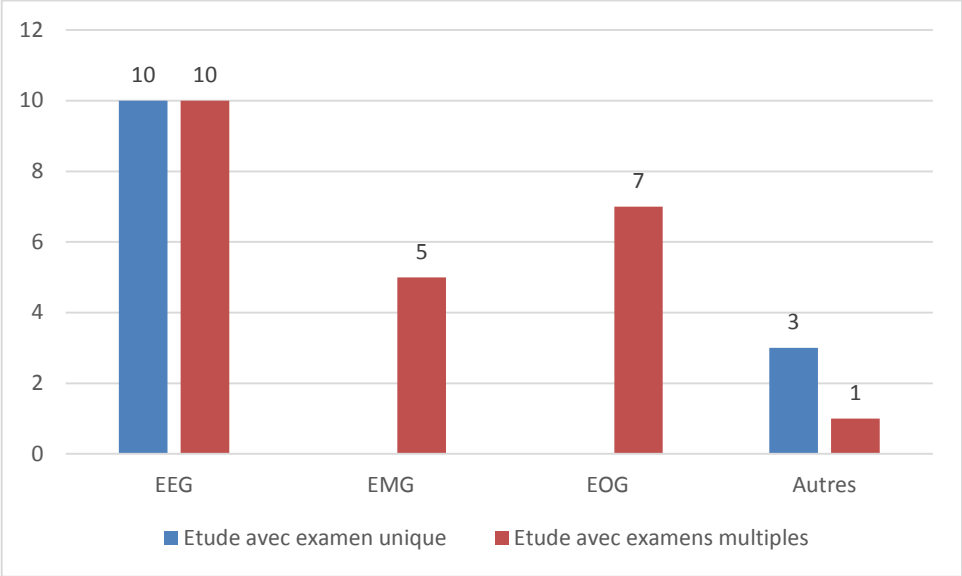
Tableau 4 : Typologie des données selon la thématique de l'article

	Total N=154		Clinique N=44		Imagerie N=79		Biologie N=19		Santé publique N=12	
	N	%	N	%	N	%	N	%	N	%
Pathologie étudiée										
Non	93	60,4%	22	50,0%	59	74,7%	7	36,8%	5	41,7%
Oui	61	39,6%	22	50,0%	20	25,3%	12	63,2%	7	58,3%
Données principales										
Autres	8	5,2%	0	0,0%	1	1,3%	0	0,0%	7	58,3%
Biologie protéomique	9	5,8%	0	0,0%	0	0,0%	8	42,1%	1	8,3%
Biologie laboratoire	12	7,8%	1	2,3%	0	0,0%	11	57,9%	0	0,0%
Clinico-comportemental	21	13,6%	20	45,5%	0	0,0%	0	0,0%	1	8,3%
Electrophysiologie	23	14,9%	23	52,3%	0	0,0%	0	0,0%	0	0,0%
Imagerie autre	1	0,6%	0	0,0%	1	1,3%	0	0,0%	0	0,0%
Imagerie IRM	74	48,1%	0	0,0%	71	89,9%	0	0,0%	3	25,0%
Imagerie multimodale	6	3,9%	0	0,0%	6	7,6%	0	0,0%	0	0,0%
Données complémentaires										
Non	106	68,8%	27	61,4%	55	69,6%	15	78,9%	9	75,0%
Oui	48	31,2%	17	38,6%	24	30,4%	4	21,1%	3	25,0%

Autres : ontologie, logiciel, coût, codes cim10, simulation de paramètres d'imagerie
Imagerie multimodale : IRM couplé à un examen électrophysiologique
Imagerie autre : TEP

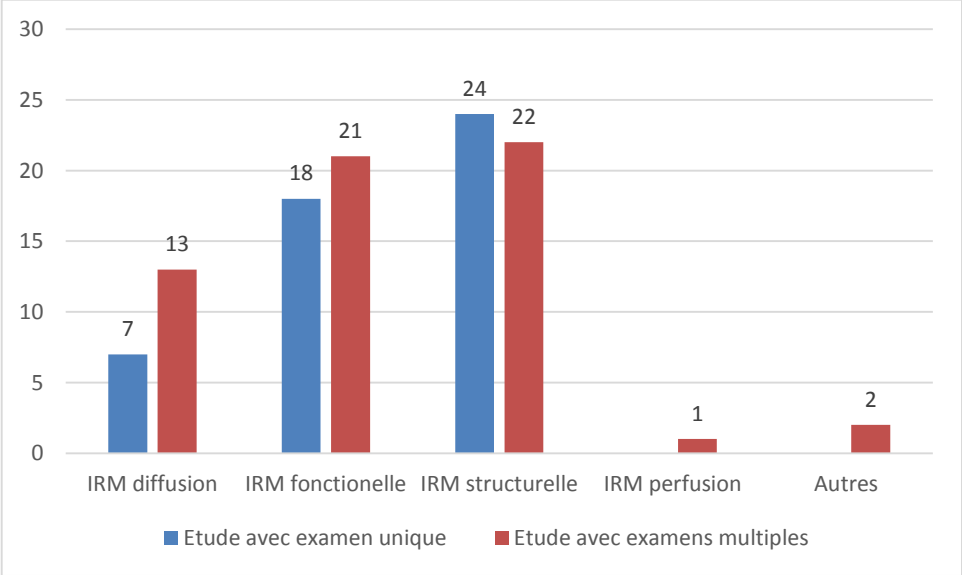
Près de 90% des données d'imagerie porte uniquement sur des données d'IRM. Dans 22 cas plusieurs types d'IRM ont été réalisées au cours de l'étude (Figure 32). Il s'agit des IRM structurelles et fonctionnelles qui sont le plus représentées. Pour 6 des articles, on retrouve une imagerie multimodale c'est-à-dire que l'IRM (principalement structurelle et/ou fonctionnelle) est couplée à un examen d'électrophysiologie (EEG et/ou magnétoencéphalographie).

Enfin, la période d'acquisition des données est peu présente. Elle est retrouvée dans seulement 14,3% des articles. Pour 10 des articles (6 santé publique, 3 cliniques et 1 biologie), la base de données décrite est une base dynamique c'est-à-dire que de nouvelles données peuvent alimenter la base ; ainsi l'article décrit la base à un instant t.



Autres : spectroscopie infrarouge, électrocorticographie, électrophysiologie intrathalamique, examen non précisé

Figure 31 : Description des données électrophysiologiques



Autres : angiographie, spectroscopie

Figure 32 : Description des données IRM

3.3.2 Accessibilité des données

La Figure 33 montre le format des données mises à disposition. Pour 22,1% des articles, les données décrites correspondent à des données résultats telles qu'on peut les retrouver dans un article classique, il s'agit en fait de compléments à l'article principal. Pour 17 des articles, les auteurs mettent à disposition d'autres types de données tel que les codes utilisés pour leur recherche mais pas de données individuelles ; les données sources étant pour certaines déjà partagées en *open access*.

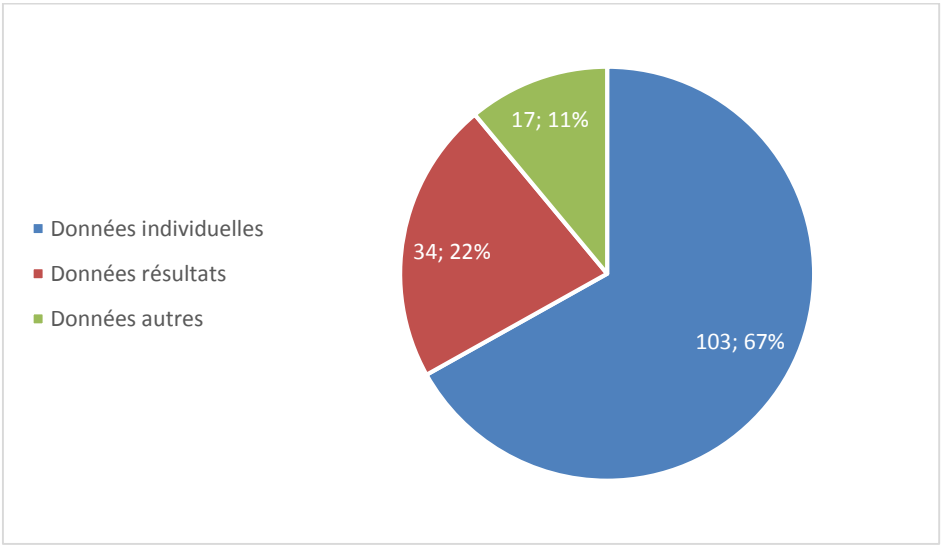


Figure 33 : Format des données

Sur les 103 articles mettant à disposition des données individuelles, 17 (16,5%) rentrent dans le cadre d'un projet de partage de données à grande échelle : 3 via *ProteomeXchange* et 14 via l'*International Neuroimaging Data-Sharing Initiative*. La majorité des jeux de données (87,4%) est accessible via un accès unique, pour les 13 autres un double accès est disponible dont 1 sans accès via un entrepôt de données (uniquement via des sites web). Les lieux d'accès aux données sont indiqués sur la Figure 34.

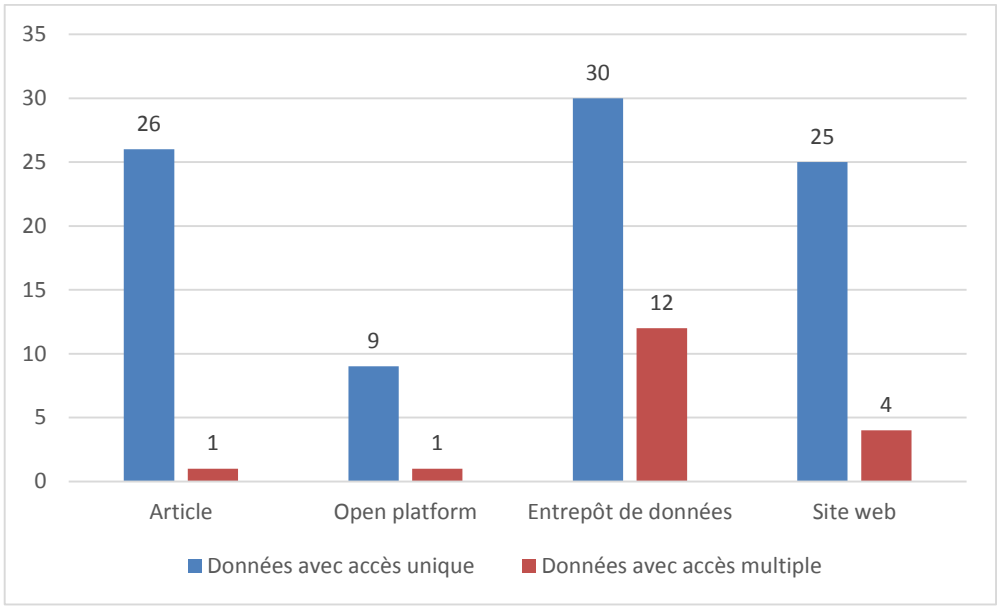
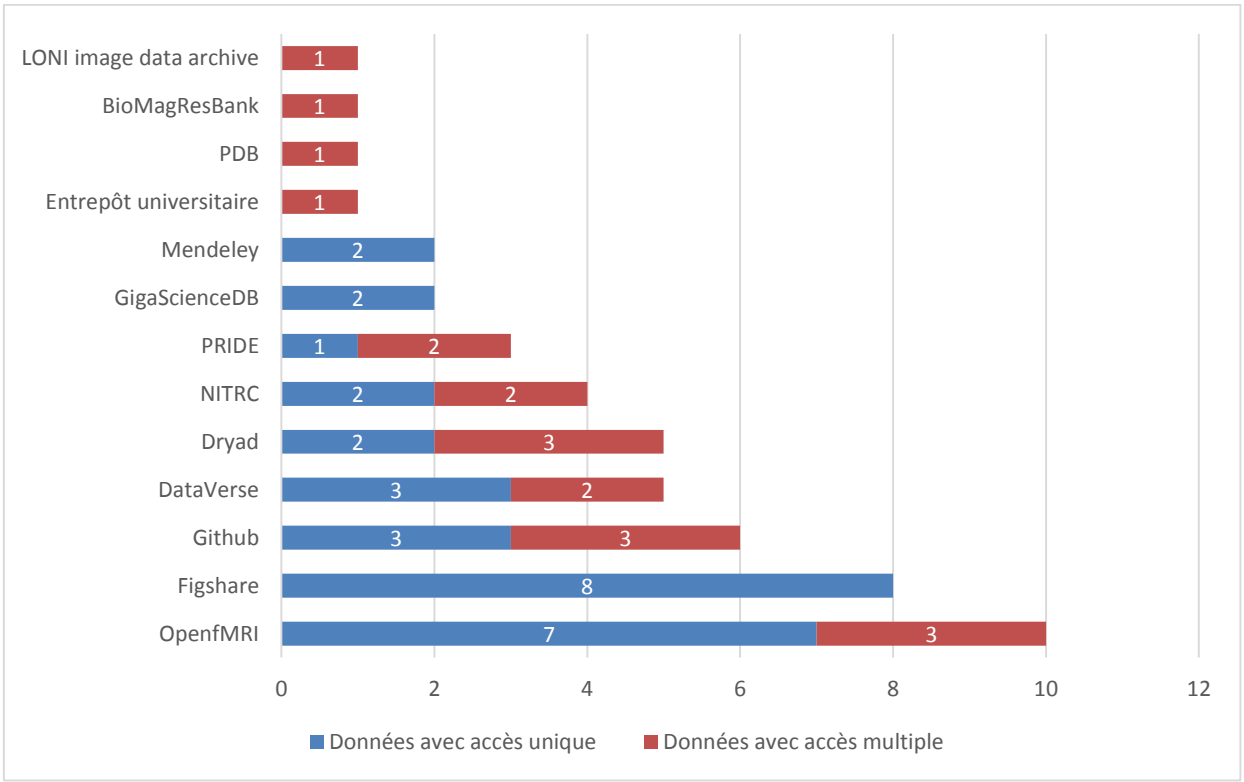


Figure 34 : Lieux d'accès aux données



PDB : Protein Data Bank
PRIDE : PRotemics IDEntifications Database
NITRC : Neuroimaging Informatics Tools and Resources Clearinghouse

Figure 35 : Entrepôts de données

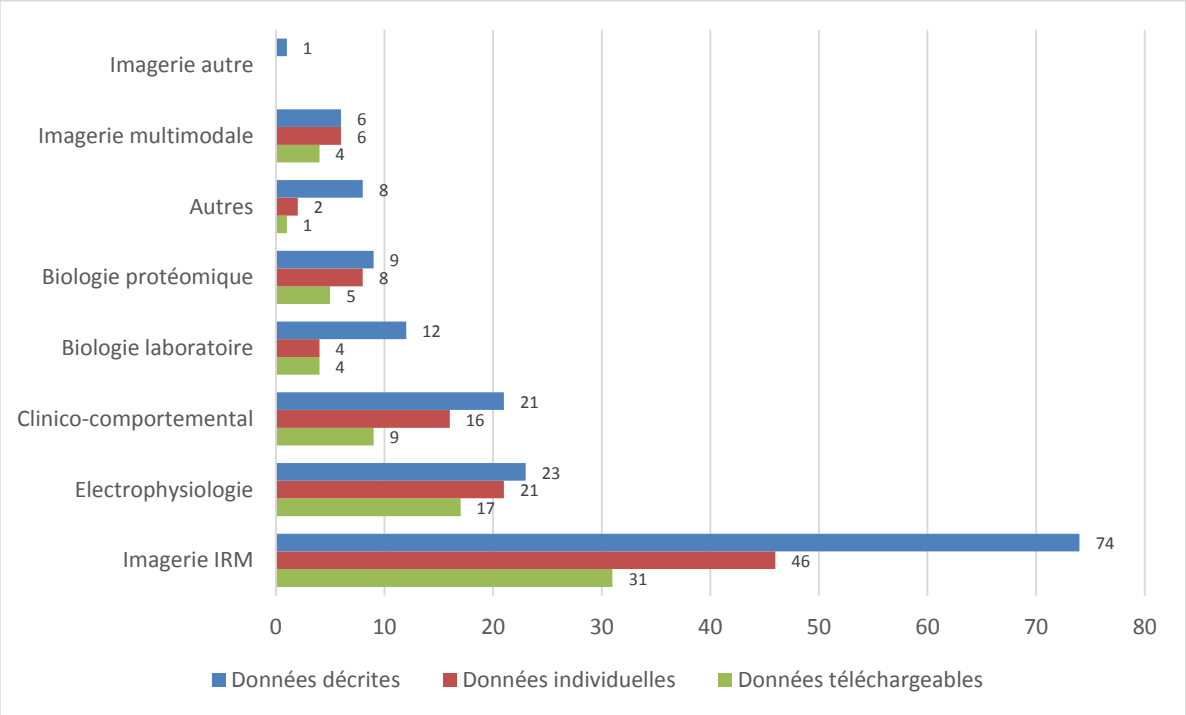
L'accès aux données au sein de l'article correspond à 3 situations : accès direct au sein même de l'article à travers des tableaux par exemple ou à l'aide d'un fichier complémentaire annexe, et accès indirect avec indication de la procédure de demande auprès de l'auteur. 40,7% des données ont été stockées au sein d'un référentiel dont on retrouve la répartition sur la Figure 35. C'est au sein du référentiel spécifique *OpenfMR*I que l'on retrouve le plus de données puis on retrouve les référentiels généraux tels que *Figshare, Github, Dryad* et *DataVerse*. A noter, 2 jeux de données disponibles au sein d'un référentiel propre à une revue (*GigaScienceDB*). Pour les 29 jeux de données disponibles à partir d'un site web (dont 1 disponible via 2 sites), 14 le sont via le site du projet de partage de l'INDI et 16 via un site spécifique. Enfin, les plateformes ouvertes mettant à disposition les données de 10 articles sont les suivantes : *Synapse Sage Bionetworks* (4 jeux), *Open Science Framework* (3 jeux) *COINS Data Exchanges* (2 jeux) et *XNAT* (1 jeux).

Pour finir, le téléchargement des données individuelles est possible dans 68,9% des cas. Pour les 32 articles dont les données ne sont pas téléchargeables, cette indisponibilité des données est dans certains cas relative du fait de la nécéssité de créer un compte pour y accéder. Pour d'autres en revanche une demande auprès des auteurs et organisme responsable est indispensable. Les informations concernant l'accessibilité des données en fonction de la thématique de l'article sont fournies dans le Tableau 5. Les Figure 36 et Figure 37 montrent respectivement le rapport entre les données décrites, l'accès aux données individuelles et leur téléchargement selon la typologie des données et l'étude d'une pathologie.

Tableau 5 : Accessibilité des données selon la thématique de l'article

	Total N=154		Clinique N=44		Imagerie N=79		Biologie N=19		Santé publique N=12	
	N	%	N	%	N	%	N	%	N	%
Format des données										
Données individuelles	103	66,9%	36	81,8%	50	63,3%	11	57,9%	6	50,0%
Données résultats	34	22,1%	5	11,4%	18	22,8%	8	42,1%	3	25,0%
Données autres	17	11,0%	3	6,8%	11	13,9%	0	0,0%	3	25,0%
Partage de données										
Partage isolé	86	83,5%	34	94,4%	41	82,0%	8	72,7%	3	50,0%
Projet de partage	17	16,5%	2	5,6%	9	18,0%	3	27,3%	3	50,0%
NA	51	/	8	/	29	/	8	/	6	/
Accès										
Multiple	13	12,6%	2	5,6%	8	16,0%	3	27,3%	0	0,0%
Unique	90	87,4%	34	94,4%	42	84,0%	8	72,7%	6	100,0%
NA	51	/	8	/	29	/	8	/	6	/
Téléchargement										
Non	32	31,1%	10	27,8%	14	28,0%	2	18,2%	6	100,0%
Oui	71	68,9%	26	72,2%	36	72,0%	9	81,8%	0	0,0%
NA	51	/	8	/	29	/	8	/	6	/

NA : Non applicable



Autres : ontologie, logiciel, coût, codes cim10, simulation de paramètres d'imagerie
Imagerie multimodale : IRM couplé à un examen électrophysiologique
Imagerie autre : TEP

Figure 36 : Accessibilité des données selon la typologie des données

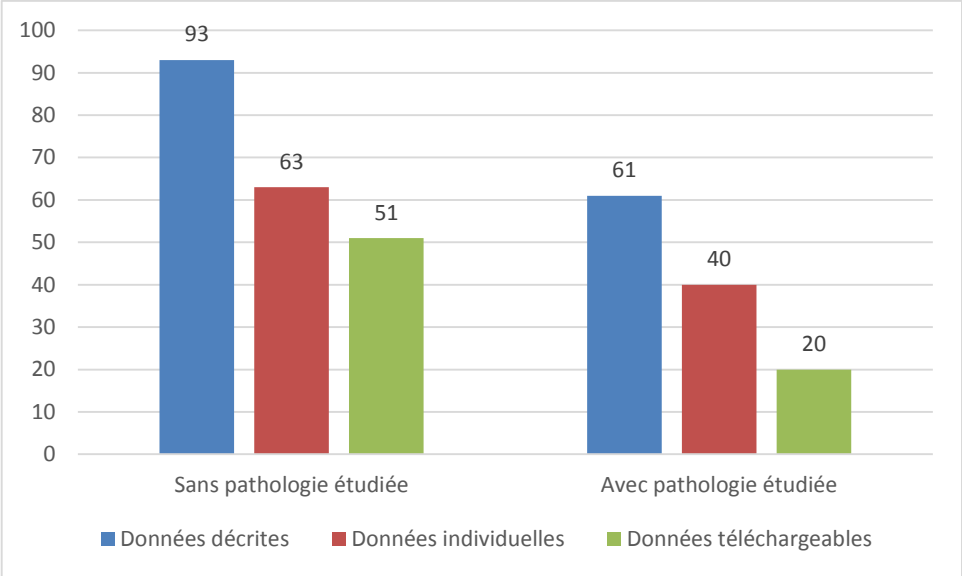


Figure 37 : Accessibilité des données selon l'étude d'une pathologie

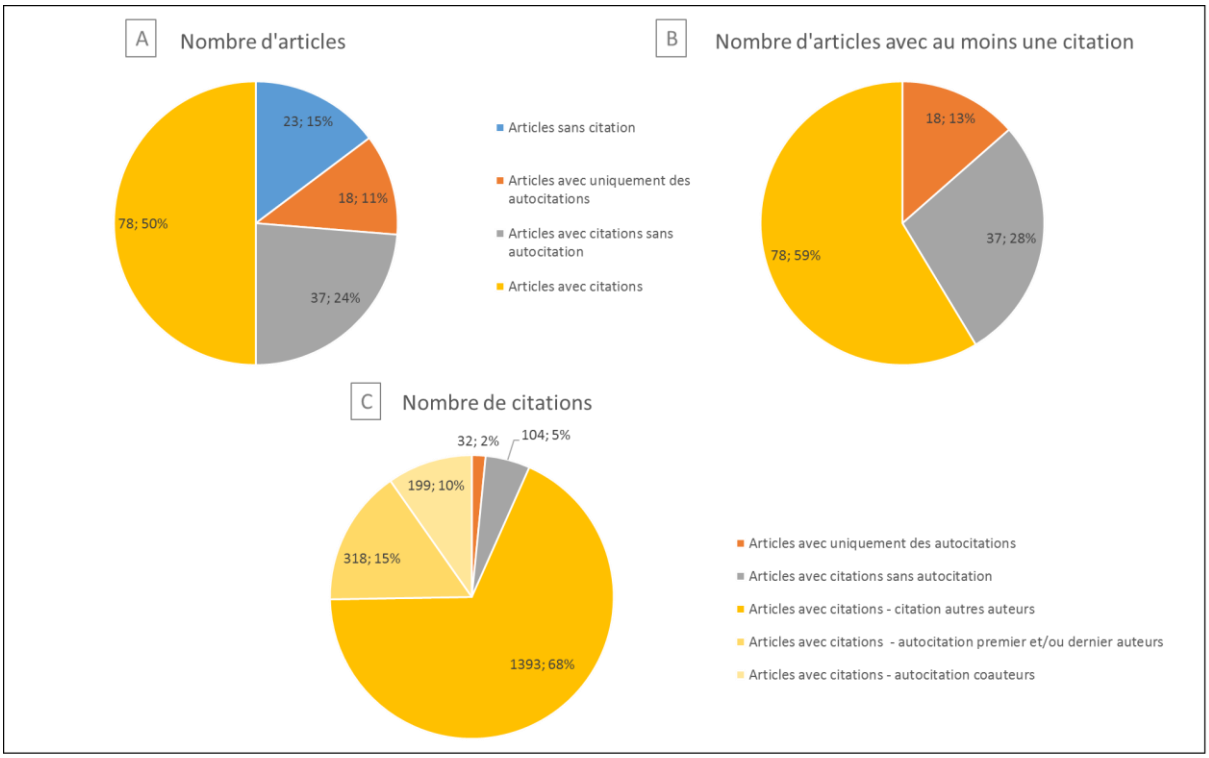
3.4 Impact

L'ensemble de la recherche concernant la bibliométrie a été réalisée via la base de données Scopus dans laquelle 7 articles n'ont pas été retrouvés indexés portant ainsi l'analyse des citations et des *altmetrics* sur 156 articles. Sur les 7 articles exclus, 6 sont publiés dans un *data journal*.

3.4.1 Etude des citations

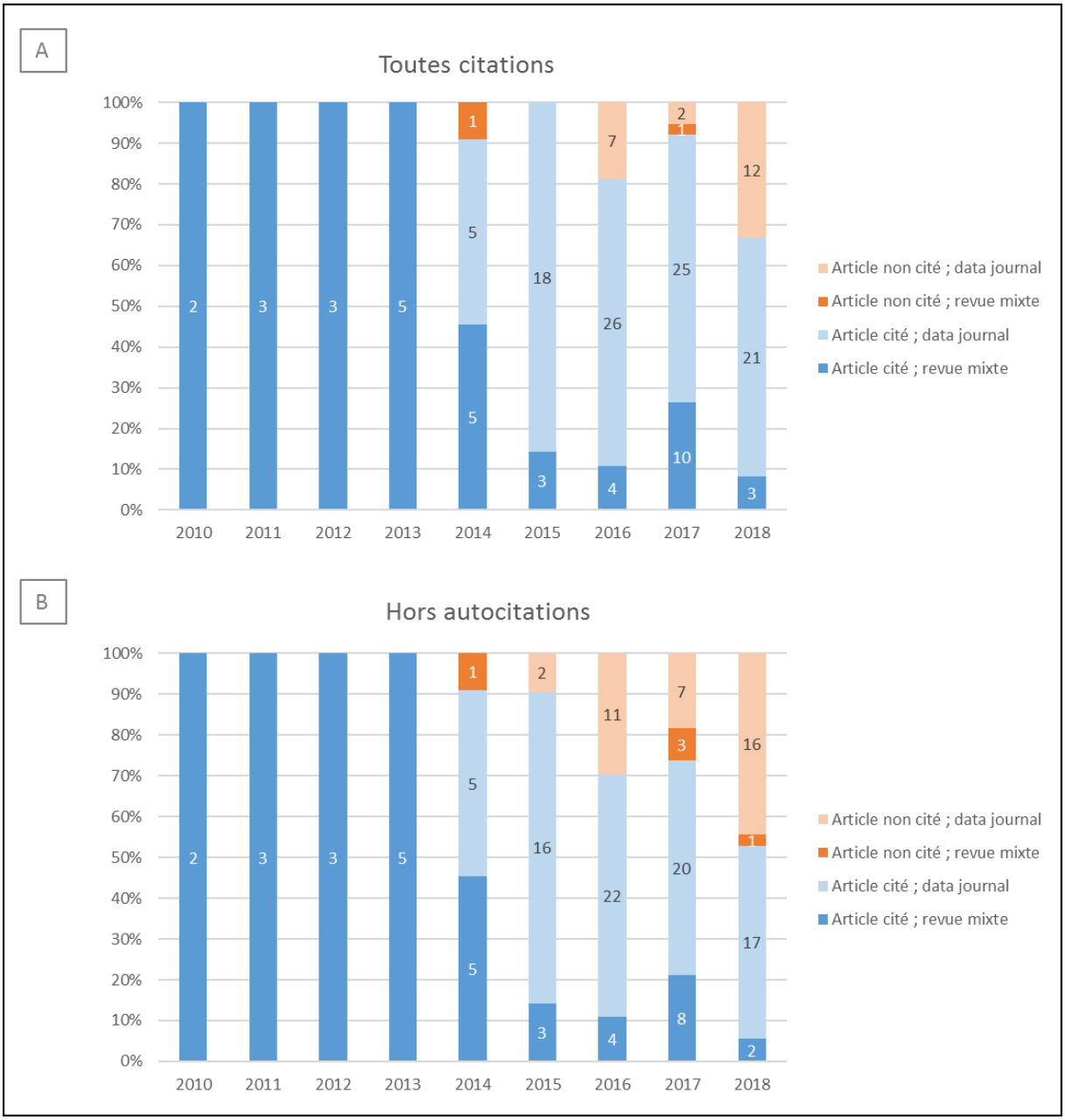
85,3% des articles ont été cités au moins une fois. Toutefois, en ne prenant pas en compte les autocitations, ce nombre est porté à 73,7%. 37 articles pour lesquels il n'y a aucune autocitation (Figure 38 A et B). Comme le montre la Figure 39 le fait qu'un article soit cité au moins une fois est plus faible pour les articles publiés récemment (année 2017 - 2018). Cette différence est significative lorsque les autocitations sont exclues (p=0,010 vs p=0,10). De même, les articles publiés au sein d'un *data journal*

sont moins cités que les articles publiés dans une revue mixte (hors autocitations p=0,037 ; avec autocitations p=0,079). Il n'existe pas de différence entre les articles portant sur des données cliniques et ceux portant sur des données d'imagerie (p=0,37).



<u>Figure 38</u> : *Data papers* et citations

Ces articles ont été cités 2046 fois soit une médiane de 3 citations (min 0 ; Q1 1 ; Q3 13 ; max 520). Pour prendre en compte le temps depuis publication nous avons calculé pour chaque article le rapport entre le nombre de citations reçu et le nombre d'années depuis publication dont la moyenne s'élève à 2,7 citations année (DS 6,8 ; min 0 ; Q1 0,3 ; mediane 1 ; Q3 2,8 ; max 74,3). A noter, 26,8% des citations concernent une autocitation par l'un des auteurs dont 17% une autocitation du premier ou dernier auteur (Figure 38 C).



<u>Figure 39</u> : Nombre de *data papers* cités selon l'année de publication

En excluant l'article publié dans *Molecular psychiatry* pour lequel le nombre de citations est fort important (520 citations), les trois journaux qui ont publié le plus de *data papers*, *Data in Brief*, *Scientific Data* et *Neuroinformatics*, ont également reçu le plus de citations avec un total de 1261 citations (Tableau 6). Bien que le nombre d'articles publiés au sein d'une revue mixte soit inférieur à celui des *data journals*, le nombre de citations reçu est cependant statistiquement plus important (moyenne

respectivement égale à 26,1 et 8,6 ; p <0,001). En prenant en compte le nombre d'années il existe toujours une différence statistique même si celle-ci semble moins importante (moyenne respective égale à 4,1 et 2,2 citation année ; p=0,013). 80,4% des citations des *data journals* ont été reçus par *Scientific Data* alors que ce journal a publié près de deux fois moins d'article par rapport à *Data in brief* ; une différence significative a ainsi été retrouvée entre ces deux journaux (p<0,001).

Tableau 6 : Nombre de citations selon le type de revue

Revue	N article	Article avec citation			Article avec citation hors autocitation			NA
		N article	N citation	Citation /année	N article	N citation	Citation /année	
Revue mixte	**40**	**38**	**1044**	**4,37**	**35**	**769**	**4,71**	**1**
Neuroinformatics	9	9	259	4,18	9	182	4,18	
Proteomics	7	7	73	2,06	6	42	2,34	
F1000 Research	5	4	8	0,75	3	4	0,83	1
GigaScience	5	4	55	3,30	4	36	3,30	
Spinal cord	3	3	53	3,05	2	34	4,45	
BMC Neurology	2	2	4	0,54	2	2	0,54	
Database	2	2	5	0,55	2	4	0,55	
Journal of Biomedical Semantics	2	2	23	1,74	2	17	1,74	
BMC Health Services Research	1	1	20	4,00	1	17	4,00	
BMC Neuroscience	1	1	5	1,67	1	1	1,67	
BMC Research Notes	1	1	13	1,44	1	12	1,44	
MIC and CCAI	1	1	6	1,50	1	2	1,50	
Molecular Psychiatry	1	1	520	74,29	1	416	74,29	
Data journal	**116**	**95**	**1002**	**2,65**	**80**	**728**	**3,02**	**6**
Data in brief	75	55	196	1,03	41	124	1,16	
Scientific Data	41	40	806	4,89	39	604	4,97	
Biomedical Data Journal	0	/	/	/	/	/	/	1
Journal of Open Psychology Data	0	/	/	/	/	/	/	3
Open Journal of Bioresources	0	/	/	/	/	/	/	2
Total	**156**	**133**	**2046**	**3,14**	**115**	**1497**	**3,53**	**7**

NA : Non applicable car article non référencé dans Scopus

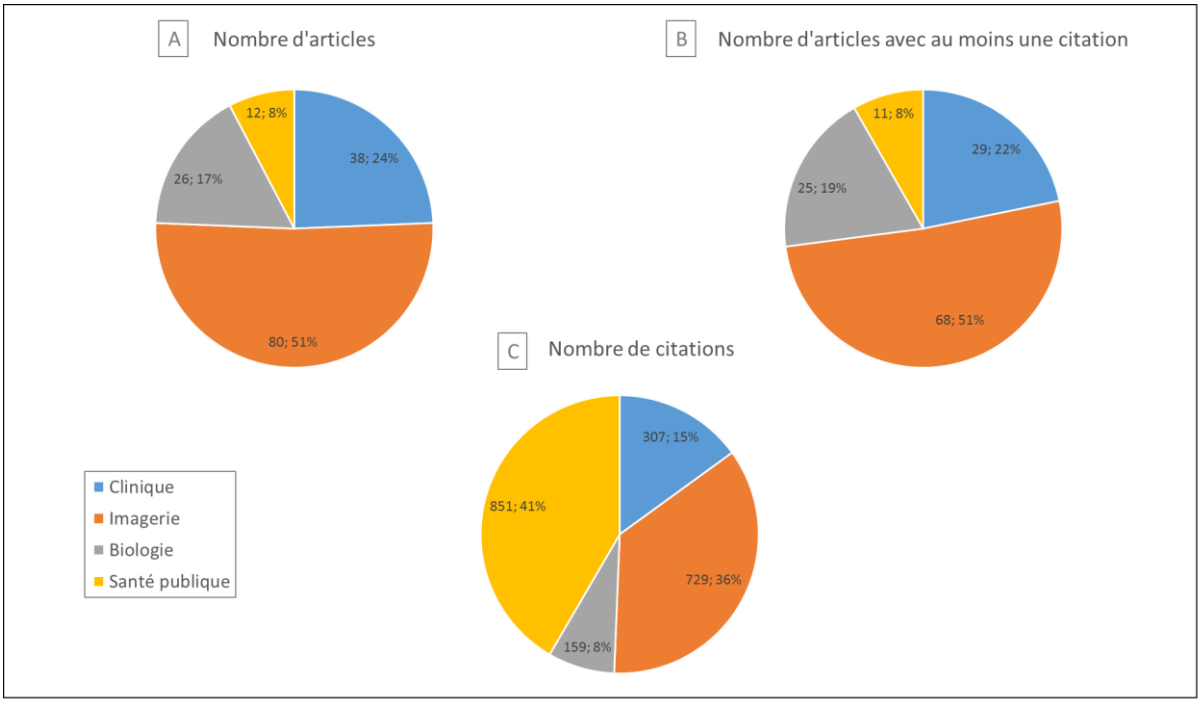
Comme il a été vu précédemment, le nombre de *data papers* publié augmente chaque année notamment au sein des *data journals* (Figure 20). De la même manière, il existe une augmentation du nombre de citations annuelles concernant ces articles pouvant témoigner d'une certaine « popularité croissante » de ce nouveau type de communication scientifique (Tableau 7). A noter que le nombre de citations pour l'année 2019 est incomplet puisque l'étude s'est terminée au 19 juillet 2019. Il n'existe pas de différence statistiquement significative concernant le nombre de citations par année entre les articles publiés récemment (2017 – 2018) et ceux publiés avant 2017 (moyenne respective égale à 1,8 et 3,5 citation/année ; p=0,24) alors que cette différence existe si on ne prend pas en considération le nombre d'années depuis la publication de l'article (moyenne respective égale à 4,9 et 20,6 citations ; p <0,001).

Tableau 7 : Evolution du nombre de citations

Année de citation	Revue mixte			Data journal			Total		
	N citation	N article citable	N article cité	N citation	N article citable	N article cité	N citation	N article citable	N article cité
2010	0	2	0	0	0	0	0	2	0
2011	4	5	3	0	0	0	4	5	3
2012	14	8	5	0	0	0	14	8	5
2013	28	14	9	0	0	0	28	14	9
2014	71	19	15	4	5	3	75	24	18
2015	144	22	18	40	23	11	184	45	29
2016	166	27	20	121	56	27	287	83	47
2017	232	37	27	218	83	47	450	120	74
2018	224	40	26	356	116	69	580	156	95
2019	161	40	29	263	116	62	424	156	91

41% des citations concernent les articles typés santé publique avec une médiane à 16,5 citations (min 0 ; Q1 6,8 ; Q3 60,5 ; max 520) alors qu'ils ne représentent que 8% des articles (Figure 40). Cette thématique regroupe notamment

des articles décrivant des référentiels qui correspondent généralement à des projets à grande échelle pouvant regrouper plusieurs types de données provenant de plusieurs études, ce qui peut expliquer un nombre de citations élevé. Il n'a pas été retrouvé de différence significative entre le nombre de citations et la typologie des données, imagerie vs clinique, avec respectivement une médiane à 3 et 2 (p=0,19).



<u>Figure 40</u> : *Data papers* et citations selon la thématique de l'article

Par ailleurs, des corrélations significatives ont été retrouvées entre le nombre de citations et le nombre d'auteurs (r=0,29 ; p<0,001), le nombre de pays (r=0,23 ; p=0,004) et le nombre d'affiliations (r=0,26 ; p=0,001). La force de ces liaisons reste néanmoins faible. Au vue des faibles niveaux de corrélation, il n'a pas été mené d'analyse multivariée.

Si on regarde en termes de délai de citation en prenant en compte uniquement la première citation de chaque article, le délai médian est de 7,9 mois (IC 95% 6,9 –

10,3). En excluant les autocitations des premiers et derniers auteurs, le délai médian de citation s'allonge en passant à 12 mois (IC 95% 10,6 – 16,7) (Figure 41). Les taux de citations à 3, 6, 12 et 18 mois sont indiqués dans Tableau 8.

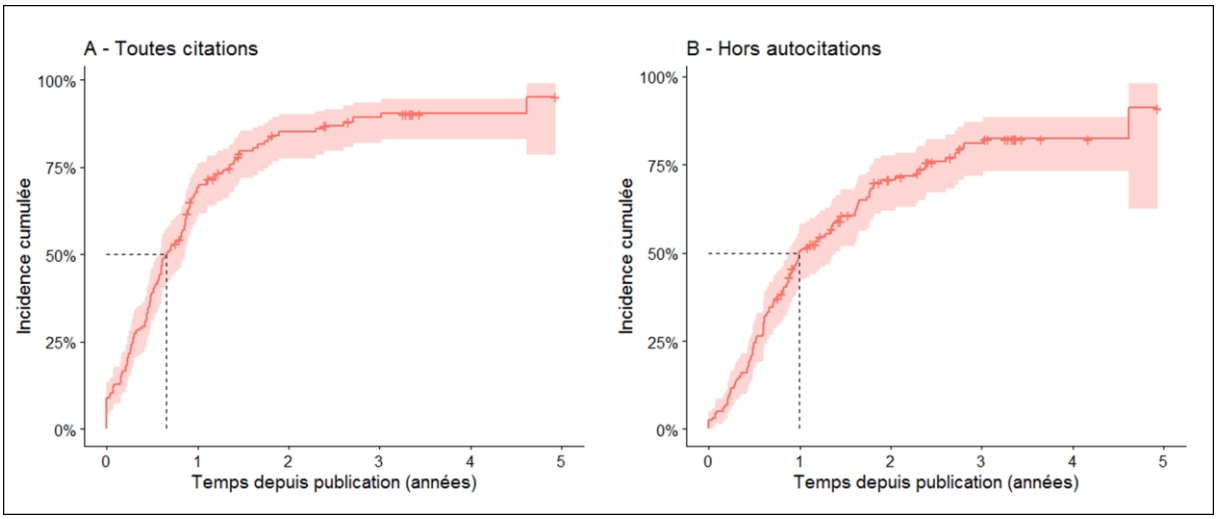


Figure 41 : Incidence cumulée de citation

Tableau 8 : Taux de citation

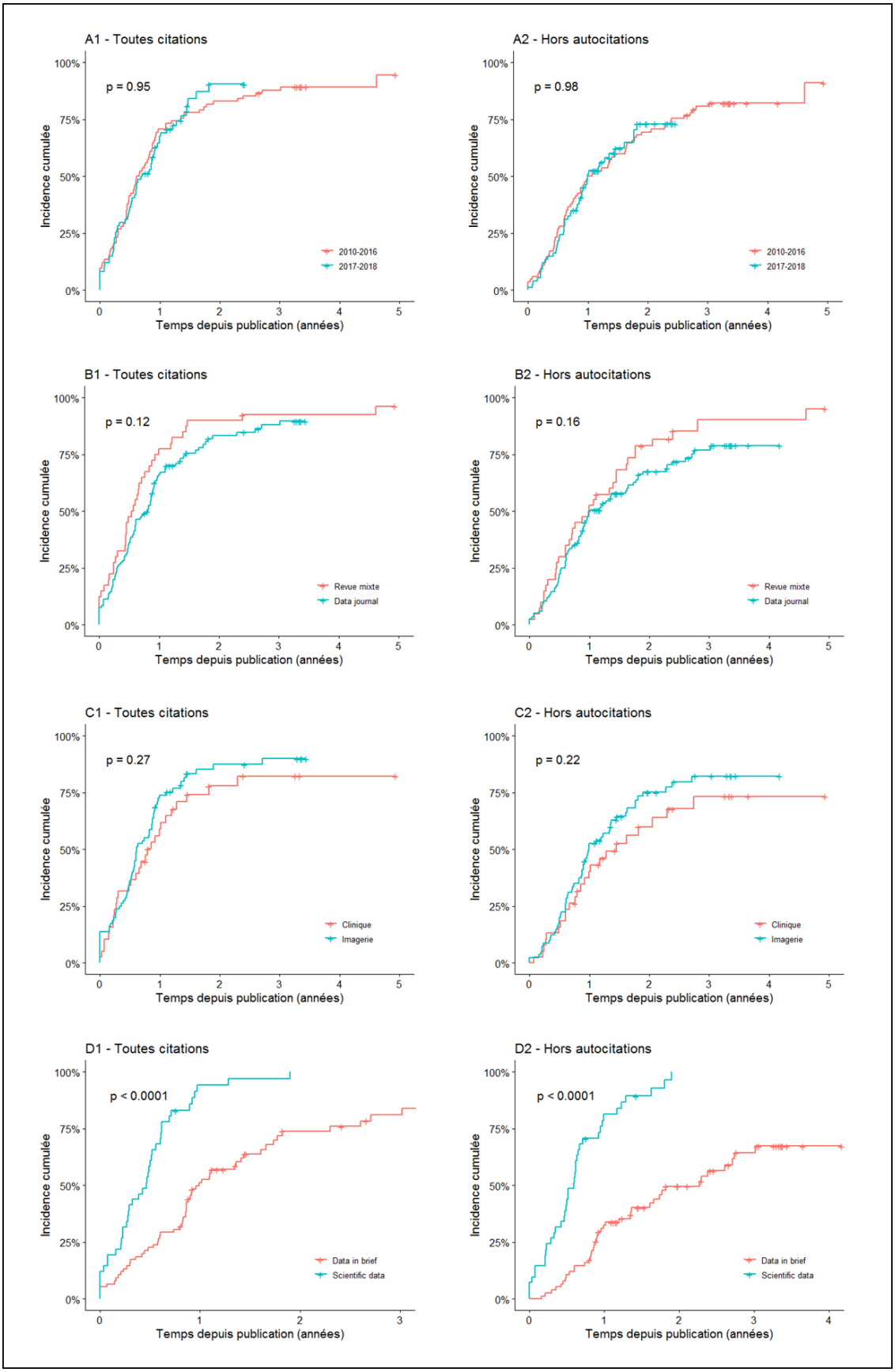
	Temps	N à risque	Taux de citations	IC 95%
Toutes citations	3 mois	123	21,8%	15% - 28%
	6 mois	96	39,1%	31% - 46%
	12 mois	46	69,2%	61% - 76%
	18 mois	24	79,8%	72% - 85%
Hors autocitations	3 mois	139	11,5%	6% - 16%
	6 mois	119	24,4%	17% - 31%
	12 mois	75	50,4%	42% - 58%
	18 mois	51	60,7%	52% - 68%

Il n'existe pas de différence significative entre le délai de citation des articles publiés avant 2017 et ceux publiés après, médiane respectivement de 7,6 mois IC95% (5,8 – 10,4) et 8,4 mois IC95% (6,41 – 11,5) (p=0,95) (Figure 42 A1). Le délai médian de citation des articles publiés dans une revue mixte est de 6,7 mois IC95% (5,29 –

10,1) contre 9,5 mois IC95% (7,16 – 11) pour ceux publiés dans un *data journal* mais cette différence n'est pas significative (p=0,12) (Figure 42 B1). De même, aucune différence significative n'a été retrouvée (p=0,27) entre les articles avec des données cliniques (médiane 9,5 mois IC95% (6,2 – 15,4)) par rapport à ceux avec des données d'imagerie (médiane 7,3 mois IC95% (6,7 – 10,6)) (Figure 42 C1). L'exclusion des autocitations montre un allongement des délais de citation pour chacun des groupes mais sans différence statistique retrouvée (Tableau 9 et Figure 42 A2, B2, C2). En revanche, la comparaison du délai de citation entre les deux *data journals* ; *Scientific data* et *Data in brief*, a montré une différence significative (p<0,001) dont les médianes respectives sont de 5,6 mois IC95% (3,4 – 6,7) et de 11,9 mois IC95% (10,4 – 17,2). Cette différence est plus importante passant respectivement à 7,1 mois IC95% (5,65 – 8,02) et 27,3 mois IC95% (16,4 – 36,2) lorsqu'on exclut les autocitations des premiers et derniers auteurs (Figure 42 D1, D2).

Tableau 9 : Délai médian de citation hors autocitations en fonction des groupes

Groupes	N	Evénement	Médiane	IC 95%
Année publication				
2010 - 2016	82	68	12,4	9,5 - 19,5
2017 - 2018	74	47	12	10,35 - 19,3
Typologie journal				
Revue mixte	40	35	11,9	7,95 - 19,5
Data journal	116	80	12,2	10,8 - 19,9
Thématique				
Clinique	38	24	17,4	11,1 - NA
Imagerie	80	59	11,8	10,6 - 16,4
Journal				
Data in brief	75	41	27,3	16,4 - 36,2
Scientific Data	41	39	7,13	5,6 - 8,0



A Année de publication ; B Typologie du journal ; C Thématique ; D Data journal

<u>Figure 42</u> : Incidence cumulée de citation en fonction des groupes

Il paraissait intéressant de décrire le cadre de ces 2046 citations. 69% des citations correspondent à des articles, 15% à des revues, 10% à des conférences et 6% correspondent à d'autres types de documents tels que des chapitres, des livres, des éditoriaux, des notes, de courtes études ou encore des lettres. Ceci laisse supposer que certains *data papers* sont cités sans qu'il y ai une réutilisation réelle des données. Par ailleurs, ces citations ont été rattaché à 26 domaines d'étude (Figure 43). Il n'est pas étonnant de voir que 51% des citations concerne le champ des neurosciences et 40% celui de la médecine mais il est intéressant de remarquer que certaines citations font partie d'un domaine d'étude différent de celui de la santé.

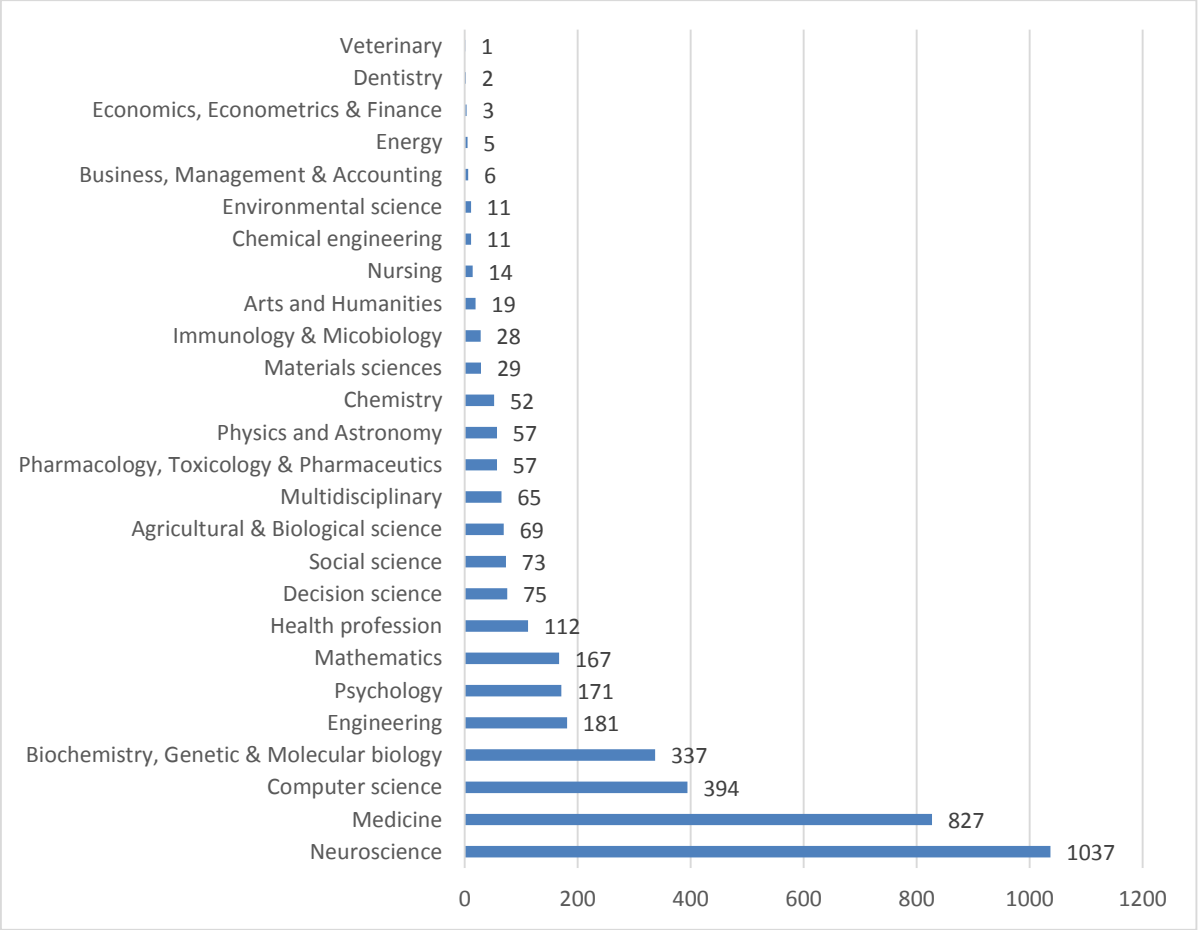


Figure 43 : Domaine d'étude des citations

3.4.2 Etude des altmetrics

Pour rappel, les scores *altmetrics* ont également été obtenus à partir de la base de données Scopus qui fournit les scores de 32 indicateurs issus de *PlumX Metrics* (Annexe 2). Ces scores permettent d'évaluer l'attractivité en ligne des articles. Comme pour l'analyse des citations, l'analyse des scores *altmetrics* a porté sur 156 articles.

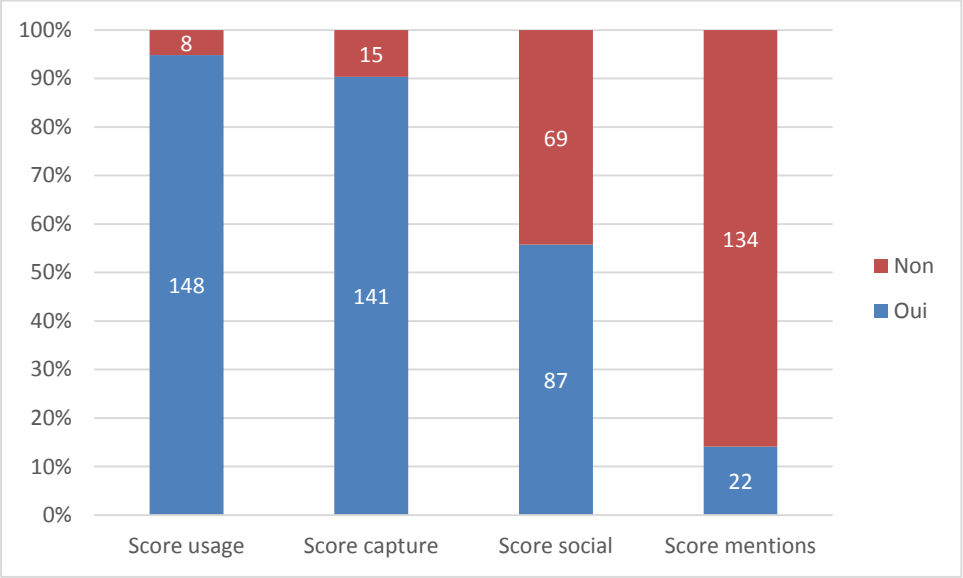


Figure 44 : Couverture des scores *altmetrics*

Un score *altmetric* a été retrouvé pour 98,1% des articles mais avec des variations pour chacun des quatres éléments constituant le score (Figure 44). Sur les 32 indicateurs proposés par *PlumX Metrics*, seul 14 d'entre eux ont présenté un résultat. Hormis la vue des *abstracts*, Twitter et les sites de *bookmarking* sociaux sont les indicateurs avec la plus grande couverture, respectivement 53% et 86% (Tableau 10). Lorsqu'il est partagé, le *data paper* l'est en moyenne 11,1 fois sur Twitter (DS 22,2 ; médiane 3 ; min 1 ; max 129 ; Q1 1 ; Q3 7,5), 26,6 fois sur Facebook (DS 42,7 ; médiane 11 ; min 2 ; max 187 ; Q1 2 ; Q3 22) et 43,8 fois sur les sites de *bookmarking* sociaux (DS 65,5 ; médiane 25 ; min 1 ; max 617 ; Q1 10 ; Q3 56,5). Ces résultats

indiquent que le *data paper* a une certaine visibilité sur les médias sociaux pouvant témoigner de l'intérêt quant aux partages et à la discussion d'un jeu de données mis à disposition. Néanmoins, comme le montrent les résultats du Tableau 10 et du Tableau 11, il existe une grande variabilité pour chacun des indicateurs *altmetrics*.

Tableau 10 : Couverture, densité et intensité des scores *altmetrics*

Indicateur	Couverture		Densité		Intensité	
	N	%	Moyenne	DS	Moyenne	DS
Score total	**153**	**98,1%**	**105,1**	**263,7**	**107,1**	**265,8**
Score usage	**148**	**94,9%**	**52,4**	**193,1**	**55,2**	**197,8**
Abstract views	148	94,9%	37,3	141,8	39,4	145,3
Clicks	16	10,3%	2,9	16,2	28,4	43,9
Downloads	6	3,8%	1,3	8,2	32,7	28,9
Full text views	12	7,7%	8,2	47	106,2	140,5
Links outs	79	50,6%	2,7	8	5,4	10,6
Score capture	**141**	**90,4%**	**42,2**	**74,8**	**46,7**	**77,3**
Exports/Saves	60	38,5%	4,3	17,2	11,3	26,4
Readers	135	86,5%	37,9	62,7	43,8	65,5
Score social	**87**	**55,8%**	**10,2**	**28,2**	**18,3**	**35,9**
Facebook	25	16,0%	4,3	19,4	26,6	42,7
Twitters	83	53,2%	5,9	17,1	11,1	22,2
Score mentions	**22**	**14,1%**	**0,2**	**0,8**	**1,7**	**1,4**
Blog mentions	5	3,2%	0	0,2	1,2	0,4
Comments	2	1,3%	0,1	0,6	4	4,2
News mentions	12	7,7%	0,1	0,4	1,2	0,6
Q&A site mentions	3	1,9%	0	0,1	1	0
References	5	3,2%	0	0,2	1	0

Couverture : proportion d'article avec un évènement
Densité : moyenne sur l'ensemble des articles
Intensité : moyenne sur les articles avec au moins un évènement
DS : Déviation standard

Le Tableau 12 montre qu'il n'existe pas de différence statistiquement significative entre le score total *altmetric* et le type de données (clinique vs imagerie) (p=0,64), une différence à la limite de la significativité pour la typologie du journal (p=0,055) et l'ancienneté de l'article (p=0,07) mais qui n'existe plus après correction des p-valeurs, et une différence significative entre les deux *data journals* (p<0,01). Les

articles publiés dans *Scientific Data* semblent avoir une attractivité sur les médias sociaux plus importante que ceux publiés dans *Data in brief* alors que le score d'usage reflétant le nombre de vues et de téléchargements ne semble pas différent. La typologie du journal, l'année de publication et la thématique de l'article ne semblent pas avoir un impact sur l'intérêt social. Des travaux sont encore nécessaires pour comprendre avec plus de précision ces nouveaux indicateurs, ces résultats sont donc à interpréter avec précaution.

Tableau 11 : Paramètres descriptifs des scores *altmetrics*

		Minimum	Q1	Médiane	Q3	Maximum
Score total	*Densité*	0	18	39.0	97.0	2983
	Intensité	1	18.0	40.0	100.0	2983
Score usage	*Densité*	0	3	10.0	27.5	2157
	Intensité	1	4.8	10.0	31.2	2157
Score capture	*Densité*	0	7	21.0	53.5	789
	Intensité	1	10.0	26.0	59.0	789
Score social	*Densité*	0	0	1.0	5.0	217
	Intensité	1	1.0	4.0	12.5	217
Score mention	*Densité*	0	0	0.0	0.0	7
	Intensité	1	1.0	1.0	2.0	7

Couverture : proportion d'article avec un évènement
Densité : moyenne sur l'ensemble des articles
Intensité : moyenne sur les articles avec au moins un évènement

Tableau 12 : Scores *altmetrics* en fonction des groupes

Score	Typologie journal		Année publication		Thème		Journal	
	Revue mixte	Data journal	2010-2016	2017-2018	Clinique	Imagerie	Data in brief	Scientific data
	N=40	N=116	N=82	N=74	N=38	N=80	N=75	N=41
Total								
Moyenne	219,4	65,6	153,5	51,4	117,2	69,2	29,7	131,2
DS	487,7	83,2	354,3	52,8	183,9	75,1	25,7	109
Médiane	63	34,5	56,5	31	36,5	39,5	23	96
p¹	0,055		0,07		0,64		<0,001*	
p²	0,22		0,25		1,00		<0,001*	
Usage								
Moyenne	157,3	16,2	85,5	15,7	69,2	19,5	12,1	23,9
DS	361,8	26,5	261,8	22	140,8	38,9	15,5	38,5
Médiane	26,5	8	12	8	12	8	8	11
p¹	<0,001*		0,049*		0,034*		0,10	
p²	0,0027*		0,25		0,17		0,10	
Capture								
Moyenne	54,3	38,1	53,1	30,2	41,8	36,1	15,9	78,6
DS	125,8	45,9	97,6	32,1	54,1	38,8	16,7	54,2
Médiane	24,5	19,5	26,5	18	19	23	11	70
p¹	0,743		0,54		1		<0,001*	
p²	1		1,00		1		<0,001*	
Social								
Moyenne	7,7	11	14,6	5,3	6,1	13,2	1,6	28,3
DS	18,3	30,9	35	16,8	21,4	33,7	3,2	47,5
Médiane	1,5	1	1	1	0	1	0	5
p¹	0,429		0,06		0,10		<0,001*	
p²	1		0,25		0,40		<0,001*	
Mentions								
Moyenne	0,2	0,3	0,3	0,2	0,1	0,3	0,1	0,5
DS	0,4	0,9	0,9	0,6	0,5	1	0,8	1
Médiane	0	0	0	0	0	0	0	0
p¹	0,965		0,83		0,18		<0,001*	
p²	1		1		0,53		<0,001*	

DS : Déviation standard
¹ p-valeur des tests t
² p-valeur corrigée par la méthode de Holm
** significativité statistique*

Discussion

Notre travail avait pour but d'étudier l'intérêt de l'utilisation des *data papers* dans la communauté médicale. Bien qu'il s'agisse d'un élément nouveau de la littérature scientifique, le nombre de *data papers* issus du domaine médical ne cesse d'augmenter avec, selon nos critères de recherche, 745 articles publiés dont 70% entre 2016 et 2018. On retrouve la même tendance pour les 163 *data papers* de neuroscience. Cette évolution témoigne d'un attrait grandissant pour ce nouveau type de publication. En parallèle, ces articles présentent une certaine attractivité sur les médias sociaux et le nombre de citations annuelles faisant référence à ces articles est en augmentation, confirmant cet intérêt naissant. Cependant, nos résultats montrent une hétérogénéité sur le format de ce nouveau document avec des différences sur leur dénomination, leur référencement, leur contenu et l'accès aux données décrites. Cette hétérogénéité confirme qu'il s'agit d'une pratique récente pour laquelle des travaux d'harmonisation de la part de l'ensemble des acteurs sont nécessaires afin d'avoir une utilisation optimale.

1 Une nouvelle forme de communication scientifique

Dans la littérature, nous n'avons retrouvé que très peu d'études s'intéressant spécifiquement au *data papers* et aucune, à notre connaissance, n'a été menée sur un domaine particulier. En ce sens, notre travail est original.

Le domaine médical apparait comme une discipline s'intéressant à la publication de *data papers*. En effet, on retrouve une majorité de revues publiant ce type d'article rattachées au domaine de la santé (128,143,144). Toutefois, comme le signalait déjà

Candela (144), même si les revues promeuvent la publication de *data papers*, peu d'entre elles en publient. Dans son étude, seules 52% des revues, toutes disciplines confondues, ont publié au moins un *data paper* entre 2000 et 2013. De la même manière, notre étude montre que parmi les 85 revues sélectionnées comme pouvant publier des *data papers* en rapport avec la médecine, jusqu'au 31 octobre 2018 seules 47% en ont publié au moins un. Chaque année, jusqu'en 2014, le nombre de revues publiant au moins un *data paper* n'a cessé de croître avant de devenir globalement stable. Nous nous sommes intéressés à tous types de journaux publiant des *data papers* (journal traditionnel, *data journal*, journal spécifique et multidisciplinaire) afin d'essayer d'avoir une idée représentative de ce phénomène au sein du domaine médical. Ainsi, nous avons démontré qu'il existe une grande variation en terme de volumétrie en fonction des journaux allant de quelques articles à plusieurs centaines. 78% des articles ont été publiés par l'un des 8 *data journals*. Même au sein de cette catégorie de journaux, il existe une hétérogénéité puisque la quasi-totalité des articles (90%) ont été publiés dans l'une des deux revues multidisciplinaires que sont *Data in Brief* et *Scientific Data*. Contrairement aux autres *data journals*, les articles de ces deux revues sont indexés dans les bases bibliographiques (Pubmed, Scopus, Web of Science) ce qui facilite leur visibilité et donc l'intérêt probable des auteurs à y publier leur article. D'autre part, la création de ces journaux dédiés aux données étant relativement récente, il peut être difficile de les faire connaître et de les pérenniser surtout dans un environnement de littérature scientifique dense. Ceci est d'autant plus vrai si la revue est spécifique et si elle n'est pas issue d'un grand éditeur tel que *Elsevier* ou *Nature*. C'est probablement une des raisons de l'arrêt de deux d'entre elles, *Dataset Papers in Science* et *Biomedical Data journal*. En parallèle, bien que la publication y reste faible, de plus en plus de revues traditionnelles autorisent la

publication de ce type de document. Ceci peut expliquer la faible évolution du nombre de *data journals*.

En terme de publication, nous avons identifié entre janvier 2004 et octobre 2018 745 *data papers* en rapport avec le domaine médical. Chaque année le nombre de publications a augmenté avec notamment une forte croissance à partir de 2014 pour atteindre près de 180 articles publiés par an dès 2016. La croissance semble se poursuivre puisque 180 articles ont été publiés au cours des 10 premiers mois de 2018. Cette évolution, bien que spécifique au domaine de la santé, semble cohérente avec l'évolution globale des *data papers* décrite par Candela (144). Ces résultats montrent qu'il existe un intérêt grandissant de la part des chercheurs médicaux probablement en lien avec le développement de l'*Open Science* et de l'*Open Data* qui ont contribué au développement de ces nouveaux articles et journaux. Néanmoins, il s'agit d'une pratique récente (71% des articles publiés au cours de ces deux dernières années) qui ne représente qu'une faible proportion de la littérature publiée (166). A noter que seuls les articles typés par les revues comme des *data papers* ont été pris en compte. Il est donc possible que certains journaux publient ce genre de document sous un format classique (notamment avant 2014) à l'instar de l'article de Sandercock (167) typé comme un article de recherche. Ces articles, probablement peu nombreux, ont le mérite d'exister mais sont difficilement repérables.

L'identification des articles représente une des grandes difficultés de notre étude. En effet, la recherche initiale sur les bases bibliographiques, basée sur la typologie des articles, ne s'est pas révélée aussi pertinente que nous l'avions envisagé et nous avons donc réalisé une recherche au sein même des revues. Parmi l'ensemble des revues existantes, nous nous sommes restreint à celles publiant des *data papers* recensées par différentes équipes ayant travaillé sur le sujet (Annexe 1). L'obtention

d’une liste exhaustive est difficile, voire impossible, mais nous avons pris en compte plusieurs listes afin d’avoir un maximum de ces revues. Pour avoir une certaine représentativité, nous avons sélectionné toutes les revues pouvant publier un article en rapport avec la médecine. Au sein même des revues, il n’est pas toujours aisé d’identifier quels sont les articles étant des *data papers* et ceux rattachés au domaine médical. Il est donc possible que certains articles n’aient pas été identifiés. De la même manière, des erreurs de classement concernant le typage des articles ont pu avoir lieu. En effet, il n’y a pas eu de croisement d’information, la recherche ayant été menée par un seul investigateur. Toutefois, l’objectif n’était pas d’être exhaustif mais d’avoir une idée de la tendance de ce nouveau phénomène au sein de notre discipline scientifique.

Concernant les différentes spécialités médicales, de façon surprenante la quasi-totalité des spécialités étaient représentées, même s’il existe de grandes différences quant au volume de publications. Ceci était peu envisageable il y a encore quelques années car les données en médecine sont souvent personnelles et le risque d’identification représente un frein majeur pour leur publication (114,116,118–120,123). Ces résultats montrent que les pratiques évoluent.

Dans le domaine médical c’est en cancérologie qu’il existe un grand nombre de publications notamment en lien avec des études sur le génome. Dans notre travail, la proportion de *data paper* pour cette spécialité est sous-estimée. En effet, nous avons d’emblée exclu les données de génomiques, or, la génomique fait partie des disciplines scientifiques qui a très tôt favorisé le partage de données notamment avec la création de banque de données tels que GenBank (103). L’analyse de ces résultats montre un biais de sélection. Toutefois, cette exclusion a permis d’accroître la visibilité des données cliniques.

Dans ces conditions, notre étude montre que la neurologie est la spécialité publiant le plus de *data papers*. Plusieurs éléments peuvent expliquer une telle différence de publication avec les autres spécialités. Tout d'abord, comme pour toutes les spécialités nous avons considéré la neurologie au sens large c'est-à-dire que nous avons pris en compte l'ensemble des articles étudiant le système nerveux humain et pas uniquement ceux en lien avec une pathologie spécifique (comme l'Alzheimer ou la sclérose en plaque par exemple). Notre champ d'analyse se rapproche donc plus de la neuroscience et peut expliquer le nombre important de *data papers*. La compréhension du fonctionnement du cerveau humain normal et pathologique est encore parcellaire et c'est un autre élément de différence. Afin d'améliorer cette compréhension, des études sont lancées et engendrent de très nombreuses données de neuroscience, probablement plus que d'autre spécialités. Ensuite, la part de l'imagerie et des examens électrophysiologiques pour comprendre le fonctionnement cérébral ne cessent de croître. Avec les avancées technologiques, les examens disponibles et les paramètres utilisés ne cessent d'évoluer, engendrant de nouvelles études physiopathologiques mais également technologiques, comme les études de reproductibilité des examens radiologiques, pourvoyeurs de multiples jeux de données. Enfin, à l'instar de la génomique, plusieurs initiatives de partages de données commencent à apparaitre tels que *OpenfMRI* (168) ou l'INDI (169,170). La volonté de faire connaître la mise à disposition de ces données au plus grand nombre, au-delà des utilisateurs usuels de ces plateformes, passe par la publication d'un article, en l'occurrence un *data paper*.

Au-delà de l'évolution croissante du nombre d'articles publiés, l'origine géographique des équipes est un autre élément témoignant de l'intérêt porté par la communauté médicale à ce nouveau format de publication. Même si les Etats-Unis

restent le pays le plus représenté, ce n'est pas loin de 38 nationalités, issues de tous les continents, qui ont participé à l'écriture d'un *data papers*. Avec des équipes nationales pour chacun des continents, cela témoigne d'un vrai intérêt mondial et pas d'une pratique isolée ou d'une pratique liée uniquement à l'internationalité des équipes. De même, la publication d'articles est de façon générale liée à une carrière universitaire, il n'est donc pas surprenant de voir un grand nombre d'articles rattachés à une unité de recherche universitaire. Cependant, de manière intéressante d'autre structures telles que des fondations, des ministères, des laboratoires ou encore des instituts de recherche utilisent ce nouveau type de document.

Il semble donc y avoir un certain attrait pour les *data papers* avec des signes plutôt encourageants pour cette nouvelle forme de communication scientifique au sein de la communauté médicale même si cela reste pour le moment un phénomène à petite échelle et peu homogène.

2 Des données médicales accessibles

Les sciences médicales font partie des disciplines dont les données sont les moins partagées (118,123). Les articles tels que celui de Sandercock (167) mettant à disposition les données d'un essai clinique randomisé de plus de 19400 patients par simple téléchargement d'un fichier, reste rare. Comme nous l'avons mentionné précédemment, les considérations éthiques et plus particulièrement la confidentialité des données des participants sont un des éléments freinant le partage de données (114,116,118–120,123). Cette spécificité des données médicales rend l'étude du partage des données au sein de cette discipline scientifique d'autant plus intéressant.

Nous avons vu en introduction qu'un *data paper* est défini comme un document décrivant un jeu de données mis à disposition de la communauté scientifique. Afin de

vérifier que leur utilisation correspondait bien à la définition d'un *data paper*, nous avons jugé pertinent de mener une étude plus approfondie sur le contenu de ces articles. Autrement dit, nous souhaitions vérifier que les données médicales étaient réellement accessibles en *open access*. Il était difficile de réaliser cette partie du travail sur l'ensemble des 745 articles et nous avons donc choisi d'étudier plus spécifiquement les articles de neuroscience pour plusieurs raisons 1) la volumétrie ; il s'agissait de la spécialité la plus représentée avec 163 articles (22%) 2) les thématiques ; notamment avec des données d'imagerie et de clinique représentées dans des proportions intéressantes 3) la typologie des journaux ; les articles ont été publiés à la fois dans des *data journals* et des revues mixtes. D'autres études devront être menées sur les autres spécialités afin d'avoir une vision plus globale sur la réalité de cette nouvelle pratique au sein des sciences médicales.

Les données médicales sont-elles réellement accessibles ? Dans notre étude seuls 67% des *data papers* présentent des données de recherche pouvant être réutilisées, et mettent à disposition leurs données individuelles. Ceci montre que plus de 30% des *data papers* ne correspondent pas à leur définition et ne répondent donc pas à leur finalité (la réutilisation de données). Parmi les *data papers* décrivant réellement la mise à disposition de données, le téléchargement des données n'était pas possible dans près d'un tiers des cas. Nous avons considéré un téléchargement possible uniquement s'il l'était directement c'est-à-dire sans démarche particulière or pour certain l'accès aux données tel que décrit dans l'article nécessitait *a priori* juste la création d'un compte utilisateur sur la plateforme d'accès rendant ainsi leur indisponibilité relative. Le résultat est de ce fait probablement surestimé. En revanche, pour d'autres, une réelle demande auprès des auteurs ou de l'organisme était nécessaire. Dans tous les cas, la démarche, les liens d'accès et de demande étaient

décrits ce qui témoigne, dans une certaine mesure, de la volonté de rendre accessible les données.

Existe-t-il une différence selon la typologie des données ? Sans avoir réalisé de test statistique, il semblerait que les données électro-physiologiques soient les données les plus accessibles. Dans ce cas, le format de ces données joue un rôle important. Il peut se faire soit sous forme d'un tracé brut soit sous forme d'un fichier comportant l'ensemble des paramètres permettant la reconstruction du tracé. Parfois, ce fichier peut se trouver au sein de l'article et donc plus facilement accessible. Dans notre travail, en neuroscience nous avons observé que, même si les *data papers* en imagerie sont plus nombreux, leur accessibilité n'est pas supérieure à celle des *data papers* de données clinico-comportementales ou de données électrophysiologiques. Néanmoins, les effectifs pour chaque type de données, hors ceux d'imagerie, restent relativement petits et peut expliquer cette différence. Il est à noter que nous avons pris en considération uniquement la donnée principalement décrite pour établir le type de données or pour certains articles plusieurs types de données peuvent être disponibles. Traitant de données médicales, il est intéressant de voir que la majorité des données directement téléchargeables n'étudie pas de pathologie spécifique. Il s'agit probablement d'études sur le fonctionnement « normal » du cerveau humain et de ce fait des données plutôt axées sur les examens complémentaires. Au contraire, avec des études en lien avec une pathologie, des données cliniques sont plus à même d'être étudiées associées ou non à des examens complémentaires. Plus il existe de données sur un individu, plus il est possible de remonter à l'identité de la personne et plus le processus d'anonymisation est complexe et coûteux. A notre avis, les auteurs ne sont pas opposés à partager leurs données puisque qu'ils décrivent leurs jeux de données mais encadrent l'accès afin de protéger les individus et de juger d'une réutilisation

adéquate de leurs données. Par ailleurs, il est possible que le consentement des sujets quant à la mise à disposition de leurs données personnelles soit plus facile à recueillir pour des patients indemnes de pathologie. Il serait intéressant de mener une enquête auprès des auteurs de *data papers* afin de recueillir leur retour d'expérience et d'identifier d'autres éléments favorisant ou non la mise à disposition de leurs données en *open access*.

Les données sont-elles de qualité ? Cette question est centrale pour une réutilisation optimale des jeux de données. Cependant, notre étude ne visait pas à évaluer la qualité des données. Des études plus spécifiques seront nécessaires pour aborder cette problématique.

En résumé, en neuroscience, certaines données médicales sont accessibles. Ce sont essentiellement des données électrophysiologiques (EEG) et des données d'imagerie (IRM). Même si dans certains cas l'accès est contrôlé, la démarche de rédaction d'un *data paper* témoigne d'une volonté certaine de partage et permet de créer de nouvelles collaborations.

3 Des données médicales réutilisées

La citation reflète la façon dont une ressource est utilisée au sein de la communauté scientifique et représente actuellement le principal critère de reconnaissance. La citation de données vise à promouvoir une référence spécifique à un ensemble de données utilisé au cours d'une étude. Cependant, son utilité est limitée car les ensembles de données sont rarement cités formellement (165,171–174) par manque de norme et de reconnaissance (165,175,176). Pourtant, les chercheurs sont en faveur de la citation formelle des données (127,175,177) notamment parce que les études qui mettent à disposition leurs données reçoivent plus de citations

(57,58,178,179). Ceci laisse supposer qu'une partie des citations est liée à la réutilisation des données sans que celles-ci ne soient directement citées (180). Dans ce contexte, le *data paper* permet de lier le problème de la citation de données avec le référencement habituel reconnu ; les auteurs d'un article de recherche souhaitant citer un ensemble de données qu'ils ont utilisé ne citent pas l'ensemble de données en tant que tel mais le *data paper* qui décrit cet ensemble de données pour lequel la citation se fait de la même manière que pour tout autre article. Ainsi, l'étude des citations des *data papers* permet d'aborder la question de la citation des données. A ce jour, nous n'avons pas retrouvé de travaux s'intéressant spécifiquement à cette voie d'étude des citations de données ce qui donne un intérêt supplémentaire à notre étude.

De façon intéressante, seulement 15% des *data papers* n'ont pas été cités alors que les jeux de données ne sont pas cités dans 88% des cas comme le démontrent Robinson-Gacia et al (172), Peters et al (173) ou Park et al (174), en étudiant la citation formelle des données via le Data Citation Index (181–183). A noter que nous nous sommes intéressés uniquement aux *data papers* rattachés à la neuroscience. Chaque discipline a ses pratiques en matière de partage et de citation de données (122,123,172,182) et il est difficile de faire des comparaisons interdisciplinaires. Ce résultat suggère cependant que le *data paper* semble être un format adapté et accepté par les chercheurs pour la citation de données. L'augmentation du nombre de citations annuel reçu par ces nouveaux articles confirme leur popularité croissante et appuie les résultats globaux retrouvés dans le rapport conjoint du CWTS de l'Université de Leiden et d'*Elsevier* (128).

D'un point de vue quantitatif, nous avons constaté que le nombre médian de citation d'un *data paper* en neurosciences est de 3. Ce résultat est inférieur au nombre médian de citations pour les articles de neuroscience lié à des données et pour les

articles de neuroscience classique avec une médiane respectivement à 8 et à 6 (179). Même si les résultats obtenus restent du même ordre de grandeur, plusieurs éléments peuvent expliquer cette différence 1) Notre recherche est basée sur une recherche manuelle et est de façon certaine non exhaustive avec un échantillon de petite taille ; 2) La rédaction d'un *data paper* ainsi défini est encore une pratique récente, il est possible que des ensembles de données soient décrits dans des articles classiques ; 3) Il existe un manque de recul pour les citations des *data paper* les plus récents alors qu'ils sont les plus nombreux ; 4) Les articles liés aux données dans l'étude de Leither (179) ont été sélectionnés à partir des termes Mesh « Atlas » ou « Base de données » en tant que sujet principal mais il n'y a pas de certitude que ces articles mettent réellement à disposition des données ; 5) Nous supposons que le nombre de citations d'un *data paper* est davantage lié à la réutilisation des données qu'un article traditionnel mettant à disposition ses données pour lequel le nombre de citations reçu peut être lié à la réutilisation des données mais surtout est lié aux résultats obtenus ; 6) La neuroscience est une discipline où la pratique du partage de données est en cours de développement.

Nous nous sommes également intéressés au délai de citation de ces *data papers* et nous avons trouvé un délai médian de 8 mois. Autrement dit, la moitié des ensembles de données publiés et documentés est réutilisée moins d'un an après leur publication. Ceci montre l'intérêt et la réactivité des chercheurs quant à l'exploitation de nouvelles données mises à disposition. L'idée que les données puissent être rapidement citées peut-être un élément supplémentaire pouvant encourager à la publication des données.

Nous avons observé la portée des *data papers* et la réutilisation potentielle des ensembles de données. De façon intéressante, les publications citant les *data papers*

de neuroscience appartiennent dans certains cas à des domaines de recherche très éloignés tels que les sciences sociales, l'agriculture ou les arts. Cela rejoint les résultats observés par Chao (184), qui, en étudiant la réutilisation des ensembles de données des sciences de la terre, a montré que la génération d'ensembles de données dans une discipline pouvait avoir une application dans un autre domaine d'étude. Les ensembles de données peuvent donc avoir un impact sur un grand nombre de disciplines universitaires amenant à faire évoluer la recherche, ouvrant à de nouvelles collaborations transdisciplinaires et aboutissant ainsi à une certaine forme d'innovation.

La citation de données ne reflète pas nécessairement la réutilisation réelle des données. Elle peut aussi être utilisée dans un but informatif. Piwowar et al (178) ont montré que seul 6% des citations d'articles liés à des données relevaient du contexte de la réutilisation des données. Nous n'avons pas étudié spécifiquement le contexte de citation des *data papers* mais la typologie et le domaine d'étude des citations indiquerait qu'il existe une certaine part de citation informative. Néanmoins, avec la seule description des jeux de données contenue dans l'article et sans résultats finaux d'une recherche, nous supposons que la citation des *data papers* serait plus à même de donner une idée sur la réutilisation réelle des ensembles de données contrairement aux articles classiques mettant à disposition des données ou à la citation formelle des jeux de données actuellement faiblement utilisée. Pour confirmer cette hypothèse une étude plus approfondie du contexte de citation des *data paper* est nécessaire. Par ailleurs, afin de rendre compte de l'impact réel des ensembles de données en prenant en compte le contexte dans lequel les citations se produisent, utilisation ou information, certains chercheurs tentent de développer de nouveaux indicateurs tel que l'U-index (185).

Un autre élément à considérer également est l'autocitation des articles. L'utilisation de la base Scopus pour l'étude des citations nous a permis d'appréhender la notion d'autocitation et nous avons observé que 27% des citations correspondaient à une autocitation (17% du premier ou du dernier auteur et 10% d'un coauteur). Cela rejoint d'autres études qui retrouvent que les mêmes auteurs ont tendance à utiliser les mêmes données plusieurs fois et que l'autocitation des données est une pratique relativement courante (172,174,186). La prise en compte ou non de ces autocitations nous a montré dans certains cas des résultats différents. Cette observation laisse supposer que l'impact du partage de données est peut-être moins important qu'attendu avec une réutilisation en partie liée aux créateurs de l'ensemble de données. L'augmentation du nombre de citations n'est donc pas forcément liée à des citations de la part de nouveaux chercheurs. Cependant comme l'a déjà indiqué Park (174) « l'autocitation doit être étudiée plus en détail concernant la citation des données ».

Un ensemble de données, comme toute ressource, peut être réutilisé sans que cela génère une citation soit parce que l'auteur n'a pas cité cet ensemble de données (180) soit parce que la réutilisation a été réalisée dans un contexte autre que la recherche scientifique tel que la formation pédagogique par exemple. Afin d'évaluer l'impact d'une publication au-delà de la littérature savante, des nouveaux indicateurs basés sur le web social, appelé *altmetrics*, sont de plus en plus étudiés(187–196). L'intérêt de ces nouvelles métriques est d'obtenir des informations quasi en temps réel sur l'influence et l'évolution de l'attention portée à une publication mise en ligne, complétant ainsi la vision de l'impact à long terme établie par l'étude des citations. De façon plus globale, ils permettent d'étudier l'impact de la recherche sur la société.

Les premières études ont montré que seul 15 à 25% des publications scientifiques présentaient une activité *altmetric* avec des variations selon les disciplines

scientifiques (193,197,198). Cette couverture est encore plus faible pour les ensembles de données (entre 4 à 9%) (173). Parmi l'ensemble des plates-formes de média sociaux sur lesquels se base ces nouveaux indicateurs, Twitter et les réseaux sociaux scientifiques tel que Mendeley sont les principales sources sociales retrouvées et étudiées (193,199–207). Haustein et al (208) ont montré que la typologie du document influence la visibilité des publications sur les réseaux sociaux avec les éditoriaux, les actualités et les revues davantage présents que les articles. Avec une couverture globale à 98% (86% pour les réseaux sociaux scientifiques et 53% pour Twitter) le *data paper* semble être un document relativement populaire sur le web. La description des caractéristiques et du traitement du jeu de données peuvent être des éléments amenant à la discussion et aux partages que le jeu soit par la suite réutilisé ou non, signant ainsi un certain intérêt social. Il est probablement plus facile de discuter des forces et faiblesses d'un jeu de données à partir d'un document qui le décrit qu'à partir de l'ensemble lui-même. Ces activités peuvent aussi participer à l'évaluation de la qualité du jeu données. Par ailleurs, il s'agit d'un document court et synthétique, apprécié des réseaux sociaux (208).

L'équipe de Kratz indique que l'impact d'une même métrique peut être différent entre un article et un ensemble de données (177). Par exemple, la lecture d'un article en ligne peut aboutir à une utilisation secondaire des informations contenues dans cet article sans nécessairement avoir téléchargé la version PDF, le nombre de vues de la page peut alors être intéressant. Au contraire pour un ensemble de données, connaitre le nombre de fois que la page d'accès a été affichée n'a que peu d'intérêt par rapport au nombre de fois qu'il a été téléchargé. Etant considéré comme un article, les mesures *altmetrics* d'usage d'un *data paper* (nombre de vues de l'abstract, nombre de vues du texte entier) peuvent témoigner là aussi de l'intérêt porté à un ensemble de

données. Aucune de ces mesures d'impact alternatives mesurées à partir du *data paper* ne peuvent présumer du téléchargement du jeu de donnée et de sa réutilisation future mais elles peuvent permettre d'évaluer l'intérêt que suscite celui-ci peut être plus facilement et plus rapidement. Bien que les *almetrics* offrent une visibilité plus rapide et étudient l'impact plus largement, des obstacles théoriques, méthodologiques et techniques doivent encore être levés quant à leur utilisation et à leur interprétation (188,209).

Plusieurs équipes de recherche étudient des mesures traditionnelles, alternatives et nouvelles pour les données afin d'avoir une meilleure compréhension sur la façon de mesurer leur impact et ainsi permettre de reconnaitre les créateurs de données (130–133,185,210,211). En attendant, l'idée du *data paper* comme moyen de publier ses données semble être un moyen intermédiaire intéressant dans un système où la reconnaissance bibliométrique prime.

4 Une pratique hétérogène

Les ensembles de données sont de plus en plus considérés comme ayant une véritable valeur en soi, pouvant être publiés de façon indépendante d'un article scientifique. En attendant la systématisation et la standardisation des pratiques de citation de données (136,212–216), le *data paper* semble pouvoir répondre momentanément aux besoins de reconnaissance des auteurs de données en permettant aux chercheurs de publier leur ensemble de données en tant que publication scientifique citable. Cependant, la rédaction d'un *data paper* soulève des questions et présente certains défis (164). Les résultats de recherche de Schöepfel (143) ont montré une certaine hétérogénéité dans la pratique que nous avons également retrouvée au cours de notre étude :

- Les *data papers* ne sont pas publiés spécifiquement dans un *data journal*. Bien que la publication y soit encore faible, de plus en plus de revues classiques acceptent de les publier. Comme déjà mentionné l'un des avantages de notre étude est d'avoir pris en compte ces deux types de revues même si la liste complète des revues classiques acceptant leur publication est impossible à référencer.
- Les *data journals* acceptent également d'autres types d'articles tels que des articles de recherche ou des commentaires par exemple. Percevoir la différence entre une revue de données dite « pure » (*data journal*) et « mixte » (revue classique) n'est pas si aisé rendant la catégorisation d'une revue complexe. Il est possible que la diversification des publications donne un espoir de pérenniser les revues, la publication de *data paper* étant encore une pratique jeune.
- La terminologie utilisée varie d'une revue à l'autre, ce qu'avait déjà relevé Candela en 2014 (144). Même au sein des *data journals*, il existe des différences. Chose peut être plus étonnante, un même éditeur peut utiliser des termes différents pour chacune de ses revues. La recherche de ce type de document au sein des revues peut donc s'avérer complexe pour un chercheur novice. On peut supposer que la variété de ces dénominations est en partie liée à la diversité des communautés scientifiques qui utilisent ces documents.
- La plupart des *data journals* ne sont pas référencés dans les bases bibliométriques, limitant ainsi leur visibilité. Par ailleurs, la typologie donnée par ces différentes bases à ces nouveaux articles ne semble pas permettre d'identifier de façon exhaustive l'ensemble des *data papers* tels qu'identifiés par les revues. Scopus ne fait aucune différence avec un article de recherche. Pour

une grande partie, aucune typologie n'a été donnée par Pubmed. Seul le Web of Science semble pouvoir s'y approcher. Cette difficulté d'identification sur les grandes bases bibliométriques a été un des freins concernant l'identification de nos articles, nous amenant à réaliser une recherche au sein même des revues.

- Il n'existe pas de format standard à l'écriture d'un *data paper* et les rubriques varient selon les revues. Seule l'accessibilité à l'ensemble des données semble être un élément commun. De plus, certains articles ne décrivent pas seulement l'ensemble de données mais présentent certains résultats de l'analyse de données. D'autres présentent uniquement des résultats de recherche ou encore mettent à disposition seulement les codes et outils nouvellement créés. L'absence d'un cadre central minimal et commun à la publication d'un *data paper* est sans doute une limite importante pour promouvoir la réutilisation des données qui y sont décrites et assurer de leur qualité. Dans le cadre de la neuroimagerie, Gorgolewski a indiqué les éléments minimaux qui devraient selon lui être retrouvés (164). En parcourant les *data papers* de 18 revues, les articles édités par *Ubiquity Press* nous semblent les mieux sectorisés avec 4 parties principales (contexte, méthode, description de l'ensemble de données, réutilisation potentielle) elle-même partitionnée permettant ainsi de trouver rapidement l'essentiel des informations attendues (217–219).

Le *data paper* est une communication scientifique encore récente pouvant expliquer cette hétérogénéité de pratiques. Il faudra probablement encore plusieurs années, en associant l'ensemble des acteurs, pour harmoniser les pratiques et aboutir à un format standardisé reconnu et accepté à l'image du format IMRAD établi pour la publication d'articles scientifiques.

5 Conclusion

Notre étude a permis d'étudier l'émergence du *data paper* au sein des sciences médicales et plus particulièrement en neuroscience. La description d'ensembles de données permet de fournir de l'information afin de permettre une réutilisation optimale du jeu de données et de contribuer à la génération de connaissances. Les pratiques sont sans aucun doute différentes selon les disciplines et spécialités, et d'autres études seront nécessaires pour obtenir une meilleure vision de l'utilisation et de l'impact de cette nouvelle forme de communication scientifique. Dans un environnement où la quantité de données ne cesse d'augmenter et où les outils d'analyse sont toujours plus performants, la question du partage des données est centrale et le besoin de reconnaissance des publications de données se fait de plus en plus pressant. En attendant de nouvelles pratiques d'attribution de crédits universitaires, le *data paper* semble être un moyen alternatif acceptable dont le nombre continuera probablement à croître et à gagner en importance dans les années à venir.

Références bibliographiques

1. Elliott KC, Cheruvelil KS, Montgomery GM, Soranno PA. Conceptions of Good Science in Our Data-Rich World. BioScience. 1 oct 2016;66(10):880-9.

2. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. BioEssays News Rev Mol Cell Dev Biol. janv 2004;26(1):99-105.

3. Elliott KC. Epistemic and methodological iteration in scientific research. Stud Hist Philos Sci Part A. 1 juin 2012;43(2):376-82.

4. O'Malley MA, Elliott KC, Burian RM. From genetic to genomic regulation: iterativity in microRNA research. Stud Hist Philos Biol Biomed Sci. déc 2010;41(4):407-17.

5. Définition : Empirisme [Internet]. [cité 18 juin 2020]. Disponible sur: http://www.toupie.org/Dictionnaire/Empirisme.htm

6. Définition : Induction [Internet]. [cité 18 juin 2020]. Disponible sur: http://www.toupie.org/Dictionnaire/Induction.htm

7. Glass DJ, Hall N. A Brief History of the Hypothesis. Cell. 8 août 2008;134(3):378-81.

8. Haufe C. Why do funding agencies favor hypothesis testing?. Stud Hist Philos Sci Part A. 1 sept 2013;44(3):363-74.

9. Roberts RM. Serendipity. New York: Wiley; 1989. 290 p.

10. Waters CK. The Nature and Context of Exploratory Experimentation: An Introduction to Three Case Studies of Exploratory Research. Hist Philos Life Sci. 2007;29(3):275-84.

11. Définition : Rationalisme [Internet]. [cité 18 juin 2020]. Disponible sur: http://www.toupie.org/Dictionnaire/Rationalisme.htm

12. Définition : Déduction, raisonnement déductif ou syllogistique [Internet]. [cité 18 juin 2020]. Disponible sur: http://www.toupie.org/Dictionnaire/Deduction.htm

13. Donovan SM, O'Rourke M, Looney C. Your Hypothesis or Mine? Terminological and Conceptual Variation Across Disciplines. SAGE Open. 1 avr 2015;5(2):2158244015586237.

14. O'Malley MA, Elliott KC, Haufe C, Burian RM. Philosophies of Funding. Cell. 21 août 2009;138(4):611-5.

15. Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data Soc. 1 avr 2014;1(1):2053951714528481.

16. Bell G, Hey T, Szalay A. Beyond the Data Deluge. Science. 6 mar 2009;323(5919):1297-8.

17. Mattmann CA. A vision for data science. Nature. janv 2013;493(7433):473-5.

18. Baraniuk RG. More Is Less: Signal Processing and the Data Deluge. Science. 11 fév 2011;331(6018):717-9.

19. McCue ME, McCoy AM. The Scope of Big Data in One Medicine: Unprecedented Opportunities and Challenges. Front Vet Sci. 2017;4:194.

20. Michener WK, Jones MB. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol Evol. fév 2012;27(2):85-93.

21. Fillinger S, de la Garza L, Peltzer A, Kohlbacher O, Nahnsen S. Challenges of big data integration in the life sciences. Anal Bioanal Chem. oct 2019;411(26):6791-800.

22. Tolle KM, Tansley S, Hey T. The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. Proc IEEE. 1 août 2011;99:1334-7.

23. Hey T, Tansley S, Tolle K, eds. Jim Gray on eScience: a transformed scientific method. In: Hey T, Tansley S, Tolle K, eds. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research, 2009.

24. Leonelli S. Introduction: Making sense of data-driven research in the biological and biomedical sciences. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci. 1 mar 2012;43(1):1-3.

25. Pietsch W. Aspects of Theory-Ladenness in Data-Intensive Science. Philos Sci. 1 déc 2015;82(5):905-16.

26. Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired [Internet]. 23 juin 2008 [cité 17 juin 2020]; Disponible sur: https://www.wired.com/2008/06/pb-theory/

27. Prensky M. H. Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom. Innov J Online Educ. 2009;5(3).

28. Dyche J. Big Data “Eurekas!” Don’t Just Happen. Harvard Business Review [Internet]. 20 nov 2012 [cité 17 juin 2020]; Disponible sur: https://hbr.org/2012/11/eureka-doesnt-just-happen

29. Siegel E, Davenport TH. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. Hoboken, New Jersey: Wiley; 2013. 320 p.

30. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 14 mar 2014;343(6176):1203-5.

31. Butler D. When Google got flu wrong. Nat News. 14 fév 2013;494(7436):155.

32. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLOS ONE. 19 août 2011;6(8):e23610.

33. Smalheiser NR. Informatics and hypothesis-driven research. EMBO Rep. 15 août 2002;3(8):702.

34. Kelling S, Hochachka W, Fink D, Ridewald M, Caruana R, Ballard G, et al. Data-Intensive Science: A New Paradigm for Biodiversity Studies. BioScience. 1 juil 2009;59:613-20.

35. Calvert J. Systems biology, synthetic biology and data-driven research: A commentary on Krohs, Callebaut, and O'Malley and Soyer. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci. 1 mar 2012;43(1):81-4.

36. Miller HJ. The Data Avalanche Is Here. Shouldn't We Be Digging?. J Reg Sci. 2010;50(1):181-201.

37. Maass W, Parsons J, Purao S, Storey V, Woo C. Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research. J Assoc Inf Syst. 28 déc 2018;19(12).

38. Strasser BJ. Data-driven sciences: From wonder cabinets to electronic databases. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci. 1 mar 2012;43(1):85-7.

39. National Consortium for Data Science [Internet]. [cité 20 juin 2020]. Disponible sur: https://datascienceconsortium.org/

40. Jagadish HV. Big Data and Science: Myths and Reality. Big Data Res. 1 juin 2015;2(2):49-52.

41. Schöpfel J, Kergosien E, Prost H. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. Atelier VADOR : Valorisation et Analyse des Données de la Recherche; INFORSID 2017. Toulouse, France. mai 2017

42. Gaillard R. De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ? [Mémoire]. Enssib [Internet]. janv 2014. [cité 13 juin 2020]. Disponible sur: https://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche

43. Organisation de Coopération et de Développement Economique. Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. 2007.

44. Rosemberg N. De la définition des données de la recherche. En quête des données [Internet]. [cité 15 juin 2020]. Disponible sur: https://donneesshs.hypotheses.org/39

45. Salami S. Ouverture des données de la recherche : de quoi parle-t-on ?. Le magazine [Internet]. 2018 [cité 15 juin 2020]. Disponible sur: https://www.didaktic.fr/science-ouverte/ouverture-des-donnees-de-la-recherche-de-quoi-parle-t-on/

46. Fournier T. « Les données de la recherche : définitions et enjeux ». Arabesques. 2014;73:4-5.

47. Australian National Data Service. ANDS guide "What is research data?" [Internet]. 11 janv 2017 [cité 15 juin 2020]. Disponible sur: https://www.ands.org.au/guides/what-is-research-data

48. University of Bristol. What counts as research data? [Internet]. [cité 15 juin 2020]. Disponible sur: https://data.blogs.bristol.ac.uk/bootcamp/data/

49. UK Data Service. Research data lifecycle [Internet]. [cité 15 juin 2020]. Disponible sur: https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx

50. Institut de l'Information Scientifique et Technique. Une introduction à la gestion et au partage des données de la recherche [Internet]. 16 sept 2014 [cité 15 juin 2020]. Disponible sur: https://www.inist.fr/wp-content/uploads/donnees/co/Donnees_recherche_web.html

51. Prost H, Schöpfel J. Les données de la recherche en SHS. Une enquête à l'Université de Lille 3 : rapport final [Rapport de recherche]. Lille; 2015.

52. Institut de l'Information Scientifique et Technique. Gestion et diffusion des données de la recherche [Internet]. 22 mars 2016 [cité 15 juin 2020]. Disponible sur: https://urfist.univ-lyon1.fr/files/2016/06/Urfist_Lyon_Donnees_de_la_recherche_gestion_diffusion.pdf

53. Lawrence B, Jones C, Matthews B, Pepler S, Callaghan S. Citation and Peer Review of Data: Moving Towards Formal Data Publication. Int J Digit Curation. 7 oct 2011;6(2):4-37.

54. Reilly S, Schallier W, Schrimpf S, Smit E, Wilkinson M. Report on Integration of Data and Publications. Opportunities for Data Exchange (ODE). 2011.

55. Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE, Martone ME. Big data from small data: data-sharing in the « long tail » of neuroscience. Nat Neurosci. nov 2014;17(11):1442-7.

56. Wallis JC, Rolando E, Borgman CL. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PloS One. 2013;8(7):e67332.

57. Piwowar HA, Day RS, Fridsma DB. Sharing Detailed Research Data Is Associated with Increased Citation Rate. Ioannidis J, ed. PLoS ONE. 21 mars 2007;2(3):e308.

58. Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. PLOS ONE. 22 avr 2020;15(4):e0230416.

59. Belter CW. Global-level data sets may be more highly cited than most journal articles. Impact of Social Sciences [Internet]. 2014 [cité 15 juin 2020]. Disponible sur: https://blogs.lse.ac.uk/impactofsocialsciences/2014/05/15/global-level-data-sets-highly-cited/

60. Belter CW. Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. PLOS ONE. 26 mar 2014;9(3):e92590.

61. Wallis JC, Rolando E, Borgman CL. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLOS ONE. 23 juil 2013;8(7):e67332.

62. Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, et al. The Availability of Research Data Declines Rapidly with Article Age. Curr Biol. 6 janv 2014;24(1):94-7.

63. Centre National de la Recherche Scientifique. Les données de la recherche dans le contexte de l'Open Science [Internet]. [cité 15 juin 2020]. Disponible sur: https://doranum.fr/wp-content/uploads/ANF_RENATIS_2016_FANDRE_1.pdf

64. Un historique du libre accès aux publications et aux données [Internet]. [cité 13 juin 2020]. Disponible sur: https://www.ouvrirlascience.fr/un-historique-du-libre-acces-aux-publications-scientifiques-et-aux-donnees

65. Déclaration de Berlin sur le Libre Accès à la Connaissance en sciences exactes, sciences de la vie, sciences humaines et sociales. 22 oct 2003.

66. Déclaration de l'OCDE sur l'accès aux données de la recherche financée par des fonds publics. 30 janv 2004.

67. Charte du G8 pour l'Ouverte des Données Publiques. 18 juin 2013.

68. LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique. 2016-1321 oct 7, 2016.

69. Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Le Plan national pour la science ouverte. 4 juil 2018.

70. NIH. NOT-OD-03-032: FINAL NIH STATEMENT ON SHARING RESEARCH DATA [Internet]. [cité 13 juin 2020]. Disponible sur: https://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html

71. NIH. NIH Grants Policy Statement [Internet]. [cité 13 juin 2020]. Disponible sur: https://grants.nih.gov/grants/policy/nihgps_2012/nihgps_ch8.htm#_Toc271264950

72. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nat News. 30 janv 2014;505(7485):612.

73. NIH. NOT-OD-08-013: Implementation Guidance and Instructions for Applicants: Policy for Sharing of Data Obtained in NIH-Supported or Conducted Genome-Wide Association Studies (GWAS) [Internet]. [cité 13 juin 2020]. Disponible sur: https://grants.nih.gov/grants/guide/notice-files/not-od-08-013.html

74. National Science Foundation. Dissemination and Sharing of Research Results [Internet]. [cité 13 juin 2020]. Disponible sur: https://www.nsf.gov/bfa/dias/policy/dmp.jsp

75. Gordon and Betty Moore Foundation. Data sharing philosophy [Internet]. [cité 13 juin 2020]. Disponible sur: https://www.moore.org/docs/default-source/Grantee-Resources/data-sharing-philosophy.pdf

76. Bill and Melinda Gates Foundation. Open Access Policy [Internet]. [cité 13 juin 2020]. Disponible sur: https://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy

77. Medical Research Council. Data Sharing Policy [Internet]. [cité 13 juin 2020]. Disponible sur: https://mrc.ukri.org/documents/pdf/mrc-data-sharing-policy/

78. Welcome Trust. Sharing research data to improve public health statement [Internet]. [cité 13 juin 2020]. Disponible sur: https://wellcome.ac.uk/what-we-do/our-work/sharing-research-data-improve-public-health-full-joint-statement-funders-health

79. Burton PR, Banner N, Elliot MJ, Knoppers BM, Banks J. Policies and strategies to facilitate secondary use of research data in the health sciences. Int J Epidemiol. 1 déc 2017;46(6):1729-33.

80. Commission Européenne. Horizon 2020 en bref - Le programme-cadre de l'UE pour la recherche et l'innovation. 2014.

81. European Commission. H2020 Programme - Guidelines on FAIR Data Managament in Horizon 2020. 26 juil 2016.

82. National Academy of Sciences, National Academy of Engineering and Institute of Medicine. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Washington, DC: The National Academies Press; 2009.

83. Institute of Medecine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: The National Academies Press; 2015.

84. Hampton T. European Drug Agency Works to Improve Transparency, but Skepticism Remains. JAMA. 5 sept 2012;308(9):850-1.

85. Koenig F, Slattery J, Groves T, Lang T, Benjamini Y, Day S, et al. Sharing clinical trial data on patient level: Opportunities and challenges. Biom J. 2015;57(1):8-26.

86. Krumholz HM, Ross JS. A Model for Dissemination and Independent Analysis of Industry Data. JAMA. 12 oct 2011;306(14):1593-4.

87. Lin J, Strasser C. Recommendations for the Role of Publishers in Access to Data. PLOS Biol. 28 oct 2014;12(10):e1001975.

88. Hrynaszkiewicz I. Publishers' Responsibilities in Promoting Data Quality and Reproducibility. In: Bespalov A, Michel MC, Steckler T, eds. Good Research Practice in Non-Clinical Pharmacology and Biomedicine. Handbook of Experimental Pharmacology, vol 257. Cham: Springer; 2019. p. 319-48.

89. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu Y-L, Snyders LN, et al. Data sharing in PLOS ONE: An analysis of Data Availability Statements. PloS One. 2018;13(5):e0194768.

90. PLOS ONE. Data availability [Internet]. [cité 13 juin 2020]. Disponible sur: https://journals.plos.org/plosone/s/data-availability

91. Nature Research. Reporting standards and availability of data, materials, code and protocols [Internet]. [cité 13 juin 2020]. Disponible sur: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-data

92. Hanson B, Sugden A, Alberts B. Making Data Maximally Available. Science. 11 fév 2011;331(6018):649-649.

93. Science AAAS. Science Journals: editorial policies [Internet]. [cité 13 juin 2020]. Disponible sur: https://www.sciencemag.org/authors/science-journals-editorial-policies

94. Groves T. BMJ policy on data sharing. BMJ. 29 janv 2010;340.

95. BMJ. Data sharing - BMJ Author Hub [Internet]. [cité 13 juin 2020]. Disponible sur: https://authors.bmj.com/policies/data-sharing/

96. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data Sharing Statements for Clinical Trials — A Requirement of the International Committee of Medical Journal Editors. N Engl J Med. 8 juin 2017;376(23):2277-9.

97. Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. Reproducible and reusable research: are journal data sharing policies meeting the mark?. PeerJ. 25 avr 2017;5:e3208.

98. Piwowar H, Chapman W. A review of journal policies for sharing research data. Nat Preced. 20 mar 2008;1-1.

99. Resnik DB, Morales M, Landrum R, Shi M, Minnier J, Vasilevsky NA, et al. Effect of Impact Factor and Discipline on Journal Data Sharing Policies. Account Res. avr 2019;26(3):139-56.

100. Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ Open. 1 oct 2016;6(10):e011784.

101. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. Public Availability of Published Research Data in High-Impact Journals. PLOS ONE. 7 sept 2011;6(9):e24357.

102. Swan A, Brown S. To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. 1 juin 2008.

103. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. janv 2010;38(suppl_1):D46-51.

104. Clough E, Barrett T. The Gene Expression Omnibus database. Methods Mol Biol Clifton NJ. 2016;1418:93-110.

105. Sardanelli F, Alì M, Hunink MG, Houssami N, Sconfienza LM, Di Leo G. To share or not to share? Expected pros and cons of data sharing in radiological research. Eur Radiol. 1 juin 2018;28(6):2328-35.

106. Hrynaszkiewicz I, Khodiyar V, Hufton AL, Sansone S-A. Publishing descriptions of non-public clinical datasets: proposed guidance for researchers, repositories, editors and funding organisations. Res Integr Peer Rev. 22 juin 2016;1(1):6.

107. Toga AW, Dinov ID. Sharing big biomedical data. J Big Data. 27 juin 2015;2(1):7.

108. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. Nat Neurosci. nov 2014;17(11):1510-7.

109. Poline J-B, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, et al. Data sharing in neuroimaging research. Front Neuroinformatics. 2012;6:9.

110. Ross JS, Lehman R, Gross CP. The Importance of Clinical Trial Data Sharing: Toward More Open Science. Circ Cardiovasc Qual Outcomes. 1 mar 2012;5(2):238-40.

111. Commission Européenne. Pour un meilleur accès aux informations scientifiques: dynamiser les avantages des investissements publics dans le domaine de la recherche.17 juil 2012.

112. Bull S, Roberts N, Parker M. Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public Health Research: A Systematic Scoping Review. J Empir Res Hum Res Ethics. juil 2015;10(3):225-38.

113. Vickers AJ. Whose data set is it anyway? Sharing raw data from randomized trials. Trials. 16 mai 2006;7:15.

114. Federer LM, Lu Y-L, Joubert DJ, Welsh J, Brandys B. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. Kanungo J, ed. PLOS ONE. 24 juin 2015;10(6):e0129506.

115. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. PLOS ONE. 29 juin 2011;6(6):e21101.

116. Rathi V, Dzara K, Gross CP, Hrynaszkiewicz I, Joffe S, Krumholz HM, et al. Sharing of clinical trial data among trialists: a cross sectional survey. BMJ. 20 nov 2012;345:e7570.

117. Wiley. Researcher data sharing insights [Internet]. 2014 [cité 10 juin 2020]. Disponible sur: https://3spxpi1radr22mzge33bla91-wpengine.netdna-ssl.com/wp-content/uploads/2014/11/researcher-data-insights-infographic-final.pdf

118. Ferguson L. How and why researchers share data (and why they don't) [Internet]. 3 nov 2014 [cité 10 juin 2020]. Disponible sur: https://www.wiley.com/network/researchers/licensing-and-open-access/how-and-why-researchers-share-data-and-why-they-dont

119. Majumder K. Researchers' attitudes towards data sharing [Internet]. Editage Insights. 30 déc 2014 [cité 10 juin 2020]. Disponible sur: https://www.editage.com/insights/researchers-attitudes-towards-data-sharing

120. Measdows A. To Share or not to Share? That is the (Research Data) Question… [Internet]. The Scholarly Kitchen. 11 nov 2014 [cité 10 juin 2020]. Disponible sur: https://scholarlykitchen.sspnet.org/2014/11/11/to-share-or-not-to-share-that-is-the-research-data-question/

121. Wiley. Global Data Sharing Trends [Internet]. juin 2017 [cité 10 juin 2020]. Disponible sur: https://authorservices.wiley.com/asset/photos/licensing-and-open-access-photos/Wiley%20Global%20Data%20Sharing%20Infographic%20June%202017.pdf

122. Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE. 26 août 2015;10(8):e0134826.

123. Stuart, D., Baynes, G., Hrynaszkiewicz, I, Allin, K., Penny, D., Lucraft, M, et al. Whitepaper: Practical challenges for researchers in data sharing. Figshare, Journal Contribution. 21 mar 2018.

124. Hahnel M, Treadway J, Fane B, Kiley R, Peters D, et al. The State of Open Data Report 2017. Digital Science. 23 oct 2017.

125. Rathi VK, Strait KM, Gross CP, Hrynaszkiewicz I, Joffe S, Krumholz HM, et al. Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey. Trials. 2 oct 2014;15(1):384.

126. Piwowar HA, Chapman WW. Public sharing of research datasets: A pilot study of associations. J Informetr. avr 2010;4(2):148-56.

127. Kratz JE, Strasser C. Researcher Perspectives on Publication and Peer Review of Data. PLOS ONE. 23 fév 2015;10(2):e0117619.

128. Meijer I, Berghmans S, Cousijn H, Tatum C, Deakin G, Plume A, et al. Open Data: the researcher perspective. 2017.

129. Budin-Ljøsne I, Isaeva J, Maria Knoppers B, Marie Tassé A, Shen H, McCarthy MI, et al. Data sharing in large research consortia: experiences and recommendations from ENGAGE. Eur J Hum Genet. mar 2014;22(3):317-21.

130. Kalager M, Adami H-O, Bretthauer M. Recognizing Data Generation. N Engl J Med. 12 mai 2016;374(19):1898-1898.

131. Bierer BE, Crosas M, Pierce HH. Data Authorship as an Incentive to Data Sharing. N Engl J Med. 27 avr 2017;376(17):1684-7.2.

132. Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. Nature. juin 2019;570(7759):30-2.

133. Meijer I, Costas R, Zahedi Z, Wouters P. The Value of Research Data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report. avr 2013.

134. Kratz J, Strasser C. Data publication consensus and controversies. F1000Research. 16 oct 2014;3:94.

135. Green T. We need publishing standards for datasets and data tables. Learn Publ. 2009;22(4):325-7.

136. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M, ed. San Diego CA: FORCE11; 2014.

137. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 15 mar 2016;3(1):1-9.

138. Chavan VS, Ingwersen P. Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. BMC Bioinformatics. nov 2009;10(S14):S2.

139. Onsrud H, Campbell J. Big Opportunities in Access to « Small Science » Data. Data Sci J. 2007;6:OD58-66.

140. Chavan V, Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics. 15 déc 2011;12(15):S2.

141. Kennedy DN, Ascoli GA, De Schutter E. Next Steps in Data Publishing. Neuroinformatics. 1 déc 2011;9(4):317-20.

142. Callaghan S, Donegan S, Pepler S, Thorley M, Cunningham N, Kirsch P, et al. Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. Int J Digit Curation. 10 mar 2012;7(1):107-13.

143. Schöpfel J, Farace D, Prost H, Zane A. Data Papers as a New Form of Knowledge Organization in the Field of Research Data. Knowl Organ. 2019;46(8):622-38.

144. Candela L, Castelli D, Manghi P, Tani A. Data journals: A survey. J Assoc Inf Sci Technol. 2015;66(9):1747-62.

145. L'Hostis D, Hamelin M, Lelievre V, Aventurier P. Publier un Data Paper pour valoriser ses données. Support de formation Infodoc Express. octobre 2016.

146. Li K, Greenberg J, Dunic J. Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. J Assoc Inf Sci Technol. 2020;71(2):172-82.

147. Smith M. Data Papers in the Network Era. Proceedings of the Charleston Library Conference; 2011.

148. Li K, Chen P-Y. The narrative structure as a citation context in data papers: A preliminary analysis of Scientific Data. Proc Assoc Inf Sci Technol. 2018;55(1):856-8.

149. Chen Y-N. An analysis of characteristics and structures embedded in data papers: a preliminary study. Libellarium J Res Writ Books Cult Herit Inst. 31 déc 2016;9:145-56.

150. Reymonet N. Améliorer l'exposition des données de la recherche : la publication de data papers. 2017.

151. Cloern J, Soranno P. ASLO Takes a Next Step toward Open Science: Introducing Data Papers, a New Article Type in Limnology & Oceanography Letters. Limnol Oceanogr Bull. 2019;28(4):142-3.

152. Newman P, Corke P. Data Papers - Peer Reviewed Publication of High Quality Data Sets. J Robot Res. 1 mai 2009;28:587.

153. Blanc AK, Ngo TD. Data Papers. Stud Fam Plann. 5 avr 2019.

154. Friedman R, Psaki S, Bingenheimer JB. Announcing a New Journal Section: Data Papers. Stud Fam Plann. 2017;48(3):291-2.

155. Le Deuff O. Une nouvelle rubrique pour la RFSIC : Le Data Paper. Rev Fr Sci L'information Commun [Internet]. 31 déc 2018 [cité 12 juin 2020];(15). Disponible sur: http://journals.openedition.org/rfsic/5275

156. Special Collection of Data Papers from Long-term Watershed Research. CSA News. 2014;59(8):10-1.

157. García-García A, López-Borrull A, Peset F. Data journals: eclosión de nuevas revistas especializadas en datos. El Prof Inf. 1 déc 2015;24(6):845-54.

158. Hrynaszkiewicz I, SHINTANI Y. Scientific Data : An open access and open data publication to facilitate reproducible research. J Inf Process Manag. 1 janv 2014;57:629-40.

159. Mourouzis I, Pantos C. The Biomedical Data Journal in the New Era of Open Access to Scientific Information. Biomed Data J. 2015; 1(1): 5-10.

160. De Schutter E. Data Publishing and Scientific Journals: The Future of the Scientific Paper in a World of Shared Data. Neuroinformatics. 1 oct 2010;8(3):151-3.

161. Parsons M, Fox P. Is Data Publication the Right Metaphor?. Data Sci J. 31 janv 2013;12(0):WDS32-46.

162. Mayernik MS, Callaghan S, Leigh R, Tedds J, Worley S. Peer Review of Datasets: When, Why, and How. Bull Am Meteorol Soc. 1 fév 2015;96(2):191-201.

163. Borgman CL. The conundrum of sharing research data. J Am Soc Inf Sci Technol. 2012;63(6):1059-78.

164. Gorgolewski K, Margulies D, Milham M. Making Data Sharing Count: A Publication-Based Solution. Front Neurosci. 6 fév 2013;7:9.

165. Parsons MA, Duerr R, Minster J-B. Data Citation and Peer Review. Eos Trans Am Geophys Union. 2010;91(34):297-8.

166. Poline J-B. From data sharing to data publishing. MNI Open Res. 31 janv 2019;2:1.

167. Sandercock PA, Niewada M, Członkowska A, the International Stroke Trial Collaborative Group. The International Stroke Trial database. Trials. 21 avr 2011;12(1):101.

168. Poldrack RA, Gorgolewski KJ. OpenfMRI: Open sharing of task fMRI data. NeuroImage. 1 janv 2017;144:259-61.

169. International Neuroimaging Data-sharing Initiative [Internet]. [cité 10 juin 2020]. Disponible sur: http://fcon_1000.projects.nitrc.org/

170. Mennes M, Biswal BB, Castellanos FX, Milham MP. Making data sharing work: The FCP/INDI experience. NeuroImage. 15 nov 2013;82:683-91.

171. Konkiel S. Tracking the impacts of data – beyond citations [Internet]. Our Research blog. 2014 [cité 9 juin 2020]. Disponible sur: https://blog.ourresearch.org/data-impact-metrics/

172. Robinson-García N, Jiménez-Contreras E, Torres-Salinas D. Analyzing data citation practices using the data citation index. J Assoc Inf Sci Technol. déc 2016;67(12):2964-75.

173. Peters I, Kraker P, Lex E, Gumpenberger C, Gorraiz J. Research data explored: an extended analysis of citations and altmetrics. Scientometrics. 1 mai 2016;107(2):723-44.

174. Park H, Wolfram D. An examination of research data sharing and re-use: implications for data citation practice. Scientometrics. 1 avr 2017;111(1):443-61.

175. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. PLOS ONE. 29 juin 2011;6(6):e21101.

176. Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, et al. An International Framework to Promote Access to Data. Science. 1 avr 2004;303:1777-8.

177. Kratz JE, Strasser C. Making data count. Sci Data. 4 août 2015;2(1):150039.

178. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ. 1 oct 2013;1:e175.

179. Leitner F, Bielza C, Hill SL , Larrañaga P. Data Publications Correlate with Citation Impact. Front. Neurosci. 13 sept 2016;10:419.

180. MacRoberts MH, MacRoberts BR. Problems of citation analysis: A study of uncited and seldom-cited influences. J Am Soc Inf Sci Technol. 2010;61(1):1-12.

181. Force MM, Robinson NJ. Encouraging data citation and discovery with the Data Citation Index. J Comput Aided Mol Des. 1 oct 2014;28(10):1043-8.

182. Torres-Salinas D, Jiménez-Contreras E, Robinson-García N. How many citations are there in the Data Citation Index?. Proceedings of 19th International Conference on Science and Technology Indicators; 3-5 sept 2014; Leiden, The Netherlands.

183. Torres-Salinas D, Martín-Martín A, Fuente-Gutiérrez E. Analysis of the coverage of the Data Citation Index – Thomson Reuters: disciplines, document types and repositories. Rev Esp Doc Científica. 30 mar 2014;37(1):e036.

184. Chao TC. Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. Proc Am Soc Inf Sci Technol. 2011;48(1):1-8.

185. Callahan A, Winnenburg R, Shah NH. U-Index, a dataset and an impact metric for informatics tools and databases. Sci Data. 20 mar 2018;5(1):180043.

186. Ajiferuke I, Lu K, Wolfram D. A comparison of citer and citation-based measure outcomes for multiple disciplines. J Am Soc Inf Sci Technol. 2010;61(10):2086-96.

187. Melero R. Altmetrics – a complement to conventional metrics. Biochem Medica. 2015;25(2):152-60.

188. Bornmann L. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. J Informetr. 1 oct 2014;8(4):895-903.

189. Sud P, Thelwall M. Evaluating altmetrics. Scientometrics. 1 fév 2014;98(2):1131-43.

190. Priem J, Piwowar H, Hemminger B. Altmetrics in the wild: Using social media to explore scholarly impact. 20 mar 2012.

191. Priem J, Hemminger B. Scientometrics 2.0: New metrics of scholarly impact on the social Web. First Monday. 8 juil 2010;15.

192. Konkiel S. Altmetrics a 21 st-century solution to determining research quality. Online. 1 juil 2013;37:10-5.

193. Zahedi Z, Costas R, Wouters P. How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics. 1 nov 2014;101(2):1491-513.

194. Powell AGMT, Bevan V, Brown C, Lewis WG. Altmetric Versus Bibliometric Perspective Regarding Publication Impact and Force. World J Surg. 1 sept 2018;42(9):2745-56.

195. Azer SA, Azer S. Top-cited articles in medical professionalism: a bibliometric analysis versus altmetric scores. BMJ Open. juil 2019;9(7):e029433.

196. Elmore SA. The Altmetric Attention Score: What Does It Mean and Why Should I Care?. Toxicol Pathol. 1 avr 2018;46(3):252-5.

197. Costas R, Zahedi Z, Wouters P. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. J Assoc Inf Sci Technol. oct 2015;66(10):2003-19.

198. Ortega JL. Disciplinary differences of the impact of altmetric. FEMS Microbiol Lett. 1 avr 2018;365(7).

199. Thelwall M, Haustein S, Larivière V, Sugimoto CR. Do Altmetrics Work? Twitter and Ten Other Social Web Services. Bornmann L, ed. PLoS ONE. 28 mai 2013;8(5):e64841.

200. Haustein S, Peters I, Sugimoto CR, Thelwall M, Larivière V. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. J Assoc Inf Sci Technol. 2014;65(4):656-69.

201. Haustein S, Siebenlist T. Applying social bookmarking data to evaluate journal usage. J Informetr. 1 juil 2011;5(3):446-57.

202. Li X, Thelwall M. F1000, Mendeley and traditional bibliometric indicators. In: E. Archambault, Y. Gingras, V. Larivière, eds. Proceedings of the 17th International Conference on Science and Technology Indicators ; 2012 ; Montreal, Canada.

203. Tonia T, Van Oyen H, Berger A, Schindler C, Künzli N. If I tweet will you cite? The effect of social media exposure of articles on downloads and citations. Int J Public Health. 1 mai 2016;61(4):513-20.

204. Eysenbach G. Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. J Med Internet Res. 2011;13(4):e123.

205. Klar S, Krupnikov Y, Ryan JB, Searles K, Shmargad Y. Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work. PLOS ONE. 6 avr 2020;15(4):e0229446.

206. Paradis N, Knoll MA, Shah C, Lambert C, Delouya G, Bahig H, et al. Twitter: A Platform for Dissemination and Discussion of Scientific Papers in Radiation Oncology. Am J Clin Oncol. juin 2020;43(6):442–445.

207. Robinson-Garcia N, Torres-Salinas D, Zahedi Z, Costas R. New data, new possibilities: Exploring the insides of Altmetric.com. El Prof Inf. 1 juil 2014;23:359-66.

208. Haustein S, Costas R, Larivière V. Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns. PLOS ONE. 17 mar 2015;10(3):e0120495.

209. Torres-Salinas D, Cabezas-Clavijo Á, Jiménez-Contreras E. Altmetrics: nuevos indicadores para la comunicación científica en la Web 2.0. Comun Rev Científica Comun Educ. 2013;21(41):53-60.

210. Ingwersen P, Chavan V. Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. BMC Bioinformatics. 15 déc 2011;12(15):S3.

211. Weber N, Thomer A, Mayernik M, Dattore B, Ji Z, Worley S. The Product and System Specificities of Measuring Curation Impact. Int J Digit Curation. 21 nov 2013;8:223-34.

212. CODATA-ICSTI Task Group on Data Citation Standards and Practices. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Sci J. 2013;12: 1-75.

213. National Research Council, Policy and Global Affairs, Board on Research Data and Information. For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Uhlir PE, ed. Washington, D.C: National Academies Press; 2012. 238 p.

214. Mayernik MS. Session summary: The RDAP12 data citation panel practitioners. Bull Am Soc Inf Sci Technol. 2012;38(5):31-31.

215. Mooney H, Newton M. The Anatomy of a Data Citation: Discovery, Reuse, and Credit. J Librariansh Sch Commun. 18 mai 2012;1.

216. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.0. 2016;45 p.

217. Journal of Open Psychology Data [Internet]. [cité 11 juin 2020]. Disponible sur: http://openpsychologydata.metajnl.com/about/submissions/

218. Open Journal of Bioresources [Internet]. [cité 11 juin 2020]. Disponible sur: http://openbioresources.metajnl.com/about/submissions/

219. Open Health Data [Internet]. [cité 11 juin 2020]. Disponible sur: http://openhealthdata.metajnl.com/about/submissions/

Annexes

1 Liste des revues publiant des data papers

➔ Les 8 projets ayant listé des revues qui publient des *data papers* :

Akers K. A Growing List of Data Journals [Internet]. Data@MLibrary. 2014 [cité 3 juill 2020]. Disponible sur: https://mlibrarydata.wordpress.com/2014/05/09/data-journals/

Candela L, Castelli D, Manghi P, Tani A. Data journals: A survey. J Assoc Inf Sci Technol. 2015;66(9):1747-62.

Dedieu L. Revues publiant des Data papers. 2018. Disponible sur: https://collaboratif.cirad.fr/alfresco/s/d/workspace/SpacesStore/a9da20c0-b4ae-4f1d-ab52-0dab82fed1ec/Revues%20publiant%20des%20datapapers%20-janvier%202018.pdf

Finnish Committee for Research Data. Data journal directory [Internet]. [cité 3 juill 2020]. Disponible sur: https://www.fcrd.fi/data-journal-directory/

Forschungsdaten. Data journals [Internet]. [cité 3 juill 2020]. Disponible sur: https://www.forschungsdaten.org/index.php/Data_Journals

Institut National de recherche pour l'Agriculture, l'Alimentation et l'Environnement. Datapartage - Data Journals et facteur d'impact ? [Internet]. [cité 3 juill 2020]. Disponible sur: https://www6.inrae.fr/datapartage/Partager-Publier/Publier-un-Data-Paper/FAQ-Datapaper/Data-Journals-et-facteur-d-impact

PREPARDE. A list of Data Journals [Internet]. [cité 3 juill 2020]. Disponible sur: http://artefacts.ceda.ac.uk/frozen_sites/preparde/preparde/blog/DataJournalsList.html

University of Strathclyde. Data Journals [Internet]. [cité 3 juill 2020]. Disponible sur: https://www.strath.ac.uk/openaccess/researchdatamanagment/datajournals/

➔ La liste des 193 revues publiant des data papers issue de la fusion des 8 projets sus mentionnés :

Access Microbiology
Acta Crystallographica E: Structure Reports Online
African Invertebrates
Annals of Forest Science
AoB Plants
Applied Informatics
Applied Vegetation Science
Aquatic Biosystems
Aquatic Data
Atomic Data and Nuclear Data Tables
Biodiscovery
Biodiversity Data Journal
Bioinvasion Records
Biology of Sex Differences
Biomedical Data Journal
BioRisk
BMC Anesthesiology
BMC Biochemistry
BMC Bioinformatics
BMC Biology
BMC Biophysics
BMC Biotechnology
BMC Cancer
BMC Cardiovascular Disorders
BMC Cell Biology
BMC Clinical Pathology
BMC Complementary & Alternative Medicine
BMC Dermatology
BMC Developmental Biology
BMC Ear, Nose and Throat Disorders
BMC Ecology
BMC Emergency Medicine
BMC Endocrine Disorders
BMC Evolutionary Biology
BMC Family Practice
BMC Gastroenterology
BMC Genetics
BMC Genomics
BMC Geriatrics
BMC Health Services Research
BMC Hematology
BMC Immunology
BMC Infectious Diseases
BMC International Health and Human Rights
BMC Medical Education
BMC Medical Ethics
BMC Medical Genetics
BMC Medical Genomics
BMC Medical Imaging
BMC Medical Informatics & Decision Making
BMC Medical Physics
BMC Medical Research Methodology
BMC Medicine
BMC Microbiology
BMC Molecular Biology
BMC Musculoskeletal Disorders
BMC Nephrology
BMC Neurology
BMC Neuroscience
BMC Nursing
BMC Obesity
BMC Ophthalmology
BMC Oral Health
BMC Palliative Care
BMC Pediatrics
BMC Pharmacology & Toxicology
BMC Physiology
BMC Plant Biology
BMC Pregnancy & Childbirth
BMC Psychiatry
BMC Psychology
BMC Public Health
BMC Pulmonary Medicine
BMC Research Notes
BMC Sports Science, Medicine and Rehabilitation
BMC Structural Biology
BMC Surgery
BMC Systems Biology
BMC Urology
BMC Veterinary Research
BMC Women's Health
BMC Zoology
Botanical Studies
Cahier Agricultures
Cell & Bioscience

Chemical Data Collections
Chemistry Central Journal
Chiropractic & Manual Therapies
Cybergeo (revue européenne de géographie)
Data
Data in brief
Data Science Journal (CODATA)
Database - The Journal of Biological Databases and Curation
Dataset Papers in Science
Diagnostic Pathology
Earth perspective
Earth System Science Data
Ecological Research
EcologicalArchives
Ecology
Economie Internationale (AV
Economics : the open access open assessment E journal
Economics letters
eLife
European Data Watch
European Journal of Lipid Science and Technology
EvoDevo
F1000 Research
Forest Ecosystems
Freshwater Metadata Journal
Frontiers in Cell and Developmental Biology
Frontiers in Cellular and Infection Microbiology
Frontiers in Environmental Science
Frontiers in Microbiology
Frontiers in Plant Science
Frontiers in Sociology
Genome Announcements / Microbiology ressource annoncements
Genome Medicine
Genomics Data
Geo: Geography and Environment
Geochemistry, Geophysics, Geosystems
Geoscience Data Journal
Geoscientific Model Development
GigaScience
Global Ecology and Biogeography
Health and Justice
Health Information Science and Systems
Human Genomics
In Silico Pharmacology
International Economics
International Journal of Epidemiology
International Journal of Food Contamination
International Journal of Robotic Research
Internet Archaeology
Irish Veterinary Journal
Journal of Applied Volcanology
Journal of Biomedical Semantics
Journal of Chemical and Engineering Data
Journal of Cheminformatics
Journal of Clinical Bioinformatics
Journal of Environmental Quality
Journal of Hymenoptera research
Journal of Occupational Medicine and Toxicology
Journal of Open Archaeology Data
Journal of open bioresources
Journal of Open Health Data
Journal of open humanities data
Journal of Open Psychology Data
Journal of Open Public health Data
Journal of Open Research Software
Journal of Physical and Chemical Reference Data
Journal of Physical and Chemical Research Data
Journal of Statistical Software
Journal of Systems Chemistry
Journal of the International Society of Sports Nutrition
Journal of Vegetation Science
Matters science
Microbial Genomics
Microbial Informatics and Experimentation

Molecular Ecology Resource
Movement Ecology
MycoKeys
Nature Biotechnology
Nature Conservation
NeoBiota
Neuroinformatics
Nuclear Data Sheets
Nucleic Acids Research
One Ecosystem
Open Archaeology Data
Open Data Journal for Agricultural Research
Open Geospatial Data, Software and Standards
Open Health Data
Open Journal of Bioresources
Open Network Biology
PhytoKeys
Plant and Cell Physiology
Plant Journal
Plant Methods
PLoS ONE
Proteomics
Renewable and Sustainable Energy Reviews
Research Data Journal for Humanities and Social Sciences
Research Ideas and Outcomes (RIO)
Scientific Data
SpringerPlus
Standards in Genomic Sciences
Substance Abuse Treatment, Prevention, and Policy
The astrophysics Journal : supplement series
Theoretical Biology and Medical Modelling
The Plant Phenome Journal
Water ressources research
ZooKeys

2 PlumX Metrics

URL : https://plumanalytics.com/learn/about-metrics/

https://blog.scopus.com/topics/plumx-metrics

PlumX Metrics provide insights into the ways people interact with individual pieces of research output (articles, conference proceedings, book chapters, and many more) in the online environment. Examples include, when research is mentioned in the news or is tweeted about. Collectively known as PlumX Metrics, these metrics are divided into five categories to help make sense of the huge amounts of data involved and to enable analysis by comparing like with like.

PlumX gathers and brings together appropriate research metrics for all types of scholarly research output.

We categorize metrics into 5 separate categories: Citations, Usage, Captures, Mentions, and Social Media.

The Five Categories:

Citations – This is a category that contains both traditional citation indexes such as Scopus, as well as citations that help indicate societal impact such as Clinical or Policy Citations.

Examples: citation indexes, patent citations, clinical citations, policy citations

Usage – A way to signal if anyone is reading the articles or otherwise using the research. Usage is the number one statistic researchers want to know after citations.

Examples: clicks, downloads, views, library holdings, video plays

 Captures – Indicates that someone wants to come back to the work. Captures can be an leading indicator of future citations.

Examples: bookmarks, code forks, favorites, readers, watchers

 Mentions – Measurement of activities such as news articles or blog posts about research. Mentions is a way to tell that people are truly engaging with the research.

Examples: blog posts, comments, reviews, Wikipedia references, news media

 Social media -This category includes the tweets, Facebook likes, etc. that reference the research. Social Media can help measure "buzz" and attention. Social media can also be a good measure of how well a particular piece of research has been promoted.

Examples: shares, likes, comments, tweets

Citation Metrics 

Citation counts in PlumX are measures of how many times your research has been cited by others. Including citation counts alongside the other modern metrics categories allows for side-by-side analysis. The following are the sources of citation counts that are currently in PlumX.

Metric	Source(s)	Description
Citation Indexes	Airiti Academic Citation Index	The number of Airiti ACI works that cite the artifact
Citation Indexes	CrossRef	The number of articles that cite the artifact according to CrossRef
Citation Indexes	PubMed Central	The number of PubMed Central articles that cite the artifact
Citation Indexes	PubMed Central Europe	The number of PubMed Central Europe articles that cite the artifact

Metric	Source(s)	Description
Citation Indexes	RePEc	The number of RePEc works that cite the artifact as computed by CiTEc
Citation Indexes	SciELO	The number of SciELO articles that cite the artifact
Citation Indexes	Scopus	The number of articles that cite the artifact according to Scopus
Citation Indexes	SSRN	The number of SSRN works that cite the artifact
Patent Citation	USPTO	The number of patents that reference the artifact according to the United States Patent and Trademark Office
Clinical Citation	Dynamed Plus Topics	The number of Dynamed Plus Topics that reference the artifact
Clinical Citation	PubMed Clinical Guidelines	The number of Clinical Guidelines from PubMed that reference the artifact
Clinical Citation	National Institute for Health and Care Excellence (NICE) – UK	The number of Clinical Guidelines from NICE that reference the artifact
Policy Citation	Policy document source lists curated by PlumX	The number of policy documents that reference an artifact

Usage Metrics

Article level usage metrics are the number one statistic that researchers want to know after their citation counts.

Is anyone reading our work?

Did anyone watch our videos?

PlumX is unique in combining artifact-level Usage data with other artifact-level metrics. Below is a listing of the current Usage metrics that PlumX supports, and the providers of the data.

Metric	Source(s)	Description
Abstract Views	Airiti Library, bepress, CABI, DSpace, EBSCO, ePrints, RePEc, SciELO, SSRN	The number of times the abstract of an article has been viewed
Clicks	bit.ly	The number of clicks of a URL

Metric	Source(s)	Description
Collaborators	GitHub	The number of collaborators of an artifact
Downloads	Airiti Library, bepress, Dryad, DSpace, EBSCO, ePrints, figshare, Github, Institutional Repositories, Pure (for select customers only), RePEc, Slideshare, SSRN	The number of times an artifact has been downloaded
Full Text Views	CABI, EBSCO, OJS Journals, PLOS, PubMedCentral (for PLOS articles only), SciELO	The number of times the full text of an article has been viewed
Holdings	WorldCat	The number of libraries that hold the book artifact
Link Outs	EBSCO	The number of times an outbound link has been clicked to a library catalog or link resolver
Plays	Vimeo, YouTube, SoundCloud	The number of times the video or audio has been played
Views	Dryad, figshare, Slideshare	The number of times the artifact has been viewed

Capture Metrics 

Captures track when end users bookmark, favorite, become a reader, become a watcher, etc. Captures indicate that someone wants to come back to the work. Captures are important because they are an early, leading indicator of future citations. Below is a table of the metrics sources that PlumX uses for capture metrics.

Metric	Source(s)	Description
Bookmarks	Delicious (historical only)	Number of times an artifact has been bookmarked
Favorites	Slideshare, SoundCloud, YouTube	The number of times the artifact has been marked as a favorite
Followers	GitHub	The number of times a person or artifact has been followed
Forks	Github	The number of times a repository has been forked
Readers	CiteULike, Goodreads, Mendeley, SSRN	The number of people who have added the artifact to their library/briefcase

Metric	Source(s)	Description
Exports/Saves	EBSCO, SSRN	This includes the number of times an artifact's citation has been exported direct to bibliographic management tools or as file downloads, and the number of times an artifact's citation/abstract and HTML full text (if available) have been saved, emailed or printed.
Subscribers	Vimeo, YouTube	The number of people who have subscribed for an update
Watchers	Github	The number of people watching the artifact for updates

Mention Metrics 

Mentions are the blog posts, comments, reviews, and wikipedia links about your research. This category measures when people are truly engaging with your research. Mentions are where the stories of how people are interacting with research can be discovered. The PlumX platform automatically uncovers mentions. Below is a listing of the sources of mentions that PlumX monitors.

Metric	Source(s)	Description
Blog Mentions	Blog lists curated by PlumX	The number of blog posts written about the artifact
Comments	Reddit, Slideshare, Vimeo, YouTube	The number of comments made about an artifact
Economic Blog Mentions	Blog lists curated by PlumX	The number of blog posts written about the artifact within the economics discipline
Forum Topic Count	Vimeo	The number of topics in a forum discussing the artifact
Gist Count	GitHub	The number of gists in the source code repository
News Mentions	News source lists curated by PlumX	The number of news articles written about the artifact
Q&A Site Mentions	Stack Exchange	The number of mentions found about an artifact
References	Wikipedia	The number of references found to the artifact
Reviews	Amazon, Goodreads, SourceForge	The number of reviews written about the artifact

Social Media Metrics 

Social media metrics are the +1s, likes, shares, and tweets about research.

By tracking social media metrics, you can see how well a researcher is promoting their work. This is especially important for early career researchers to measure and understand who is interacting with their work. Of course, social media also allows us to track the buzz and attention surrounding research.

The following table lists the sources that PlumX tracks for Social Media.

Metric	Source(s)	Description
Likes	Vimeo, YouTube	The number of times an artifact has been liked
Shares, Likes & Comments	Facebook	The number of times a link was shared, liked or commented on
Ratings	Amazon, Goodreads, SourceForge	The average user rating of the artifact.
Recommendations	Figshare, SourceForge	The number of recommendations an artifact has received
Scores	Reddit	The number of upvotes minus downvotes on Reddit
Tweets	Twitter via Gnip	The number of tweets and retweets that mention the artifact

3 Nombre de data papers par journal

Revue	Typologie	N article
Data in brief	Data journal	424
Scientific Data	Data journal	95
Proteomics	Revue mixte	34
Open Journal of Bioresources	Data journal	28
International Journal of Epidemiology	Revue mixte	16
Journal of Open Health Data	Data journal	14
BMC Research Notes	Revue mixte	13
GigaScience	Revue mixte	9
Journal of Open Psychology Data	Data journal	9
Neuroinformatics	Revue mixte	9
F1000 Research	Revue mixte	8
Biomedical Data Journal	Data journal	6
BMC Medical Informatics & Decision Making	Revue mixte	6
BMC Nephrology	Revue mixte	5
BMC Bioinformatics	Revue mixte	4
BMC Infectious Diseases	Revue mixte	4
BMC Musculoskeletal Disorders	Revue mixte	4
BMC Psychiatry	Revue mixte	4
BMC Dermatology	Revue mixte	3
BMC Pregnancy & Childbirth	Revue mixte	3
Data	Data journal	3
BMC Cancer	Revue mixte	2
BMC Neurology	Revue mixte	2
BMC Public Health	Revue mixte	2
Dataset Papers in Science	Data journal	2
Journal of Biomedical Semantics	Revue mixte	2
Biology of Sex Differences	Revue mixte	1
BMC Anesthesiology	Revue mixte	1
BMC Complementary & Alternative Medicine	Revue mixte	1
BMC Health Services Research	Revue mixte	1
BMC Immunology	Revue mixte	1
BMC International Health and Human Rights	Revue mixte	1
BMC Medical Education	Revue mixte	1
BMC Medical Research Methodology	Revue mixte	1
BMC Neuroscience	Revue mixte	1
BMC Oral Health	Revue mixte	1
BMC Pharmacology & Toxicology	Revue mixte	1
BMC Systems Biology	Revue mixte	1
Chiropractic & Manual Therapies	Revue mixte	1
Substance Abuse Treatment Prevention and Policy	Revue mixte	1
Applied Informatics	Revue mixte	0

Revue	Typologie	N article
Biodiscovery	Revue mixte	0
BMC Biochemistry	Revue mixte	0
BMC Biophysics	Revue mixte	0
BMC Biotechnology	Revue mixte	0
BMC Cardiovascular Disorders	Revue mixte	0
BMC Cell Biology	Revue mixte	0
BMC Clinical Pathology	Revue mixte	0
BMC Developmental Biology	Revue mixte	0
BMC Ear. Nose and Throat Disorders	Revue mixte	0
BMC Emergency Medicine	Revue mixte	0
BMC Endocrine Disorders	Revue mixte	0
BMC Evolutionary Biology	Revue mixte	0
BMC Family Practice	Revue mixte	0
BMC Gastroenterology	Revue mixte	0
BMC Geriatrics	Revue mixte	0
BMC Hematology	Revue mixte	0
BMC Medical Ethics	Revue mixte	0
BMC Medical Imaging	Revue mixte	0
BMC Medical Physics	Revue mixte	0
BMC Medicine	Revue mixte	0
BMC Molecular Biology	Revue mixte	0
BMC Nursing	Revue mixte	0
BMC Obesity	Revue mixte	0
BMC Ophthalmology	Revue mixte	0
BMC Palliative Care	Revue mixte	0
BMC Pediatrics	Revue mixte	0
BMC Physiology	Revue mixte	0
BMC Psychology	Revue mixte	0
BMC Pulmonary Medicine	Revue mixte	0
BMC Sports Science Medicine and Rehabilitation	Revue mixte	0
BMC Surgery	Revue mixte	0
BMC Urology	Revue mixte	0
BMC Womens Health	Revue mixte	0
Cell & Bioscience	Revue mixte	0
Data Science Journal (CODATA)	Revue mixte	0
European Journal of Lipid Science and Technology	Revue mixte	0
EvoDevo	Revue mixte	0
Frontiers in Cell and Developmental Biology	Revue mixte	0
Health and Justice	Revue mixte	0
Journal of Occupational Medicine and Toxicology	Revue mixte	0
Journal of the International Society of Sports Nutrition	Revue mixte	0
Nature Biotechnology	Revue mixte	0
Research Ideas and Outcomes (RIO)	Revue mixte	0
Theoretical Biology and Medical Modelling	Revue mixte	0

4 Revues supplémentaires identifiées via Pubmed

Aucune recherche spécifique n'a été réalisée dans ces revues. Nous n'avons pas vérifié comment ces revues identifiaient les *data paper,* ni si elles avaient publié d'autres *data papers* qui n'auraient pas été identifiés dans Pubmed.

Revue
Critical Care Medicine
Journal of Biomed and Health Informatics
Journal of Forensic and Legal Medicine
Medical Image Computing and Computer-Assisted Intervention
Medical Physics
Molecular Psychiatry
Ophthalmologe
Pediatric Rheumatology
Spinal cord
The Journal of Clinical Psychiatry
Transactions on Computational Biology and Bioinformatics

5 Liste des 163 data papers de neuroscience

Aine CJ, Bockholt HJ, Bustillo JR, Cañive JM, Caprihan A, Gasparovic C, et al. Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. Neuroinformatics. oct 2017;15(4):343-64.

Alexander LM, Escalera J, Ai L, Andreotti C, Febre K, Mangone A, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. Sci Data. 19 2017;4:170181.

Arboix A, Arbe G, García-Eroles L, Oliveres M, Parra O, Massons J. Infarctions in the vascular territory of the posterior cerebral artery: clinical features in 232 patients. BMC Res Notes. 7 sept 2011;4(1):329.

Auzias G, Brun L, Deruelle C, Coulon O. Benchmark data for sulcal pits extraction algorithms. Data Brief. déc 2015;5:595-8.

Avants BB, Duda JT, Kilroy E, Krasileva K, Jann K, Kandel BT, et al. The pediatric template of brain perfusion. Sci Data. 3 fév 2015;2(1):150003.

Baker SL, Maass A, Jagust WJ. Considerations and code for partial volume correcting [18F]-AV-1451 tau PET data. Data Brief. déc 2017;15:648-57.

Basak C, Qin S, Nashiro K, O'Connell MA. Functional magnetic neuroimaging data on age-related differences in task switching accuracy and reverse brain-behavior relationships. Data Brief. 1 août 2018;19:997-1007.

Bauman WA, Wecht JM, Biering-Sørensen F. International spinal cord injury endocrine and metabolic extended data set. Spinal Cord. mai 2017;55(5):466-77.

Belyk M, Brown S, Kotz SA. Demonstration and validation of Kernel Density Estimation for spatial meta-analyses in cognitive neuroscience using simulated data. Data Brief. août 2017;13:346-52.

Berkovich-Ohana A, Harel M, Hahamy A, Arieli A, Malach R. Data for default network reduced functional connectivity in meditators, negatively correlated with meditation expertise. Data Brief. sept 2016;8:910-4.

Boayue NM, Csifcsák G, Puonti O, Thielscher A, Mittner M. Head models of healthy and depressed adults for simulating the electric fields of non-invasive electric brain stimulation. F1000Research. 2018;7:704.

Book GA, Anderson BM, Stevens MC, Glahn DC, Assaf M, Pearlson GD. Neuroinformatics Database (NiDB)--a modular, portable database for the storage, analysis, and sharing of neuroimaging data. Neuroinformatics. oct 2013;11(4):495-505.

Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 3 mar 2016;3(1):160011.

Brantley JA, Luu TP, Nakagome S, Zhu F, Contreras-Vidal JL. Full body mobile brain-body imaging data during unconstrained locomotion on stairs, ramps, and level ground. Sci Data. 10 juil 2018;5(1):180133.

Brůha P, Mouček R, Vacek V, Šnejdar P, Černá K, Řehoř P. Collection of human reaction times and supporting health related data for analysis of cognitive and physical performance. Data Brief. 1 avr 2018;17:469-511.

Bunevicius A, Kazlauskas H, Raskauskiene N, Janusonis V, Bunevicius R. Thyroid Hormone and C-Reactive Protein Serum Concentrations, Disease Severity and Discharge Outcomes of Ischemic Stroke Patients: A Dataset. Biomed Data J. 1 jan 2015;1:13-8.

Cao F. fMRI data from Korean, Chinese and English subjects in a word rhyming judgment task. Data Brief. 9 mar 2016;7:591-4.

Cao F. Brain MRI data in Chinese dyslexic children performing an auditory rhyming judgment task. Data Brief. avr 2017;11:473-8.

Carass A, Roy S, Jog A, Cuzzocreo JL, Magrath E, Gherman A, et al. Longitudinal multiple sclerosis lesion segmentation data resource. Data Brief. 1 juin 2017;12:346-50.

Chang L, Akazawa K, Yamakawa R, Hayama S, Buchthal S, Alicata D, et al. Delayed early developmental trajectories of white matter tracts of functional pathways in preterm-born infants: Longitudinal diffusion tensor imaging data. Data Brief. mar 2016;6:1007-15.

Chen Y, Juhas M, Greenshaw AJ, Hu Q, Meng X, Cui H, et al. Data on the impact of SSRIs and depression symptoms on the neural activities in obsessive–compulsive disorder at rest. Data Brief. 1 juin 2016;8:324-8.

Cho H, Ahn M, Ahn S, Kwon M, Jun SC. EEG datasets for motor imagery brain-computer interface. GigaScience. 01 2017;6(7):1-8.

Condon DM, Revelle W. Selected ICAR Data from the SAPA-Project: Development and Initial Validation of a Public-Domain Measure. J Open Psychol Data. 26 jan 2016;4(1):e1.

Cserhati M, Pandey S, Beaudoin J, Baccaglini L, Guda C, Fox H. The National NeuroAIDS Tissue Consortium (NNTC) Database: an integrated database for HIV-related studies Downloaded from. Database J Biol Databases Curation. 3 août 2015;2015

Dasenbrock HH, Cote DJ, Pompeu Y, Vasudeva VS, Smith TR, Gormley WB. Validation of an International Classification of Disease, Ninth Revision coding algorithm to identify decompressive craniectomy for stroke. BMC Neurol. 26 juin 2017;17(1):121.

Datta A, Sze SK. Data for iTRAQ profiling of micro-vesicular plasma specimens: In search of potential prognostic circulatory biomarkers for Lacunar infarction. Data Brief. sept 2015;4:510-7.

Di Martino A, Yan C-G, Li Q, Denio E, Castellanos FX, Alaerts K, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry. juin 2014;19(6):659-67.

Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data. 14 mar 2017;4(1):170010.

Di Scala C, Mazzarino M, Yahi N, Varini K, Garmy N, Fantini J, et al. Anandamide-ceramide interactions in a membrane environment: Molecular dynamic simulations data. Data Brief. oct 2017;14:163-7.

Dickson PI, Kaitila I, Harmatz P, Mlikotic A, Chen AH, Victoroff A, et al. Data from subjects receiving intrathecal laronidase for cervical spinal stenosis due to mucopolysaccharidosis type I. Data Brief. déc 2015;5:71-6.

Dominik T, Dostál D, Zielina M, Šmahaj J, Sedláčková Z, Procházka R. EEG data and introspective reports from the Libet's experiment replication. Data Brief. 27 sept 2018;20:2040-4.

Ducharme S, Albaugh MD, Nguyen T-V, Hudziak JJ, Mateos-Pérez JM, Labbe A, et al. Trajectories of cortical surface area and cortical volume maturation in normal brain development. Data Brief. 10 nov 2015;5:929-38.

DuPre E, Luh W-M, Spreng RN. Multi-echo fMRI replication sample of autobiographical memory, prospection and theory of mind reasoning tasks. Sci Data. 20 déc 2016;3(1):160116.

Dvorak MF, Itshayek E, Fehlings MG, Vaccaro AR, Wing PC, Biering-Sorensen F, et al. International Spinal Cord Injury: Spinal Interventions and Surgical Procedures Basic Data set. Spinal Cord. fév 2015;53(2):155-65.

El-Ansary A. Data of multiple regressions analysis between selected biomarkers related to glutamate excitotoxicity and oxidative stress in Saudi autistic patients. Data Brief. juin 2016;7:111-6.

Emelyanov A, Andoskin P, Pchelina S. Dataset of total, oligomeric alpha-synuclein and hemoglobin levels in plasma in Parkinson's disease. Data Brief. 28 nov 2016;10:182-5.

Emmerling TC, Zimmermann J, Sorger B, Frost M, Goebel R. Time-resolved searchlight analysis of imagined visual motion using 7 T ultra-high field fMRI: Data on interindividual differences. Data Brief. juin 2016;7:468-71.

Esteban O, Zosso D, Daducci A, Bach-Cuadra M, Ledesma-Carbayo MJ, Thiran J-P, et al. Data on the verification and validation of segmentation and registration methods for diffusion MRI. Data Brief. 1 sept 2016;8:871-6.

Fan Q, Nummenmaa A, Wichtmann B, Witzel T, Mekkaoui C, Schneider W, et al. A comprehensive diffusion MRI dataset acquired on the MGH Connectome scanner in a biomimetic brain phantom. Data Brief. juin 2018;18:334-9.

Faraut MCM, Carlson AA, Sullivan S, Tudusciuc O, Ross I, Reed CM, et al. Dataset of human medial temporal lobe single neuron activity during declarative memory encoding and recognition. Sci Data. 13 fév 2018;5(1):180010.

Filevich E, Lisofsky N, Becker M, Butler O, Lochstet M, Martensson J, et al. Day2day: investigating daily variability of magnetic resonance imaging measures over half a year. BMC Neurosci. 24 août 2017;18(1):65.

Föcking M, Chen W-Q, Dicker P, Dunn MJ, Lubec G, Cotter DR. Proteomic analysis of human hippocampus shows differential protein expression in the different hippocampal subfields. Proteomics. août 2012;12(15-16):2477-81.

Forstmann BU, Keuken MC, Schafer A, Bazin P-L, Alkemade A, Turner R. Multi-modal ultra-high resolution structural 7-Tesla MRI data repository. Sci Data. 2014;1:140050.

Fracasso A, van Veluw SJ, Visser F, Luijten PR, Spliet W, Zwanenburg JJM, et al. Myelin contrast across lamina at 7T, ex-vivo and in-vivo dataset. Data Brief. 1 sept 2016;8:990-1003.

Franz D, Olsen HL, Klink O, Gimsa J. Automated and manual patch clamp data of human induced pluripotent stem cell-derived dopaminergic neurons. Sci Data. 25 2017;4:170056.

Gentillon H, Stefańczyk L, Strzelecki M, Respondek-Liberska M. Prenatal brain MRI samples for development of automatic segmentation, target-recognition, and machine-learning algorithms to detect anatomical structures. F1000Research. 8 sept 2017;6:93.

Gollub RL, Shoemaker JM, King MD, White T, Ehrlich S, Sponheim SR, et al. The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. Neuroinformatics. juil 2013;11(3):367-88.

González-Burguera I, Ricobaraza A, Aretxabala X, Barrondo S, García Del Caño G, López de Jesús M, et al. Data for the morphometric characterization of NT2-derived postmitotic neurons. Data Brief. juin 2016;7:1349-54.

Goren N, Avery J, Dowrick T, Mackle E, Witkowska-Wrobel A, Werring D, et al. Multi-frequency electrical impedance tomography and neuroimaging data in stroke patients. Sci Data. 3 juil 2018;5(1):180112.

Gorgolewski KJ, Durnez J, Poldrack RA. Preprocessed Consortium for Neuropsychiatric Phenomics dataset. F1000Research. 22 sept 2017;6:1262.

Gorgolewski KJ, Mendes N, Wilfling D, Wladimirow E, Gauthier CJ, Bonnen T, et al. A high resolution 7-Tesla resting-state fMRI test-retest dataset with cognitive and physiological measures. Sci Data. 2015;2:140054.

Gorgolewski KJ, Storkey A, Bastin ME, Whittle IR, Wardlaw JM, Pernet CR. A test-retest fMRI dataset for motor, language and spatial attention functions. GigaScience. 29 avr 2013;2(1):6.

Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics. oct 2012;10(4):331-9.

Hanke M, Adelhöfer N, Kottke D, Iacovella V, Sengupta A, Kaule FR, et al. A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. Sci Data. 25 oct 2016;3(1):160092.

Hanke M, Baumgartner FJ, Ibe P, Kaule FR, Pollmann S, Speck O, et al. A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. Sci Data. 27 mai 2014;1(1):140003.

Hanke M, Dinga R, Häusler C, Guntupalli JS, Casey M, Kaule FR, et al. High-resolution 7-Tesla fMRI data on the perception of musical genres – an extension to the studyforrest dataset. F1000Research. 29 juin 2015;4:174.

Hansson KT, Skillbäck T, Pernevik E, Kern S, Portelius E, Höglund K, et al. Expanding the cerebrospinal fluid endopeptidome. Proteomics. 2017;17(5).

Harteveld AA, Denswil NP, Van Hecke W, Kuijf HJ, Vink A, Spliet WGM, et al. Data on vessel wall thickness measurements of intracranial arteries derived from human circle of Willis specimens. Data Brief. 2 mai 2018;19:6-12.

Harvey BM, Dumoulin SO. Data describing the accuracy of non-numerical visual features in predicting fMRI responses to numerosity. Data Brief. 1 fév 2018;16:193-205.

He Y, Luu TP, Nathan K, Nakagome S, Contreras-Vidal JL. A mobile brain-body imaging dataset recorded during treadmill walking with a brain-computer interface. Sci Data. 24 avr 2018;5(1):180074.

Holmes AJ, Hollinshead MO, O'Keefe TM, Petrov VI, Fariello GR, Wald LL, et al. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. Sci Data. 7 juil 2015;2(1):150031.

Horn A. A structural group-connectome in standard stereotactic (MNI) space. Data Brief. déc 2015;5:292-6.

Huang L, Taicheng H, Zhen Z, Liu J. A test-retest dataset for assessing long-Term reliability of brain morphology and resting-state brain activity. Sci Data. 15 mar 2016;3:160016.

Jafari-Khouzani K, Elisevich KV, Patel S, Soltanian-Zadeh H. Dataset of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques. Neuroinformatics. déc 2011;9(4):335-46.

Jennum P, Iversen HK, Ibsen R, Kjellberg J. Cost of stroke: a controlled national study evaluating societal effects on patients and their partners. BMC Health Serv Res. 13 oct 2015;15:466.

Jensen M, Cox AP, Chaudhry N, Ng M, Sule D, Duncan W, et al. The neurological disease ontology. J Biomed Semant. 6 déc 2013;4(1):42.

Jimura K, Hirose S, Wada H, Yoshizawa Y, Imai Y, Akahane M, et al. Data for behavioral results and brain regions showing a time effect during pair-association retrieval. Data Brief. 1 sept 2016;8:891-3.

Kaya M, Binli MK, Ozbay E, Yanar H, Mishchenko Y. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. Sci Data. 16 oct 2018;5(1):180211.

Kennedy KM, Rieck JR, Boylan MA, Rodrigue KM. Functional magnetic resonance imaging data of incremental increases in visuo-spatial difficulty in an adult lifespan sample. Data Brief. avr 2017;11:54-60.

Keuken MC, Forstmann BU. A probabilistic atlas of the basal ganglia using 7 T MRI. Data Brief. 31 juil 2015;4:577-82.

Kodiweera C, Wu Y-C. Data of NODDI diffusion metrics in the brain and computer simulation of hybrid diffusion imaging (HYDI) acquisition scheme. Data Brief. juin 2016;7:1131-8.

Kondylis ED, Randazzo MJ, Alhourani A, Wozny TA, Lipski WJ, Crammond DJ, et al. High frequency activation data used to validate localization of cortical electrodes during surgery for deep brain stimulation. Data Brief. 1 mar 2016;6:204-7.

Koush Y, Ashburner J, Prilepin E, Sladky R, Zeidman P, Bibikov S, et al. Real-time fMRI data for testing OpenNFT functionality. Data Brief. oct 2017;14:344-7.

Kulaga-Yoskovitz J, Bernhardt BC, Hong S-J, Mansi T, Liang KE, van der Kouwe AJW, et al. Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset. Sci Data. 10 nov 2015;2(1):150059.

Kumar S, Cieplak P. CaspNeuroD: A knowledgebase of predicted caspase cleavage sites in human proteins related to neurodegenerative diseases. Database. 6 oct 2016;2016.

Lacerda EM, Bowman EW, Cliff JM, Kingdon CC, King EC, Lee J-S, et al. The UK ME/CFS Biobank for biomedical research on Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS) and Multiple Sclerosis. Open J Bioresour. 2017;4.

Langer N, Ho EJ, Alexander LM, Xu HY, Jozanovic RK, Henin S, et al. A resource for assessing information processing in the developing brain using EEG and eye tracking. Sci Data. 11 2017;4:170040.

Lanka P, Deshpande G. Resting state fMRI data from subjects scanned with the EPI-PACE (Echoplanar Imaging - Prospective Acquisition CorrEction) sequence. Data Brief. oct 2018;20:2072-5.

Lesjak Ž, Galimzianova A, Koren A, Lukin M, Pernuš F, Likar B, et al. A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. Neuroinformatics. 2018;16(1):51-63.

Li K, Jiang J, Qiu L, Yang X, Huang X, Lui S, et al. A multimodal MRI dataset of professional chess players. Sci Data. 1 sept 2015;2(1):150044.

Liao W, Hamel REG, Olde Rikkert MGM, Oosterveld SM, Aalten P, Verhey FRJ, et al. A profile of The Clinical Course of Cognition and Comorbidity in Mild Cognitive Impairment and Dementia Study (The 4C study): two complementary longitudinal, clinical cohorts in the Netherlands. BMC Neurol. 25 nov 2016;16(1):242.

Liew S-L, Anglin JM, Banks NW, Sondag M, Ito KL, Kim H, et al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci Data. 20 2018;5:180011.

Lin Q, Dai Z, Xia M, Han Z, Huang R, Gong G, et al. A connectivity-based test-retest dataset of multi-modal magnetic resonance imaging in young healthy adults. Sci Data. 27 oct 2015;2(1):150056.

Liu D, Dong Z, Zuo X, Wang J, Zang Y. Eyes-open/eyes-closed dataset sharing for reproducibility evaluation of resting state fMRI data analysis methods. Neuroinformatics. oct 2013;11(4):469-76.

Liu W, Wei D, Chen Q, Yang W, Meng J, Wu G, et al. Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. Sci Data. 14 fév 2017;4(1):170017.

Luciw MD, Jarocka E, Edin BB. Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. Sci Data. 25 nov 2014;1(1):140047.

Lüsebrink F, Sciarra A, Mattern H, Yakupov R, Speck O. T1-weighted in vivo human whole brain MRI dataset with an ultrahigh isotropic resolution of 250 μm. Sci Data. 14 2017;4:170032.

Maclaren J, Han Z, Vos SB, Fischbein N, Bammer R. Reliability of brain volume measurements: a test-retest dataset. Sci Data. 2014;1:140037.

Martins-de-Souza D, Guest PC, Guest FL, Bauder C, Rahmoune H, Pietsch S, et al. Characterization of the human primary visual cortex and cerebellum proteomes using shotgun mass spectrometry-data-independent analyses. Proteomics. fév 2012;12(3):500-4.

Martins-de-Souza D, Guest PC, Steeb H, Pietsch S, Rahmoune H, Harris LW, et al. Characterizing the proteome of the human dorsolateral prefrontal cortex by shotgun mass spectrometry. PROTEOMICS. 2011;11(11):2347-53.

McCray AT, Trevvett P, Frost HR. Modeling the autism spectrum disorder phenotype. Neuroinformatics. avr 2014;12(2):291-305.

McManus CA, Polden J, Cotter DR, Dunn MJ. Two-dimensional reference map for the basic proteome of the human dorsolateral prefrontal cortex (dlPFC) of the prefrontal lobe region of the brain. Proteomics. juil 2010;10(13):2551-5.

Meffert H, Hwang S, Nolan ZT, Chen G, Blair JR. BOLD data representing activation and connectivity for rare no-go versus frequent go cues. Data Brief. 1 juin 2016;7:66-70.

Mei N, Grossberg MD, Ng K, Navarro KT, Ellmore TM. A high-density scalp EEG dataset acquired during brief naps after a visual working memory task. Data Brief. juin 2018;18:1513-9.

Meng Y, Li G, Wang L, Lin W, Gilmore J. Discovering Cortical Folding Patterns in Neonatal Cortical Surfaces Using Large-Scale Dataset. In 2016. p. 10-8.

Metzler-Baddeley C, Caeyenberghs K, Foley S, Jones DK. Longitudinal data on cortical thickness before and after working memory training. Data Brief. juin 2016;7:1143-7.

Minjarez B, Calderón-González KG, Valero Rustarazo ML, Herrera-Aguirre ME, Labra-Barrios ML, Rincon-Limas DE, et al. Data set of interactomes and metabolic pathways of proteins differentially expressed in brains with Alzheimer′s disease. Data Brief. 1 juin 2016;7:1707-19.

Mouček R, Vařeka L, Prokop T, Štěbeták J, Brůha P. Event-related potential data from a guess the number brain-computer interface experiment on school children. Sci Data. 28 mar 2017;4(1):160121.

Mushtaq F, Guillen PP, Wilkie RM, Mon-Williams MA, Schaefer A. Feedback-related potentials in a gambling task with randomised reward. Data Brief. 1 mar 2016;6:378-85.

Nah Y, Shin N-Y, Yi S, Lee S-K, Han S. Data on subjective recollection effects reflected in large-scale functional connectivity patterns in postpartum women. Data Brief. 1 août 2018;19:1142-7.

Nichols BN, Mejino JL, Detwiler LT, Nilsen TT, Martone ME, Turner JA, et al. Neuroanatomical domain of the foundational model of anatomy ontology. J Biomed Semant. 8 jan 2014;5(1):1.

Nih LR, Moshayedi P, Llorente IL, Berg AR, Cinkornpumin J, Lowry WE, et al. Engineered HA hydrogel for stem cell transplantation in the brain: Biocompatibility data using a design of experiment approach. Data Brief. 24 nov 2016;10:202-9.

O'Connor D, Potler NV, Kovacs M, Xu T, Ai L, Pellman J, et al. The Healthy Brain Network Serial Scanning Initiative: a resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions. GigaScience. 01 2017;6(2):1-14.

Oikari LE, Okolicsanyi RK, Griffiths LR, Haupt LM. Data defining markers of human neural stem cell lineage potential. Data Brief. 19 fév 2016;7:206-15.

Opsahl JA, Vaudel M, Guldbrandsen A, Aasebø E, Van Pesch V, Franciotta D, et al. Label-free analysis of human cerebrospinal fluid addressing various normalization strategies and revealing protein groups affected by multiple sclerosis. Proteomics. avr 2016;16(7):1154-65.

Orban P, Madjar C, Savard M, Dansereau C, Tam A, Das S, et al. Test-retest resting-state fMRI in healthy elderly persons with a family history of Alzheimer's disease. Sci Data. 2015;2:150043.

Os HJA van, Ruigrok YM, Manniën J, Dijk EJ van, Koudstaal PJ, Luijckx GJ, et al. Dutch Parelsnoer Institute-Cerebrovascular Accident (CVA) Study: A Large Multicenter Clinical Biobank with Standardized Collection and Storage. Open J Bioresour. 19 juil 2018;5(0):8.

Panwalkar V, Schulte M, Lecher J, Stoldt M, Willbold D, Dingley AJ. Data describing the solution structure of the WW3* domain from human Nedd4-1. Data Brief. 1 sept 2016;8:605-12.

Pauli WM, Nili AN, Tyszka JM. A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei. Sci Data. 17 avr 2018;5(1):180063.

Peter F, Trilck M, Rabenstein M, Rolfs A, Frech MJ. Dataset in support of the generation of Niemann-Pick disease Type C1 patient-specific iPS cell lines carrying the novel NPC1 mutation c.1180T>C or the prevalent c.3182T>C mutation - Analysis of pluripotency and neuronal differentiation. Data Brief. juin 2017;12:123-31.

Ping L, Duong DM, Yin L, Gearing M, Lah JJ, Levey AI, et al. Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's Disease. Sci Data. 13 mar 2018;5(1):180036.

Pinho AL, Amadon A, Ruest T, Fabre M, Dohmatob E, Denghien I, et al. Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. Sci Data. 12 juin 2018;5(1):180105.

Poldrack RA, Congdon E, Triplett W, Gorgolewski KJ, Karlsgodt KH, Mumford JA, et al. A phenome-wide examination of neural and cognitive function. Sci Data. 6 déc 2016;3(1):160110.

Potvin O, Mouiha A, Dieumegarde L, Duchesne S, Alzheimer′s Disease Neuroimaging Initiative. FreeSurfer subcortical normative data. Data Brief. déc 2016;9:732-6.

Rangaprakash D, Dretsch MN, Yan W, Katz JS, Denney TS, Deshpande G. Hemodynamic response function parameters obtained from resting-state functional MRI data in soldiers with trauma. Data Brief. oct 2017;14:558-62.

Rangaprakash D, Wu G-R, Marinazzo D, Hu X, Deshpande G. Parameterized hemodynamic response function data of healthy individuals obtained from resting-state functional MRI in a 7T MRI scanner. Data Brief. 1 avr 2018;17:1175-9.

Rezaei M, Mohammadi H, Khazaie H. EEG/EOG/EMG data from a cross sectional study on psychophysiological insomnia and normal sleep subjects. Data Brief. 1 déc 2017;15:314-9.

Robbins K, Su K, Hairston WD. An 18-subject EEG data collection using a visual-oddball task, designed for benchmarking algorithms and headset performance comparisons. Data Brief. 1 fév 2018;16:227-30.

Rosenke M, Weiner KS, Barnett MA, Zilles K, Amunts K, Goebel R, et al. Data on a cytoarchitectonic brain atlas: effects of brain template and a comparison to a multimodal atlas. Data Brief. juin 2017;12:327-32.

S Cassoli J, Brandão-Teles C, G Santana A, H M F Souza G, Martins-de-Souza D. Ion Mobility-Enhanced Data-Independent Acquisitions Enable a Deep Proteomic Landscape of Oligodendrocytes. Proteomics. nov 2017;17(21).

Savazzi F, Isernia S, Jonsdottir J, Di Tella S, Pazzi S, Baglio F. Design and implementation of a Serious Game on neurorehabilitation: Data on modifications of functionalities along implementation releases. Data Brief. 1 oct 2018;20:864-9.

Scarpino M, Lanzo G, Lolli F, Carrai R, Moretti M, Spalletti M, et al. Data on multimodal approach for early poor outcome (Cerebral Performance Categories 3-5) prediction after cardiac arrest. Data Brief. août 2018;19:704-11.

Schauer G, Chang A, Schwartzman D, Rae CL, Iriye H, Seth AK, et al. Fractionation of parietal function in bistable perception probed with concurrent TMS-EEG. Sci Data. 16 août 2016;3(1):160065.

Schechter JR, Greene DJ, Koller JM, Black KJ. A revised method for measuring distraction by tactile stimulation. F1000Research. 2014;3:188.

Schumacher LV, Reisert M, Nitschke K, Egger K, Urbach H, Hennig J, et al. Data on the test-retest reproducibility of streamline counts as a measure of structural connectivity. Data Brief. 1 août 2018;19:1361-81.

Sengupta A, Kaule FR, Guntupalli JS, Hoffmann MB, Häusler C, Stadler J, et al. A studyforrest extension, retinotopic mapping and localization of higher visual areas. Sci Data. 25 oct 2016;3(1):160093.

Sengupta A, Yakupov R, Speck O, Pollmann S, Hanke M. Ultra high-field (7 T) multi-resolution fMRI data for orientation decoding in visual cortex. Data Brief. 24 mai 2017;13:219-22.

Serrao M, Chini G, Bergantino M, Sarnari D, Casali C, Conte C, et al. Dataset on gait patterns in degenerative neurological diseases. Data Brief. 1 févr 2018;16:806-16.

Shin J, von Lühmann A, Kim D-W, Mehnert J, Hwang H-J, Müller K-R. Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset. Sci Data. 13 fév 2018;5(1):180003.

Shiraishi N, Katayama A, Nakashima T, Yamada S, Uwabe C, Kose K, et al. Three-dimensional morphology of the human embryonic brain. Data Brief. 1 sept 2015;4:116-8.

St John-Williams L, Blach C, Toledo JB, Rotroff DM, Kim S, Klavins K, et al. Targeted metabolomics and medication classification data from participants in the ADNI1 cohort. Sci Data. 17 2017;4:170140.

Stancak A, Cook S, Wright H, Fallon N. Data to support observation of late and ultra-late latency components of cortical laser evoked potentials. Data Brief. déc 2015;5:1031-4.

Sweeney-Reed CM, Zaehle T, Voges J, Schmitt FC, Buentjen L, Kopitzki K, et al. Clinical, neuropsychological, and pre-stimulus dorsomedial thalamic nucleus electrophysiological data in deep brain stimulation patients. Data Brief. sept 2016;8:557-61.

Tam A, Dansereau C, Badhwar A, Orban P, Belleville S, Chertkow H, et al. A dataset of multiresolution functional brain parcellations in an elderly population with no or mild cognitive impairment. Data Brief. déc 2016;9:1122-9.

Tegeder I. Yeast-2-Hybrid data file showing progranulin interactions in human fetal brain and bone marrow libraries. Data Brief. déc 2016;9:1060-2.

Tiberti N, Sanchez J-C. Comparative analysis of cerebrospinal fluid from the meningo-encephalitic stage of T. b. gambiense and rhodesiense sleeping sickness patients using TMT quantitative proteomics. Data Brief. sept 2015;4:400-5.

Tjew-A-Sin M, Tops M, Heslenfeld DJ, Koole SL. Data on simulated interpersonal touch, individual differences and the error-related negativity. Data Brief. 1 juin 2016;7:1327-30.

Tsapanou A, Gu Y, O'Shea DM, Yannakoulia M, Kosmidis MH, Dardiotis E, et al. Dataset on the associations between sleep quality/duration and cognitive performance in cognitively healthy older adults. Data Brief. oct 2017;14:720-3.

Tullo S, Devenyi GA, Patel R, Park MTM, Collins DL, Chakravarty MM. Warping an atlas derived from serial histology to 5 high-resolution MRIs. Sci Data. 19 juin 2018;5(1):180107.

Turi Z, Schäfer SA, Antal A, Paulus W, Mittner M. Data from 'Placebo Enhances Reward Learning in Healthy Individuals'. J Open Psychol Data. 20 avr 2018;6(1):2.

Van der Meer J, Pampel A, van Someren E, Ramautar J, van der Werf Y, Gomez-Herrero G, et al. « Eyes Open - Eyes Closed » EEG/fMRI data set including dedicated « Carbon Wire Loop » motion detection channels. Data Brief. juin 2016;7:990-4.

Van der Zwaag W, Buur PF, Fracasso A, van Doesum T, Uludağ K, Versluis M, et al. Examples of sub-millimeter, 7T, T1-weighted EPI datasets acquired with the T123DEPI sequence. Data Brief. 16 août 2018;20:415-8.

Vařeka L, Brůha P, Mouček R. Event-related potential datasets based on a three-stimulus paradigm. GigaScience. 1 déc 2014;3:35.

Vareka L, Bruha P, Moucek R, Mautner P, Cepicka L, Holecková I. Developmental coordination disorder in children - experimental work and data annotation. GigaScience. 01 2017;6(4):1-6.

Vialaret J, Schmit P-O, Lehmann S, Gabelle A, Wood J, Bern M, et al. Identification of multiple proteoforms biomarkers on clinical samples by routine Top-Down approaches. Data Brief. 31 mar 2018;18:1013-21.

Villena-González M, López V, Rodríguez E. Data of ERPs and spectral alpha power when attention is engaged on visual or verbal/auditory imagery. Data Brief. juin 2016;7:882-8.

Wakeman DG, Henson RN. A multi-subject, multi-modal human neuroimaging dataset. Sci Data. 20 jan 2015;2(1):150001.

Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 11 sept 2018;5(1):180185.

Wang SH, Lobier M, Siebenhühner F, Puoliväli T, Palva S, Palva JM. Hyperedge bundling: Data, source code, and precautions to modeling-accuracy bias to synchrony estimates. Data Brief. juin 2018;18:262-75.

Watson PD, Paul EJ, Cooke GE, Ward N, Monti JM, Horecka KM, et al. Cognitive and anatomical data in a healthy cohort of adults. Data Brief. 5 avr 2016;7:1221-7.

Wei D, Zhuang K, Ai L, Chen Q, Yang W, Liu W, et al. Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset. Sci Data. 17 juil 2018;5(1):180134.

Widerström-Noga E, Biering-Sørensen F, Bryce TN, Cardenas DD, Finnerup NB, Jensen MP, et al. The International Spinal Cord Injury Pain Basic Data Set (version 2.0). Spinal Cord. avr 2014;52(4):282-6.

Wiens S, Szychowska M, Eklund R, Nilsson ME. Data on the auditory duration mismatch negativity for different sound pressure levels and visual perceptual loads. Data Brief. avr 2017;11:159-64.

Wilke M. A spline-based regression parameter set for creating customized DARTEL MRI brain templates from infancy to old age. Data Brief. fév 2018;16:959-66.

Willén RM, Granhag PA. Data from Interviews with 95 Respondents Recollecting Repeated Dental Visits. J Open Psychol Data. 23 nov 2015;3(1):e7.

Wu D, Chang L, Akazawa K, Oishi K, Skranes J, Ernst T, et al. Change-point analysis data of neonatal diffusion tensor MRI in preterm and term-born infants. Data Brief. juin 2017;12:453-8.

Xiao Y, Fonov V, Chakravarty MM, Beriault S, Al Subaie F, Sadikot A, et al. A dataset of multi-contrast population-averaged brain MRI atlases of a Parkinson's disease cohort. Data Brief. juin 2017;12:370-9.

Yadollahpour A, Bayat A, Rashidi S, Saki N, Karimi M. Dataset of acute repeated sessions of bifrontal transcranial direct current stimulation for treatment of intractable tinnitus: A randomized controlled trial. Data Brief. 13 sept 2017;15:40-6.

Yadollahpour A, Mayo M, Saki N, Rashidi S, Bayat A. A chronic protocol of bilateral transcranial direct current stimulation over auditory cortex for tinnitus treatment: Dataset from a double-blinded randomized controlled trial. F1000Research. 2018;7:733.

Zhang Y, Guo Z, Zou L, Yang Y, Zhang L, Ji N, et al. Data for a comprehensive map and functional annotation of the human cerebrospinal fluid proteome. Data Brief. 1 juin 2015;3:103-7.

Zhang Y, Wei H, Cronin MJ, He N, Yan F, Liu C. Longitudinal data for magnetic susceptibility of normative human brain development and aging over the lifespan. Data Brief. oct 2018;20:623-31.

Zuo X-N, Anderson JS, Bellec P, Birn RM, Biswal BB, Blautzik J, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci Data. 9 déc 2014;1(1):140049.

Zwoliński P, Roszkowski M, Zygierewicz J, Haufe S, Nolte G, Durka PJ. Open database of epileptic EEG with MRI and postoperational assessment of foci--a real world verification for the EEG inverse solutions. Neuroinformatics. déc 2010;8(4):285-99.

AUTEUR : Nom : PARENT **Prénom : Audrey**

Date de soutenance : 9 octobre 2020

Titre de la thèse : Valorisation des jeux de données issus de la recherche médicale : la publication de *data papers*, analyse bibliométrique et score *altmetrics*. Application aux données de la neuroscience.

Thèse - Médecine - Lille 2020

Cadre de classement : Médecine

DES + spécialité : Santé publique et médecine sociale

Mots-clés : *Open science*, *open data*, données de la recherche, citation de données, *altmetric*, communication scientifique, *data paper*

Résumé :

Contexte : Les données scientifiques sont devenues tout aussi importantes que le résultat de leur analyse. La réutilisation efficace des données rend la question de leur partage centrale mais leur publication est limitée par manque de reconnaissance spécifique. Le *data paper* semble pouvoir répondre en partie à ce problème. L'objectif de notre travail était d'étudier la publication des *data papers* dans les sciences de la santé et plus particulièrement en neuroscience.

Méthodes : Nous avons mené une étude empirique au sein des revues reconnues comme publiant des *data papers* et sélectionné les *data papers* du domaine médical publiés jusqu'au 31 octobre 2018. Les articles ont été classés selon leur spécialité médicale et leur thématique. Pour les *data papers* de neuroscience, des informations générales et spécifiques au contenu de l'article ont été relevés et l'accès aux données décrites testé. Nous avons recherché leur typologie d'indexation au sein des trois grandes bases bibliographiques (Pubmed, Web Of Science, Scopus). Enfin, les données de bibliométrie et les scores *altmetrics* disponibles sur la base Scopus ont été recueillis.

Résultats : 745 *data papers* médicaux ont été identifiés dont 163 en neuroscience. Leur publication a fortement augmenté ces dernières années mais reste hétérogène. 68,9% des données médicales décrites sont directement accessibles, essentiellement des données d'électrophysiologie et d'imagerie. Le nombre de citations annuelles augmente. 75% des *data papers* ont été cités pour une moyenne de 2,7 citations/an/article et un délai médian de citation à 7,9 mois (IC95% 6,7 – 10,3). Une activité *altmetric* a été retrouvés pour 53% des articles sur Twitter et pour 86% sur les sites de *bookmarking* sociaux. Des différences existent selon la revue.

Conclusion : Le *data paper* reste pour le moment un phénomène récent, à petite échelle et peu homogène. En attendant une reconnaissance spécifique des citations de données, le *data paper* semble être un moyen alternatif acceptable dont le nombre continuera probablement à croître dans les années à venir.

<u>Composition du Jury</u> :

Président :	Monsieur le Professeur Emmanuel CHAZARD
Assesseurs :	Monsieur le Professeur Stefan DARMONI Monsieur le Professeur Cristian PREDA Monsieur le Docteur Vincent CHOURAKI
Directeur de thèse :	Madame le Docteur Amélie LANSIAUX