Université de Montréal

Qui fait quoi ?
Analyse des libellés de contribution dans les articles savants

par
Benoit Macaluso

École de bibliothéconomie et des sciences de l'information
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences

en vue de l'obtention du grade de maîtrise
en sciences de l'information

Avril 2015

Résumé

Qui fait quoi au sein d'une équipe de recherche scientifique? Ce mémoire propose de définir les rôles de chacun des auteurs à l'aide d'une méthodologie originale. L'ordre des auteurs est un indicateur utile de l'importance relative de chacun des co-auteurs, mais il ne révèle pas la nature précise de leur rôle au sein de l'équipe. Ainsi, notre recherche démontre que l'utilisation des libellés de contributions aux articles scientifiques à des fins bibliométriques est possible et ce, même à grande échelle. Nos résultats montrent qu'il existe des différences dans les contributions apportées aux articles scientifiques et que ces différences s'appliquent aussi à la répartition du travail au sein des équipes de recherche selon le sexe des auteurs.

Mots-clés : Bibliométrie, Femme, Rôle, Contribution, Autorat, Activité scientifique

Abstract

Who does what in a scientific research team? This work proposes to define the roles of each author with an original methodology. The order of the authors is a useful indicator of the relative importance of each of the co-authors, but it does not reveal the precise nature of their respective role. Our research shows that using statements of contributions from scientific articles for bibliometric purposes is possible, even on a large scale. Our results show that there are differences in the contributions to scientific articles and these differences also apply to the distribution of work among research teams by gender.

Keywords : Bibliometric, Woman, Role, Contribution, Authorship, Gender, Scientific Activity

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles

API – Application Programming Interface

DOI – Digital Object Identifier

ICMJE – International Committee of Journal Medical Editors

IMRaD – Introduction, Methods, Results and Discussion

MSSQL – Microsoft SQL

NSF – National Science Foundation

OST – Observatoire des sciences et des technologies

PDF – Portable Document Format

PLoS – Public Library of Science

RIS – Research Information Systems

SSIS – SQL Server Integration Services

UNESCO – Organisation des Nations Unies pour l'Éducation, la Science et la Culture

URL – Uniform Resource Locator

XML – Extensible Markup Language

À Josianne, mon âme sœur.
Au prochain banquet!

Remerciements

Je considérais les études supérieures comme un défi, un jeu auquel je n'avais pas encore joué et qui m'apparaissait, après avoir collaboré aux études de plusieurs personnes de passage à l'OST, digne d'intérêt! Je suis très heureux d'avoir terminé la partie avec le sourire et le goût de poursuivre la saison, malgré les embûches et les déceptions.

Je veux remercier Jean-Pierre Robitaille, coordonnateur de l'OST, pour ses précieux conseils, les nombreuses discussions de corridor et la lecture attentive de mon mémoire. Je suis aussi très reconnaissant envers mon directeur de maîtrise Vincent Larivière; son support, autant financier que moral, a permis de réaliser ce projet dans des conditions optimales. Ses encouragements dignes des entraîneurs sportifs de haut niveau pourraient passer à l'histoire!

Je tiens aussi à remercier mes collègues de travail à l'OST; Marie-Claude Laframboise, Mario Rouette et, plus particulièrement, Pascal Lemelin pour avoir pris les bouchées doubles pendant quelques mois.

Je voudrais aussi remercier l'OST pour les données utilisées dans ce mémoire.

Enfin, je ne peux m'empêcher de souligner les doléances de mes deux filles, Mathilde et Justine, pour que je termine ce mémoire avant l'été! Je suspecte une tactique de mon amour Mélanie, pour qui me voir passer tout ce temps devant un écran a été un supplice. Merci d'être aussi compréhensive et résiliente.

Introduction

Être ou ne pas être mentionné dans la liste des auteurs d'un article scientifique dépend en principe de l'importance de la contribution que nous y avons apportée. Toutefois, à moins d'être le seul auteur, la nature précise de notre contribution, notre apport à la recherche et à la conception de l'article demeurent habituellement invisibles aux yeux de la communauté scientifique. Dans certains cas toutefois, l'importance relative de notre rôle peut être estimée en fonction de notre position dans la liste des co-auteurs. En sciences naturelles et génie tout comme en sciences de la santé, par exemple, il est généralement admis que le premier et le dernier auteur sont les principaux contributeurs d'un article scientifique (Pontille, 2004). L'ordre des auteurs devient ainsi un indicateur utile de l'importance relative de chacun des co-auteurs, mais il ne révèle pas la nature précise de leur contribution respective. Qui a défini le problème et l'approche, qui a réalisé les expériences, qui a interprété les résultats, qui a rédigé le texte, etc.?

Comme nous allons le voir, certaines revues rendent disponible depuis quelques années une description explicite des contributions de chacun des co-auteurs. Ces libellés de contributions—*contributorship statements* en anglais—précisent, par exemple, que la rédaction de l'article a été effectuée par les auteurs A, B, et C, que l'analyse des données a été faite par les auteurs, A, C et D, etc. Cette divulgation permet d'évaluer la participation à la recherche de façon beaucoup plus précise et explicite que le permet l'ordre des auteurs, et permet de connaître le rôle précis joué par chacun. Par exemple, il est possible d'identifier

clairement ceux qui ont rédigé l'article, ceux qui ont effectué les expériences, qui ont recruté des volontaires, qui ont réalisé les analyses, etc.

Destinée à assurer une distribution plus équitable des crédits et des responsabilités liés à la publication d'un article scientifique, la divulgation de la contribution de chacun des co-auteurs permet également de mieux apprécier leurs rôles respectifs dans la production des résultats de recherches publiés. Il est bien connu en effet que les co-auteurs d'un article scientifique donné n'y participent pas tous de la même façon, que la définition des approches, et la rédaction du texte ne sont généralement assurées que par un petit nombre d'individus, que l'analyse du matériel empirique est souvent confiée à des spécialistes de la mesure et que le travail de laboratoire ou de terrain est produit le plus souvent par des catégories de personnel situées plus bas dans la hiérarchie des organisations scientifiques. Les libellés de contribution permettent donc de distinguer clairement dans la liste des co-auteurs les diverses catégories de personnels scientifiques.

Les objectifs de la recherche

Les libellés de contributions constituent un indicateur relativement fiable de la division du travail dans le cadre d'activités scientifiques. Toutefois, comme leur divulgation demeure une pratique récente, il n'existe pas à l'heure actuelle de méthodologie pour les rendre accessibles à l'analyse bibliométrique. Dans ce contexte, le premier objectif de notre étude consistera à développer une méthode de traitement de l'information permettant d'opérationnaliser l'usage des libellés de contribution comme indicateur de la division du travail au sein des équipes de recherche. Dans un second temps, nous démontrerons la faisabilité de notre méthode de même que l'utilité des indicateurs qu'elle permet de produire à travers l'exploration des relations entre les différentes contributions effectuées par les auteurs en fonction de la taille des équipes de recherche. Enfin, nous effectuerons une étude en profondeur des disparités hommes/femmes dans la division du travail de recherche. Le choix de ce sujet tient à l'intérêt grandissant pour la mesure de la présence des femmes en science. Comme nous le verrons dans notre revue de littérature, de nombreuses études se sont penchées sur cette question depuis plus de quarante ans et, encore aujourd'hui, l'Organisation des Nations Unies pour l'Éducation, la Science et la Culture (UNESCO) consacre une grande part de ses efforts à la promotion de la place des femmes en sciences et en fait un de ses domaines prioritaires[1].

Dans l'étude originale présentée ici, nous tenterons premièrement de déterminer la nature de la division du travail telle que mesurée par les libellés de contribution. Dans un second temps, nous examinerons les rôles joués par les auteurs en fonction de leur sexe. Plus précisément,

[1] Voir : http://www.unesco.org/new/fr/natural-sciences/priority-areas/gender-and-science/.

nous voulons comparer la nature et la diversité de la contribution des co-auteurs en fonction de leur sexe tout en ventilant ces résultats en fonction de critères discriminants comme la position dans la liste des auteurs, le statut d'auteur de correspondance et l'âge académique.

L'opérationnalisation de nos objectifs de recherche se fera à travers les deux questions principales suivantes :

1- Quelle est la proportion des auteurs qui ont participé aux différentes contributions selon la nature des contributions?
    a. Existe-t-il des différences entre les disciplines?
    b. La taille de la liste des auteurs a-t-elle une incidence sur la nature et la diversité des contributions des différents auteurs?

2- Existe-t-il des différences entre les types de contributions faites par les auteurs masculins et féminins?
    a. La taille de l'équipe de recherche, le statut des auteurs et leur âge académique ont-ils une incidence sur la nature et la diversité de leurs contributions respectives?

Revue de la littérature

La bibliométrie

Les termes bibliométrie et scientométrie ont été introduits à la fin des années 60 par Pritchard et par Nalimov et Mulchenko. Pritchard définit la bibliométrie comme : « the application of mathematics and statistical methods to books and other media of communication » (1969). Nalimov et Mulchenko utilisent le terme scientométrie et le définissent comme : « the application of those quantitative methods which are dealing with the analysis of science viewed as an information process » (Nalimov et Mulchenko, 1969).

La création du *Science Citation Index* par Eugene Garfield en 1963 permet la mise en relation de documents publiés dans des revues scientifiques, grâce aux références faites à ceux-ci par d'autres documents sources, indexés dans la base de données (Price, 1965). C'est de cette relation que naît le fameux Facteur d'impact (Garfield, 1979) ainsi que tous les autres indicateurs bibliométriques utilisant le lien entre les documents cités et les documents citants, de l'analyse de réseaux de co-citations (Small, 1973) à l'indice H (Hirsch, 2005). L'autre caractéristique originale du *Science Citation Index*, la présence des adresses institutionnelles complètes—départements, institutions, villes, provinces, et pays—de chacun des auteurs, permet aussi de mesurer la distribution géographique de l'activité scientifique.

Au fil des ans, plusieurs autres corpus sont venus se greffer au *Science Citation Index* dont le *Social Science Citation Index*, le *Conference Proceedings Citation Index*, le *Book Citation Index* ainsi que la version *Expanded* du *Science Citation Index*. Ces corpus sont intégrés dans

une plateforme multi-sources connue sous l'appellation *Web of Science* et est maintenant détenue par la firme Thomson Reuters.

Aujourd'hui, d'autres bases de données permettent aussi la production d'indicateurs bibliométriques dont *Google Scholar, Microsoft Academic* et *Scopus* d'Elsevier. Alors que les deux premières sources permettent des analyses bibliométriques limitées aux auteurs—et, ainsi, ne permettent pas la régionalisation de l'information via l'analyse des adresses institutionnelles—la base de données *Scopus* permet, grosso modo, le même type d'analyse que le *Web of Science*. Ainsi, le *Web of Science* et *Scopus* peuvent être utilisées afin d'effectuer des analyses bibliométriques avancées.

Le contexte de l'utilisation des libellés de contributions

Hyperautorat[2] et fraudes

Depuis le milieu des années 90, les articles signés par plus d'un auteur représentent plus de 90% de l'ensemble des publications en sciences naturelles et génie (Larivière, Gingras et Archambault, 2006) et près de 70% en sciences sociales et humaines (Larivière, 2007). Parmi ces articles, certains sont signés par plusieurs dizaines d'auteurs, voire des centaines, particulièrement en physique et en astronomie.

Ce phénomène, nommé hyperautorat, tend à modifier la signification de la signature scientifique (Cronin, 2001) et, par conséquent, le crédit donné aux auteurs ainsi que les responsabilités liées à ce statut (Biagioli et Galison, 2003; Birnholtz, 2006; Claxton, 2005a,

[2] Traduction libre de *Hyperauthorship* proposé par Cronin (2001).

2005b; Cronin, 2001; Pontille, 2004; Rennie, Yank et Emanuel, 1997). Dans un tel contexte de collaboration entre de nombreux individus, le rôle joué par chacun des auteurs et la répartition du travail de recherche entre chacun d'eux deviennent en fait difficiles à apprécier, si bien qu'il est même parfois permis de douter de la contribution réelle de certains signataires. À cet égard, de très nombreuses et importantes lacunes ont pu être identifiées dans la littérature. Ainsi, après une analyse de près de 445 revues scientifiques et plus de 650 codes d'éthique proposés par les organisations professionnelles, Bošnak et Marušić (2012) arrivent à la conclusion que l'absence ou la trop grande variété des définitions de l'autorat mènent à des pratiques de signature pour le moins discutables.

En effet, outre l'augmentation du nombre d'auteurs par article, d'autres phénomènes plus préoccupants ont également mis en lumière la nécessité d'une plus grande transparence dans la façon d'attribuer le crédit aux auteurs des articles scientifiques. Les cas de fraudes scientifiques par exemple, pour être élucidés et traités de façon équitable, demandent une connaissance fiable du rôle de chacun des co-auteurs engagés dans la recherche mise en cause. Dans la même veine, la présence d'auteurs « honorifiques » qui n'ont apporté aucune contribution réelle à la recherche et qui n'apparaissent donc que pour des motifs inavouables dans la liste des auteurs ou, encore, l'omission volontaire de certains contributeurs. (Benos et al., 2005; Claxton, 2005a; Flanagin et al., 1998; Rennie et Flanagin, 1994). Dans un contexte où il faut justifier et expliciter la contribution de chacun, ces fraudes sont difficilement camouflables.

Au cours des quinze dernières années, l'hyperautorat et les fraudes vont conduire les éditeurs de revues scientifiques à recommander l'usage d'une définition claire et précise du statut

d'auteur et des conditions pour y accéder. Les guides de bonnes pratiques de publication, qui définissent entre autre les normes de présentation des articles et des références, se présentent aussi de plus en plus sous forme de code d'éthique à l'usage des revues et des auteurs. Plus important encore, certains proposent une définition de l'autorat déterminant les contributions minimales à l'attribution du crédit lié à l'article et par le fait même au statut d'auteur (Resnik et Master, 2011; Smith et Williams-Jones, 2012; Smith, Jordan et Walsby, 2011).

Définition de l'auteur selon International Committee of Journal Medical Editors

La définition de l'auteur du International Committee of Journal Medical Editors est probablement la plus diffusée. En effet, le grand nombre de revues médicales publiées dans le monde a favorisé sa diffusion et son adoption.

Absente des premières éditions des Uniform Requirements for Manuscripts Submitted to Biomedical Journals développés par l'International Committee of Medical Journal Editors (ICMJE)[3], une définition explicite de l'autorat comprenant trois critères apparaît dans l'édition de 1988 :

> All persons designated as authors should qualify for authorship. Each author should have participated sufficiently in the work to take public responsibility for the content. Authorship credit should be based only on substantial contributions to
>
> (a) either conception and design or else analysis and interpretation of data and to
>
> (b) drafting the article or revising it critically for important intellectual content and on
>
> (c) final approval of the version to be published.
>
> All three conditions must be met. (ICMJE, 1988)

[3] À l'époque, il était aussi connu sous le nom de « Vancouver Group » ou International Steering Committee of Medical Editors.

En 2013, l'ICMJE publie ses *Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals* et bonifie sa définition de l'autorat en y ajoutant un quatrième critère qui, clairement, est une réponse à l'augmentation des cas de fraude scientifique :

> The ICMJE recommends that authorship be based on the following 4 criteria :
>
> - Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
> - Drafting the work or revising it critically for important intellectual content; AND
> - Final approval of the version to be published; AND
> - Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. (ICMJE, 2013)

L'écho de cette définition se retrouve dans certains systèmes de pointage qui tentent de déterminer la valeur de la contribution de chacun des auteurs et ainsi attribuer le crédit à travers l'ordre d'apparition dans la liste des auteurs ou, dans le cas d'une contribution jugée peu significative—voire moins intellectuelle et plus cléricale—dans les remerciements (Eggert, 2011; Hunt, 1991). Reste qu'il est difficile d'établir la contribution réelle de chacun des auteurs *a posteriori* (après la publication de l'article) car, même si certaines revues exigent leur divulgation, elles ne publient pas systématiquement cette information. Certaines études vont tenter ainsi de l'estimer ou de l'évaluer avec différentes méthodes reposant sur l'analyse de l'ordre des auteurs ou sur des sondages auprès des auteurs d'articles scientifiques (Davis et Gregerman, 1969; Hoen, Walvoort et Overbeke, 1998; Moulopoulos, Sideris et Georgilis, 1983; Shapiro, Wenger et Shapiro, 1994), mais de tels résultats demeurent toujours approximatifs et ambigus. D'une part, comme les pratiques de cosignature diffèrent d'une

discipline à l'autre (Endersby, 1996; Frandsen et Nicolaisen, 2010; Marušić, Bošnjak et Jerončić, 2011; Pontille, 2004; Zuckerman, 1968), l'ordre des auteurs ne réflètent pas toujours la nature et l'étendue de la contribution de chacun, car l'ordre alphabétique prévaut parfois. Aussi, alors que dans certaines disciplines l'ordre est fonction de l'importance de la contribution, dans d'autres, les dernières positions de la liste des auteurs constituent des places de choix. En effet, elles sont détenues par les auteurs les plus importants autant de part leur contribution mais aussi du point du hierarchique. D'autre part, les sondages ne récoltent que les impressions personnelles des répondants sur leur contribution plusieurs mois, sinon des années, après la publication de l'article. Ces deux principales méthodes n'offrent donc pas d'analyses empiriques très fiables des contributions réelles et effectivement publiées.

Divulgation des contributions

En juillet 1997, la revue *The Lancet* a mis en application pour la première fois un système proposé, entre autres, par Rennie, Yank et Emanuel (1997), et fondé sur la divulgation des contributions de chacun des auteurs. Cette divulgation permet une distribution plus équitable des crédits et des responsabilités liés à la publication d'un article scientifique. En vertu de ce système, les auteurs deviennent des *contributors* et parmi eux, certains se portent garants de l'intégrité de la recherche à titre de *guarantors*. Yank et Rennie (1999, 661) le présentent ainsi :

> … a new system that 1) acknowledges as a contributor everyone who has performed important work on a project that results in an article, 2) lists descriptions of their contributions for the reader, 3) includes on the byline the names of those who contributed most substantially, and 4) lists as guarantor those who can take responsibility for the integrity of the entire work.

Depuis, plusieurs revues scientifiques dont *Nature* (1999, 2009), le *British Medical Journal* (Smith, 1997) et le *Journal of the American Medical Association* (Rennie, Flanagin et Yank, 2000), ont commencé à publier systématiquement les contributions des auteurs[4]. Dans le même mouvement, la Public Library of Science (PLoS)[5] fournit depuis ses débuts en 2003 le libellé de contribution dans 97% des articles publiés dans ses revues en accès libre[6].

Outre une plus grande transparence relative aux contributions spécifiques de chacun des auteurs, ces pratiques ont aussi permis la réalisation d'analyses sur les pratiques d'autorat. Par exemple, une étude sur les liens entre l'ordre des auteurs, la nature des contributions et la conformité aux critères de l'ICMJE a été effectuée par Yank et Rennie (1999) sur 121 articles publiés entre juillet et décembre 1997 dans la revue *The Lancet*. Leurs résultats démontrent qu'il n'y a pas de correspondance claire entre la conformité aux critères de l'ICMJE et les contributions fournies par les auteurs. En effet, les premiers auteurs semblent avoir contribué de manière significative à l'article mais, comme le mentionnent les auteurs de l'étude, le peu de documents analysés incite à la prudence lors de l'interprétation des résultats. La même analyse a été effectuée sur un échantillon plus grand (N = 1 068) avec le même genre de résultats, c'est-à-dire des rôles plus importants chez les premiers auteurs (Hwang et al., 2003). Pour les autres auteurs, la contribution était beaucoup moins claire.

Les études de ce type sur le rôle des auteurs n'en sont donc qu'à leurs débuts et beaucoup de questions demeurent encore sans réponse. Outre des analyses sur le rôle des premiers auteurs,

[4] Pour des exemples tirés de la revue *Nature* voir : http://blogs.nature.com/nautilus/2007/11/post_12.html.
[5] Voir : http://www.plosone.org/static/editorial#authorship.
[6] Selon les données préliminaires issues du projet effectué dans le cadre du cours SCI6916 – Projet dirigé.

la divulgation des contributions permet d'identifier plus clairement le ou les responsables de chacune des tâches inhérentes aux activités de recherche : du design de l'expérimentation de la recherche à l'approbation de la version publiée de l'article, en passant par l'analyse des résultats et la rédaction de l'article. Une telle richesse d'information permet la formulation de nombreuses questions de recherche. Par exemple, comment la progression de la carrière des chercheurs affecte-t-elle l'évolution de leurs contributions aux articles qu'ils signent? Comment les diverses structures institutionnelles de la recherche affectent-elles la division du travail scientifique entre co-auteurs? Comment se manifeste la répartition géographique ou socio-économique des contributions? Les recherches effectuées en collaboration entre auteurs des pays en voie de développement et auteurs des pays développés sont-elles équitables ou assiste-t-on plutôt à une division internationale du travail qui confine les premiers à des tâches subalternes? Chacune de ces questions mériterait à elle seule une étude approfondie pour laquelle l'analyse des libellés de contribution pourrait apporter des éléments de réponse empiriques fiables et originaux. Dans le cadre de ce mémoire, nous allons nous concentrer sur deux questions intimement liées, la division du travail de recherche en général et l'analyse des contributions spécifiques de chacun des auteurs d'un article scientifique selon leur sexe. Comme nous le verrons dans la section suivante, les disparités homme/femme en science ont déjà été traitées abondamment à l'aide de méthodes bibliométriques, notamment du point de vue de la productivité et de l'impact scientifique des travaux de recherche, mais la question de la division du travail selon le sexe au sein même des équipes de recherche demeure encore inexplorée. L'analyse des libellés de contribution permet de combler cette lacune.

Disparité homme/femme : Présence, productivité et qualité

Les études sur la disparité homme/femme en science sont multiples et proviennent de différents champs d'expertise. En bibliométrie, celles-ci s'intéressent à deux dimensions de l'activité de recherche scientifique : la productivité et l'impact scientifique. Comme le sujet de notre étude de cas est la contribution spécifique des femmes aux articles scientifiques, nous nous attardons à la présence des femmes en tant qu'auteures des articles scientifiques.

Présence des femmes

Le tableau 1 présente le recensement des résultats des recherches récentes sur la productivité et la présence des femmes comme auteurs d'articles scientifiques. Il révèle que les études sur le sujet ont été assez nombreuses au cours des dernières années et qu'elles ont visé des populations et des appartenances disciplinaires variées.

Tableau 1. Présence des femmes (en %) selon l'étude, la période, le lieu/population visée et la discipline.

Étude	Période	Lieu/Population	Discipline	% Femmes
West et al., 2013	1900-2011	Monde	Ventilation disciplinaire	21,9
Keith, Layne, Babchuk & Johnson, 2002	1960-1995	États-Unis (Liste de chercheurs)	Sociologie (Revues top 3)	27,6
Aksnes, Rorstad, Piro & Sivertsen, 2011	1981-2009	Norvège	Ventilation disciplinaire	35,0
Ostby, Strand, Nordas & Gleditsch, 2013	1983-2008	Revue	Relations Internationales	40,0
Lewison & Markusova, 2011	1985;1995;2005	Russie	Ventilation disciplinaire (2005)	30,0
Borrego, Barrios, Villarroya, & Olle, 2010	1990-2002	Espagne (Liste de chercheurs)	Toutes	41,7
Gonzalez-Brambila & Veloso, 2007	1991-2002	Mexique	Toutes	27,0
Bordons, Morillo, Fernandez & Gomez, 2003	1994-1999	Espagne	Ressources naturelles	24,0
			Chimie	38,0
Bunz, 2005	1994-2004	Revue	Communication	41,6
Rigg, McCarragher & Krmenec, 2012	1995-2009	Revue	Géographie	21,7
Mauleon & Bordons, 2006	1996-2000	Espagne	Sciences des matériaux	32,0
Braisher, Symonds & Gemmel, 2005	1999-2004	Australie et Royaume-Uni	Science de la vie (*Nature* et *Science*)	33,0
Mendlowicz et al., 2010	2001-2008	Brésil	Psychiatrie	34,7
Nourmohammadi & Hodaei, 2013	2001-2010	Iran	Toutes	36,1
Bird, 2011	2005	Royaume-Uni	Sciences sociales	32,0
Sotudeh & Koshian, 2013	2005-2007	Revue	Nanoscience et technologie	13,8
Abramo, D'angelo & Murgia, 2003	2006-2010	Italie	Ventilation disciplinaire	37,0
Barrio, Villarroya & Borrego, 2013	2007	Espagne	Psychologie	42,0
Larivière et al., 2013	2008-2012	Monde (et par pays)	Toutes	30,0
Peñas & Willett, 2006	Tout	États-Unis et Royaume-Uni (Liste de chercheurs)	Bibliothéconomie et sciences de l'information	43,4

Certaines études mesurent la disparité homme/femme en utilisant comme corpus les articles d'une ou plusieurs revues spécifiques (Bunz, 2005; Ostby, Strand, Nordas et Gleditsch, 2013; Rigg, McCarragher et Krmenec, 2012; Sotudeh et Khoshian, 2014). On compte aussi plusieurs travaux effectués à partir de listes de chercheurs ou portant sur un département particulier (Borrego, Barrios, Villarroya et Olle, 2010; Eloy et al., 2013; Keith, Layne, Babchuk et Johnson, 2002; Penas et Willett, 2006). Bien sûr, il est aussi possible d'effectuer l'analyse d'un lieu géographique (Aksnes, Rorstad, Piro et Siverstsen, 2011; Lewison et Markusova, 2011; Nourmohammadi et Hodaei, 2014). Les données de ces recherches bibliométriques sont présentées avec une ventilation disciplinaire ou non. Par exemple, Mendlowicz et al. (2011) se concentrent seulement sur la psychiatrie au Brésil.

Larivière et al. (2013) présentent des résultats pour l'ensemble des pays et ce, en fonction des disciplines de recherche. En utilisant les noms et prénoms des auteurs ainsi que le pays d'appartenance, il leur a été possible de déterminer leur sexe. De fait, ils ont relevé le défi d'identifier le sexe de tous les auteurs ayant publié dans une ou autre des revues recensées dans le *Web of Science* et ce, peu importe le lieu géographique, la revue ou le département. Cet exploit en fait l'étude la plus exhaustive dans le domaine. Leurs résultats montrent entre autre que les femmes représentent 30% des auteurs et que les articles rédigés par une femme en solo ou lorsqu'elle se trouve première ou dernière auteure ont moins d'impact (en termes de citations) que ceux rédigés par un homme dans ces mêmes positions.

La part des femmes dans les articles scientifiques varie donc d'une étude à l'autre selon la nature des données recensées, des lieux géographiques et milieux professionnels étudiés.

Néanmoins, le pourcentage de femmes dans les articles scientifiques se situe, à quelques exceptions près, entre 30% et 40%. En général, ces pourcentages sont plus élevés dans les disciplines des sciences sociales, et plus bas dans les disciplines des sciences physiques et du génie.

Productivité

Le tableau 2 présente les données sur la productivité des femmes et des hommes disponibles dans les études recensées. Ces études bibliométriques montrent que la différence entre la production scientifique des femmes et des hommes se situe en moyenne autour de 30%. Ainsi, dans l'ensemble des pays et des disciplines étudiés, ces études ont montré une productivité moindre pour les femmes. D'une façon analogue à la présence des femmes, on remarque que l'écart entre la productivité des hommes et celle des femmes est plus faible dans le domaine des sciences sociales et humaines que dans le domaine des sciences naturelles, du génie, et de la médecine.

Tableau 2. Productivité des femmes, des hommes et la différence entre la productivité des femmes et des hommes (en %) selon l'étude, la période, le lieu/population visée et la discipline

				Productivité		
Référence	**Période**	**Lieu**	**Discipline**	**Femme**	**Homme**	**Différence (en %)**
Nahkaie, 2002	1984-1987	Canada	Toutes	3,0	4,5	33,3
Aksnes, Rorstad, Piro & Sivertsen, 2011	1987	Norvège	Ventilation disciplinaire	3,2	5,1	36,7
Gonzalez-Brambila & Veloso, 2007	1991-2002	Mexique	Toutes	F = 0,07 articles de moins sur la période		
Maske, Durden & Gaynor, 2003	1992-1993	États-Unis	Économie	7,0	13,9	49,6
Fox, 2005	1993-1994	États-Unis	Sciences et génie	8,9	11,4	21,9
Pripc, 2002	1993-1998	Croatie	Toutes	4,8	6,8	29,4
Xie & Shauman, 2003	1993	États-unis	Sciences et génie (Ensemble de la carrière)	22,9	39,7	42,3
			Sciences et génie (2 ans)	4,5	5,5	18,2
Bordons, Morillo, Fernandez & Gomez, 2003	1994-1999	Espagne	Ressource Naturelles	8,4	10,1	16,8
			Chimie	15,8	20,8	24,0
Bunz, 2005	1994-2004	Revue	Communication	12,4	15,6	20,5
Stack, 2004	1995	États-Unis	Sciences et génie	5,5	8,4	34,5
Mauleon & Bordons, 2006	1996-2000	Espagne	Sciences des matériaux	17,8	25,6	30,5
Braisher, Symonds & Gemmel, 2005	1999-2004	Australie et Royaume-Uni	Sciences de la vie (*Nature* et *Science*)	9,1	12,4	26,6
Gallivan & Benbunan-Fich, 2007	1999-2003	Revue	Science de l'information (auteurs avec plus de 3 articles)	4,3	4,4	2,3
Larivière et al., 2011	2000-2008	Québec	Santé	12,0	19,0	36,8
			Sciences naturelles et génie	13,0	19,0	31,6
			Sciences sociales et humanités	2,3	3,2	28,1
Corley, 2005	2001-2002	États-Unis	Sciences et génie	2,5	3,5	28,6
Sotudeh & Koshian, 2013	2005-2007	Revue	Nanoscience et technologie	4,0	4,4	8,7
Larivière et al., 2013	2008-2012	Monde(et par pays)	Toutes (1er auteur)	1,0	1,9	48,2
Keith et al., 2002	Début à 1995	États-Unis	Sociologie (revues top 3)	0,3	0,4	38,1
Leahy, 2006	Tout	États-Unis	Sociologie et Linguistique	9,5	15,5	38,7
Penas & Willett, 2006	Tout	États-Unis et Royaume-Uni	Bibliothéconomie et Sciences de l'information	2,7	4,7	42,6

Qualité et impact scientifique

Comme nous l'avons vu précédemment, Larivière et al. (2013) observent que les articles écrits par une femme en solo ou lorsqu'elle se trouve première ou dernière auteure ont moins d'impact que ceux écrits par un homme qui occupe ces mêmes positions dans la liste des auteurs. Même si la différence des comptes de citations entre les hommes et les femmes qu'elle rapporte est plus faible, l'analyse de Gonzales-Brambila et Veloso (2007) effectuée sur la production des chercheurs mexicains fournit des résultats similaires. Aksnes, Rorstad, Piro et Sivertsen (2011) arrivent aussi à une légère différence en termes d'impact en faveur des hommes. Ils attribuent celle-ci aux avantages cumulatifs (Merton, 1968) des citations associés à la productivité accrue des hommes. D'autres études rapportent des différences non-significatives (Bordons et al., 2003; Lewison, 2001; Sotudeh et Khoshian, 2014) alors que Long (1992), à partir d'une analyse des articles en biochimie, montre que les femmes reçoivent un nombre plus élevé de citations que les hommes. Une étude espagnole sur les post-doctorants suggère aussi que les femmes publient dans des revues ayant un plus grand impact que celles où les hommes publient (Borrego, Barrios, Villarroya et Ollé, 2010). En somme, aucune tendance claire ne se dégage dans la littérature concernant la relation entre le sexe des auteurs et l'impact des publications.

Méthodologie

Dans les prochains paragraphes, le processus via lequel nous avons créé notre jeu de données sera exposé en détail. Tout d'abord, les sources de données seront présentées. Par la suite, nous passerons en revue l'ensemble des traitements effectués sur ces données, du téléchargement des fiches des articles étudiés jusqu'à l'identification du sexe des auteurs. À la fin de ce parcours, nous serons en mesure d'apprécier le soin et l'effort qui ont été déployés afin de constituer le jeu de données.

Les sources de données

Nous avons utilisé deux sources afin de construire notre jeu de données : la base de données des publications de l'Observatoire de sciences et de technologies (OST) construite à partir du *Web of Science* de Thomson Reuters et les articles de la Public Library of Science (PLoS) disponible sur leur site web en format XML. Bien que les revues publiées par PLoS soient indexées dans le *Web of Science*, le libellé de la contribution n'est pas une information que cette dernière collige et, ainsi, ces informations ne sont disponibles que dans le plein texte des documents diffusés sur le site web de PLoS. Un des défis principaux de la présente étude consiste donc à lier entre elles ces deux sources d'information.

Base de données de l'OST - Web of Science

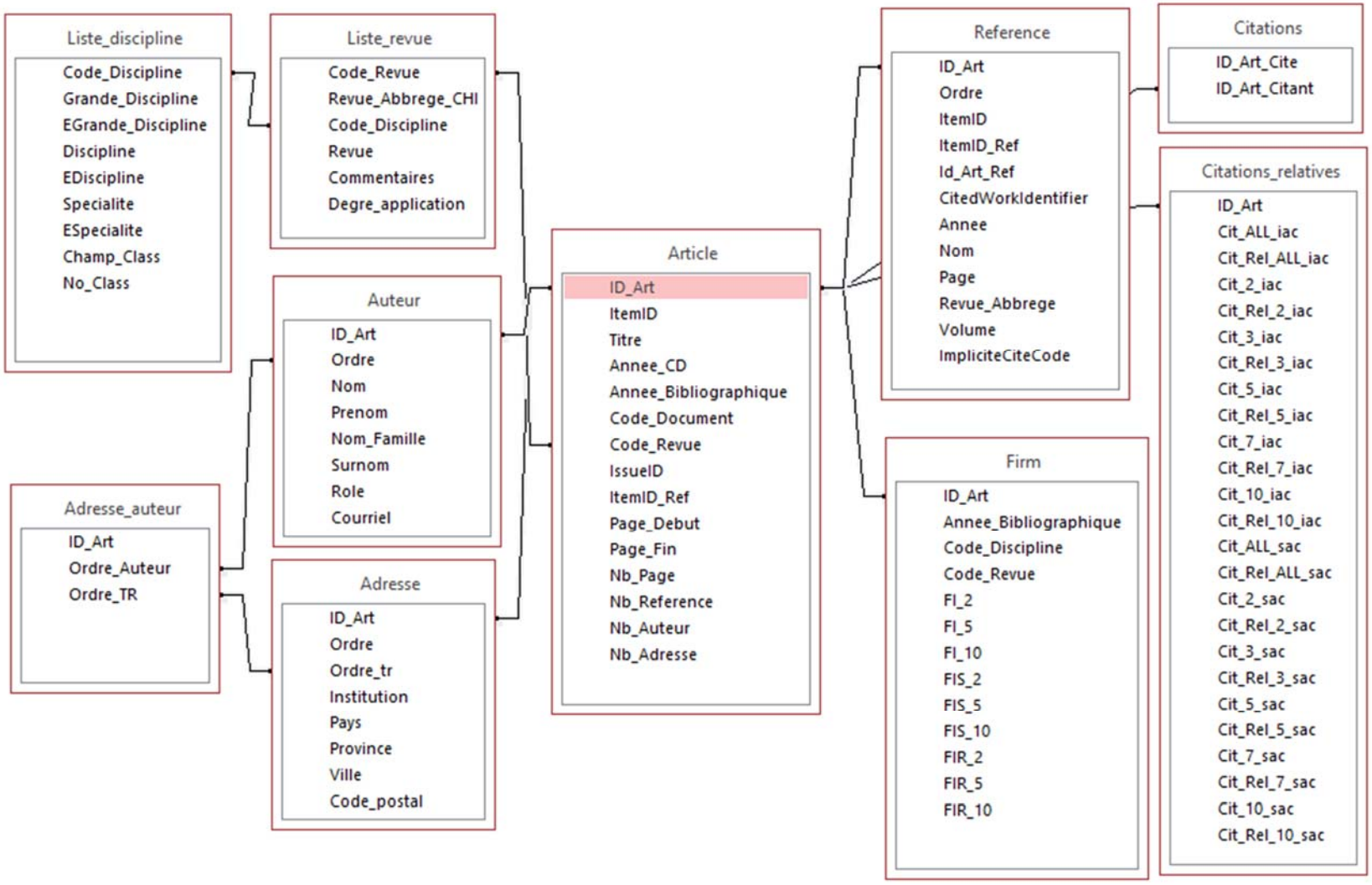
La base de données des publications—créée et maintenue par l'OST sur un serveur MSSQL—comprend l'essentiel du *Web of Science* commercialisé par Thomson Reuters, soit le *Science Citation Index Expanded*, le *Social Science Citation Index* et le *Arts & Humanities Citation*

Index. Les revues dépouillées par Thomson Reuters sont sélectionnées en fonction de divers critères dont les normes d'éditions, le contenu rédactionnel, le caractère international des contributions qu'elles publient et le nombre de citations qu'elles reçoivent dans les autres revues scientifiques (Testa, 2010). Le *Web of Science* contient à ce jour plus de 50 millions de notices bibliographiques provenant de plus de 12 000 revues scientifiques. Ces notices bibliographiques contiennent près d'un milliard de citations. Ils citent bien sûr des documents dont les notices sont présentes dans le *Web of Science*, mais aussi des documents non-recensés dans celui-ci.

La structure de la base de données de l'OST a été développée à des fins bibliométriques (cf. figure 1). La table Article contient les informations sur chacun des notices bibliographiques indexées dans la base de données comme le type de documents (article, note, etc.), l'année de publication, le titre, etc. Les noms des auteurs se trouvent dans la table Auteur et les adresses institutionnelles de chacun dans la table Adresse. Le lien entre les auteurs et leurs adresses institutionnelles, colligé dans la table Adresse_auteur, n'est présent qu'à partir de l'année 2008. Ce lien n'est pas exhaustif et, au sein d'un même article, il n'est pas nécessairement fait pour l'ensemble des co-auteurs. Malgré cela, il nous sera très utile pour l'attribution du sexe en fonction du prénom, du nom de famille et du pays.

La table Revue comprend le nom des revues dépouillées par Thomson Reuters ainsi que la catégorie dans laquelle elles sont classées au sein de la classification disciplinaire de la National Science Foundation (NSF). Le lien entre les documents cités et citant est conservé dans la table Citations. Il permet le calcul d'indicateurs construit sur le compte des citations.

Figure 1. Diagramme abrégé de la base de données des publications de l'OST



Public Library of Science (PLoS)

La Public Library of Science (PLoS) est l'éditeur de huit revues scientifiques en libre accès avec comités de pairs. *PLoS Biology* (2003) *et PloS Medicine* (2004) sont les deux premières revues publiées par PLoS. Suivent par la suite, *PloS Genetics*, *PLoS Computationnal Biology*, *PLoS Pathogens* en 2005, *PLoS One* en 2006 et *PLoS Neglected Tropical Diseases* en 2007. *PloS Clinical Trials* n'a été publiée qu'en 2006 et 2007. Entre 2003 et 2014, 127 911 documents sont publiés par PLoS (cf. tableau 3). La revue *PLoS One* est sans contredit la plus importante avec 106 460 documents entre 2006 et 2014. Le reste des documents (21 451) est réparti dans les autres revues.

Tableau 3. Nombre d'articles par revue PLoS, 2003 – 2014*

Revue	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total
PLoS ONE				137	1 230	2 716	4 405	6 750	13 797	23 464	31 524	22 437	**106 460**
PLoS Genetics			77	208	230	352	473	471	565	721	874	552	**4 523**
PLoS Pathogens			41	123	198	286	459	534	556	640	739	524	**4 100**
PLoS Biology	98	456	431	423	321	327	264	304	276	230	292	196	**3 618**
PLoS Computational Biology			72	168	251	287	376	414	418	521	553	423	**3 483**
PLoS Neglected Tropical Dis.					42	179	224	350	445	525	623	533	**2 921**
PLoS Medicine		68	434	487	346	250	199	193	206	208	219	128	**2 738**
PLoS Clinical Trials				40	28								**68**
Total	**98**	**524**	**1 055**	**1 586**	**2 646**	**4 397**	**6 400**	**9 016**	**16 263**	**26 309**	**34 824**	**24 793**	**127 911**

Source : http://article-level-metrics.plos.org/plos-alm-data/.

*2014 est incomplète

Le traitement des données

Processus de téléchargement du corpus

Afin de créer le corpus contenant la totalité des contributions de chacun des auteurs, nous avons utilisé les données produites par PLoS dans le cadre du projet Article-Level Metrics[7]. Disponibles dans un fichier Excel[8], celles-ci portent sur l'ensemble des 127 911 articles publiés durant la période 2003 – 2014. Les données sont présentées sous forme de liste d'indicateurs pour chacun des articles. Comme le Digital Object Identifier (DOI) pour chacun des articles y est présent, il nous permettra d'automatiser le téléchargement des articles et de faire le lien avec la notice bibliographique du *Web of Science*.

Il existe différentes façons de télécharger un document déposé sur un site web. Il est possible de le faire manuellement en sélectionnant les documents un à un, mais compte tenu du grand nombre de documents à traiter ici, cette option s'avère fastidieuse et comporte un risque d'erreurs élevé. Le téléchargement automatisé de fichiers à l'aide d'un programme de type « robot » ou « moissonneur » intégré dans une solution SQL Server Integration Services (SSIS)[9] nous a semblé l'option la plus appropriée.

En plus de présenter le texte intégral de ses articles sur son site web, PLoS rend disponible ceux-ci pour le téléchargement en format PDF, RIS, BibTex et XML. Pour construire les URL permettant le téléchargement de chacun des articles publiés dans ses revues, PLoS utilise

[7] Voir : http://article-level-metrics.plos.org/alm-info/ et Annexe 1 pour les variables disponibles.
[8] Disponible ici : http://article-level-metrics.plos.org/plos-alm-data/.
[9] Pour plus d'information sur ces solutions voir: http://msdn.microsoft.com/fr-ca/sqlserver/cc511477.aspx.

toujours le même format : le site de la revue, le DOI et le type de format sélectionné. Par exemple, le lien suivant : http://www.plosone.org/article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0004048&representation=XML pointe vers le format XML de l'article « The Effects of Aging on Researchers' Publication and Citation Patterns ».

À partir de la liste de DOI et du modèle standard d'URL pointant vers les fichiers XML, nous avons développé une routine de téléchargement automatisé. Pour chacun des articles, l'URL est construite[10] et interrogée par la solution SSIS à l'aide d'un script en Visual C#[11]. Le fichier est ainsi téléchargé automatiquement vers un dossier de travail sur l'ordinateur client. Le téléchargement complété, l'adresse URL, le DOI et la date du téléchargement sont conservés dans une table servant de journal de bord et le cycle se poursuit jusqu'au dernier DOI de la liste.

Il est possible d'employer l'interface de programmation (API – Application Programming Interface) de PLoS pour télécharger les informations sur les contributions[12], mais nous pensons que l'utilisation du texte intégral en format XML est d'une plus grande utilité à court et moyen terme, pour cette recherche ou pour une autre, car elle permet l'utilisation d'outils d'analyse textuel. De fait, on se sert du texte intégral du corpus téléchargé ici dans le cadre d'une recherche sur la distribution des références au sein de la structure IMRaD (Introduction, Methods, Results and Discussion) des articles publiés par PLoS (Bertin, Atanassova, Gingras

[10] Voir Annexe 2.
[11] Voir Annexe 3.
[12] Voir : http://api.plos.org/.

et Larivière, 2014). Le projet sur la recherche à facettes d'articles scientifiques, initié en 2014 par Atanassova & Bertin, emploie également le corpus créé par notre méthodologie (2014).

Extraction des données de PLoS

Les données pertinentes à notre analyse sont ensuite extraites du fichier XML de chacun des articles à l'aide d'un script en Visual C#[13] intégré à la solution SSIS. Comme les métadonnées et les informations bibliographiques sont déjà disponibles dans la version du Web of Science de l'Observatoire des sciences et des technologies, seules deux variables sont extraites du fichier XML : le DOI de chacun des articles et les libellés des contributions de chacun des auteurs. Le jumelage de chacun des articles de PLoS et sa notice correspondante dans le *Web of Science* est fait à partir des DOI. Lors du traitement du fichier XML, les données sont conservées dans une table de travail et le fichier traité est transféré vers un dossier d'archive.

La couverture des revues PLoS par le *Web of Science*

La couverture des revues PLoS par le *Web of Science* est excellente. En effet, tel que présenté dans le tableau 4, il y a 97 209 articles publiés dans les revues de PLoS sur l'ensemble de la période 2008-2013 et près de 97,6 % (94 879) sont recensés dans le *Web of Science*. Cependant, dans le *Web of Science*, tous n'ont pas leur DOI associé et comme le lien avec les données de PLoS est effectué à l'aide de celui-ci, le nombre de documents disponibles pour la construction de notre corpus s'en trouve réduit. Malgré tout, 92 845 documents ont une correspondance directe à l'aide du DOI, soit près de 96%.

[13] Voir Annexe 4

En termes de documents publiés et recensés, la revue *PLoS One* constitue la plus importante avec 82 656 documents publiés, dont 81 393 documents indexés par le *Web of Science*. C'est donc dire que la quasi-totalité des articles publiés dans PLoS One sont recensés dans le *Web of Science*. De plus, le DOI nous permet de retracer plus de 98% d'entre eux. La revue *PLoS Clinical Trials* présente une anomalie : le *Web of Science* indexe davantage d'article que ce que la revue a publié! Après analyse, la présence de doublons—articles indexés plus d'une fois dans le *Web of Science* —explique ce phénomène. L'occurrence de doublons dans le *Web of Science* est très faible[14] et cela n'affecte en rien la qualité des données.

Tableau 4. Nombre d'articles publiés dans les revues de PLoS recensés dans le *Web of Science* et taux (en %) de correspondance par revue, 2008 – 2013

Revues	PLoS	*Web of Science*	Correspondance directe sur le DOI	Correspondance (%)
PLoS ONE	82 656	81 393	81 208	98,2%
PLoS Neglected Tropical Diseases	2 346	2 197	2 098	89,4%
PLoS Genetics	3 456	3 251	2 865	82,9%
PLoS Pathogens	3 214	3 015	2 621	81,5%
PLoS Computational Biology	2 569	2 429	2 090	81,4%
PLoS Medicine	1 275	1 214	975	76,5%
PLoS Biology	1 693	1 380	988	58,4%
PLoS Clinical Trials	-	-	-	-
Total	**97 209**	**94 879**	**92 845**	**95,5%**

Sélection du corpus

Une fois les articles liés, on dispose des informations nécessaires à la sélection de notre corpus. Pour construire notre corpus de base, nous allons utiliser les documents de types article, note et article de synthèse publiés entre 2008 et 2013. Ces types de documents sont généralement admis comme étant des vecteurs de nouveaux savoirs scientifiques

[14] Sur la base de la revue, du titre et du premier auteur, les doublons représentent 0,1% entre 2000 et 2013 pour les documents de types article, note et article de synthèse.

(Moed, 1996). La période choisie permet d'utiliser le lien entre les auteurs et les adresses institutionnelles. Ces premiers critères rappellent 88 067 documents. Les articles sans libellés de contributions sont exclus du corpus (N = 692) tout comme ceux dont le DOI ne permet pas de concordance entre le *Web of Science* et PLoS (N = 369). En excluant quatre doublons, notre corpus comprend finalement 87 002 articles.

Traitement du champ contenant les libellés de contributions

Pour chaque article, les libellés des contributions de toutes les personnes ayant contribué à l'article sont contenus dans un seul champ texte. Dans notre corpus, la plupart des documents comprennent un champ dont le format est relativement standard. Voici l'exemple du format le plus commun :

> Conceived and designed the experiments: UA RMP PP FDC. Analyzed the data: UA RMP. Contributed reagents/materials/analysis tools: EG CD LS MO. Wrote the paper: PP FDC.

Dans ce premier exemple, le lien entre les contributeurs et leurs contributions est toujours construit de la même façon : le libellé de la contribution, deux points, les initiales des contributeurs et un point. Il existe cependant d'autres types de libellés de contributions qui, en raison de leur format, sont plus difficiles à traiter. Par exemple, celui-ci avec les initiales des contributeurs séparées par des virgules et devant la contribution :

> MS, CS, NC, CT, FGB, DR, NSF, MCP, HF, MPF, FB, PVA, PEC, SO, AG, FAS, PD, AM, MLA, and OS conceived, designed or performed the experiments. MS, CS, NC, CT, FGB, DR, NSF, KLR, MCP, HF, FB, PVA, PEC, SO, AG, IS, FAR, FAS, PD, AM, MLA, and OS

> analyzed the data or corrected the paper. BV, AZ, and AS contributed reagents/materials/analysis tools. MS, MLA, and OS wrote the paper.

Ou encore celui-ci avec les noms écrits au long :

> Conceived and designed the experiments: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Performed the experiments: Ohad Yogev, Orli Yogev, Esti Singer, Thomas D. Fox. Analyzed the data: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Contributed reagents/materials/analysis tools: Michal Goldberg, Thomas D. Fox, Ophry Pines. Wrote the paper: Ohad Yogev, Michal Goldberg, Ophry Pines.

Extraction des libellés de contribution et des contributeurs

L'objectif du premier traitement[15] est de séparer la contribution (*conceived and designed the expediments, analyzed the data,* etc.) et le contributeur (les noms ou initiales) tout en conservant le lien qui les unit. Il nous faut premièrement extraire chaque libellé de contribution en faisant suivre les informations sur l'identité des contributeurs (leurs initiales ou leur nom) qui sont associés à ce libellé. Pour construire notre premier ensemble à traiter, nous identifions le format le plus commun en nous limitant aux 86 725 libellés qui commencent par au moins sept caractères sans espace (après une analyse sommaire la plupart des termes utilisés au début des libellés ont plus de six caractères) ou par les termes suivants: *wrote, final, first, icmje, model, this, the, took, idea, for, gave, data, built, study.*

[15] Voir Annexe 5.

Afin de réduire les problèmes associés au traitement automatisé, nous distinguons deux sous-groupes : un sous-groupe A, défini par ceux qui contiennent un nombre de points identique au nombre de deux points et un sous-groupe B défini par ceux qui contiennent un nombre supérieur ou inférieur de points et de deux points. Cette précision permet de distinguer les libellés contenant des points utilisés à d'autres fins qu'à la fermeture d'une phrase ou des points manquants qui viendraient compromettre l'isolement des initiales dans le traitement subséquent.

Le traitement appliqué au sous-groupe A (N = 82 031) consiste d'abord à scinder le libellé à chaque point (.). Par exemple, le texte suivant :

> Conceived and designed the experiments: SD RK. Performed the experiments: SD MM MJH. Analyzed the data: SD MM. Contributed reagents/materials/analysis tools: MJH. Wrote the paper: SD RK.

est décomposé en phrases contenant le libellé de contribution et les contributeurs :

1. Conceived and designed the experiments: SD RK
2. Performed the experiments: SD MM MJH
3. Analyzed the data: SD MM
4. Contributed reagents/materials/analysis tools: MJH
5. Wrote the paper: SD RK

Lors de ce processus, nous avons fragmenté le champ des contributions du groupe A en 493 215 libellés avec les initiales. Par la suite, à l'aide du marqueur de deux points (:) dans les libellés de contributions, les initiales des noms des contributeurs sont isolées et copiées dans un nouveau champ afin d'effectuer leur traitement. À ce moment, un nettoyage de certains

caractères comme le tiret, l'espace ou l'apostrophe est réalisé sur les initiales. Nous isolons ensuite les initiales de chacun des contributeurs à l'aide de l'espace ou de la virgule.

Lors du traitement du champ des contributions du sous-groupe B (N = 4 694), c'est-à-dire des libellés qui contiennent un nombre différent de point et de deux points, nous ne pouvons pas utiliser le point comme marqueur de fin de libellé de contribution. Par exemple, le texte suivant comprend un point qui permet d'abréger le deuxième prénom du contributeur Thomas D. Fox :

> Conceived and designed the experiments: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Performed the experiments: Ohad Yogev, Orli Yogev, Esti Singer, Thomas D. Fox. Analyzed the data: Ohad Yogev, Orli Yogev, Eitan Shaulian, Michal Goldberg, Thomas D. Fox, Ophry Pines. Contributed reagents/materials/analysis tools: Michal Goldberg, Thomas D. Fox, Ophry Pines. Wrote the paper: Ohad Yogev, Michal Goldberg, Ophry Pines.

Si aucun traitement spécifique n'était effectué, ce point serait confondu avec le point permettant de clore la contribution et de passer à la suivante ou encore, de fermer le champ. Il faut alors, pour segmenter ce champ, utiliser une autre méthode. Nous avons donc introduit un nouveau marqueur pour délimiter chacune des contributions à l'intérieur du libellé. La barre verticale « | » a été insérée au début de chacune des contributions en trois étapes. Premièrement, nous avons découpé le texte en segments en utilisant le point. Par contre, il n'est pas possible, à ce moment-ci, de faire suivre correctement les initiales ou les noms

complets des contributeurs car le point (.) n'agit plus ici seulement comme un marqueur de fin de libellé, mais aussi d'abréviations des prénoms ou des préfixes des contributeurs. Deuxièmement, comme le marqueur deux points (:) de début d'énumération des contributeurs est toujours présent, nous l'avons utilisé pour extraire la contribution seulement. Enfin, nous avons introduit le nouveau marqueur de début et, par le fait même de fin, des contributions dans le texte en repérant les libellés de contributions extraits à l'étape 2. Voici un exemple de ce traitement :

> |conceived and designed the experiments: ohad yogev, orli yogev, eitan shaulian, michal goldberg, thomas d. fox, ophry pines. |performed the experiments: ohad yogev, orli yogev, esti singer, thomas d. fox. |analyzed the data: ohad yogev, orli yogev, eitan shaulian, michal goldberg, thomas d. fox, ophry pines. |contributed reagents/materials/analysis tools: michal goldberg, thomas d. fox, ophry pines. |wrote the paper: ohad yogev, michal goldberg, ophry pines.

Nous avons ainsi extrait 139 327 libellées avec les noms ou les initiales des contributeurs. Par la suite, tout comme pour le sous-groupe A, nous isolons les initiales ou noms de chacun des contributeurs à l'aide de l'espace ou de la virgule.

Comme le premier partitionnement (sous-groupe A et B) nous a permis d'établir les libellés de contribution les plus communs, nous avons utilisés les cinq principaux libellés pour délimiter le début et la fin de la contribution ainsi qu'identifier les initiales ou les noms des contributeurs pour les 227 libellés de contributions résiduels. Les libellés de contribution qui

ne se retrouvent pas parmi les cinq plus importants sont classés dans une nouvelle catégorie nommée *Autres libellés de contribution* (à ne pas confondre avec la catégorie originale *Other*). Les documents résiduels ont donc été soumis à un partitionnement plus grossier, mais permettant quand même d'associer les principales contributions aux contributeurs.

À la fin, les traitements effectués sur le libellé des contributions des 87 002 articles ont permis d'extraire 20 667 libellés distinct et non harmonisés associés à 40 356 initiales (cf. tableau 5).

Tableau 5. Nombre et taux de couples article-initiales et nombre d'article par les principaux libellés de contribution sans harmonisation

	Article-Initiales		Article	
Contribution	**Nombre distinct (N)**	**% du N total**	**Nombre distinct (N)**	**% du N total**
Analyzed the data	317 974	50,0%	85 786	98,6%
Performed the experiments	308 655	48,6%	82 172	94,4%
Conceived and designed the experiments	286 480	45,1%	84 853	97,5%
Wrote the paper	262 728	41,3%	80 370	92,4%
Contributed reagents/materials/analysis tools	218 161	34,3%	63 961	73,5%
Wrote the manuscript	18 804	3,0%	5 593	6,4%
Icmje criteria for authorship read and met	2 991	0,5%	281	0,3%
Agree with manuscript results and conclusions	2 519	0,4%	244	0,3%
Contributed to the writing of the manuscript	2 336	0,4%	275	0,3%
Other	2 071	0,3%	310	0,4%
Autres libellés de contribution (N = 20 657)	*118 895*	*18,7%*	*17 039*	*19,6%*
Total (Nombre distinct (N))	**635 679**	**-**	**87 002**	**-**

Classification et normalisation des contributions

La normalisation des libellés de contributions est une tâche relativement ardue, car l'étendue des différents types de contributions et de leurs libellés est vaste. Les cinq types de contributions les plus utilisés par les contributeurs constituent naturellement notre classification, et la classe *Other* a été ajoutée afin de tenir compte de l'ensemble des libellés de contributions que nous ne pouvons classer dans l'une ou l'autre des catégories.

Voici la classification utilisée :

1. Analyzed the data
2. Performed the experiments
3. Conceived and designed the experiments
4. Wrote the paper
5. Contributed reagents/materials/analysis tools
6. Other

Tout d'abord nous avons harmonisé manuellement les libellés de contributions jusqu'à représentant 95% du nombre total de combinaisons article – initiales. Par la suite, afin de bonifier notre travail de correspondance manuelle, nous avons comparé, à l'aide de la transformation utilisant la logique floue disponible dans la solution SSIS[16], l'ensemble des 20 669 libellés de contributions aux cinq catégories de notre classification. Le pourcentage de similarité minimum utilisé est de 70%. Nous avons ainsi harmonisé à peine plus de 3% (N = 461) de l'ensemble des libellés. Ce faible taux de correspondance confirme l'étendue des possibles. Au terme de ce travail de classement, à la fois manuel et automatique, tous les libellés de contributions se sont vus attribués une ou l'autre des catégories. Malgré le nombre

[16] Pour plus d'informations sur les transformations SSIS – Fuzzy Lookup : http://msdn.microsoft.com/fr-ca/library/ms137786(v=sql.110).aspx.

élevé de contributions dans la classe *Other* (N = 20 243) celles-ci ne représentent plus maintenant que 12,6% de l'ensemble des couples article – initiales avec 79 978 combinaisons (cf. tableau 6).

Tableau 6. Nombre distinct et pourcentage de couples initiales - article et nombre distinct d'article et pourcentage d'article par libellé de contribution harmonisé

	Article-Initiales		Article	
Contribution	**Nombre distinct (N)**	**% du N total**	**Nombre distinct (N)**	**% du N total**
Analyzed the data	320 080	50,6%	85 900	98,7%
Performed the experiments	311 679	49,3%	82 811	95,2%
Conceived and designed the experiments	288 765	45,6%	85 406	98,2%
Wrote the paper	287 796	45,5%	86 517	99,4%
Contributed reagents/materials/analysis tools	220 331	34,8%	64 444	74,1%
Other (20 243)	*79 978*	*12,6%*	*15 900*	*18,3%*
Total (Nombre distinct (N))	**632 799**	**-**	**87 002**	**-**

Lien *Web of Science* – PLoS avec le DOI et les auteurs

Le DOI—identifiant unique à chaque article—est la principale variable permettant de faire le lien avec la version du *Web of Science* de l'OST et les informations téléchargées de PLoS. Ce lien permet d'utiliser, pour chaque article, les métadonnées recensées dans le *Web of Science* et de les jumeler aux contributions décrites dans chaque article téléchargé de PLoS. Cependant, il nous faut aussi faire un deuxième lien sur le nom des auteurs présents dans le *Web of Science* nous permettant de présenter les données selon l'ordre des auteurs. De plus, le travail d'appariement des initiales et des noms contenus dans les libellés de contributions issus de PLoS à celui des noms d'auteurs du *Web of Science* est essentiel en ce qu'il nous assure de la pertinence du traitement d'extraction des informations contenues dans les libellés de contribution. En effet, si plusieurs auteurs d'un article donné n'ont pas de correspondance au

sein des libellés de contribution traités, nous serons contraints de réévaluer l'ensemble de notre traitement.

Pour faire le lien avec la table Auteur du *Web of Science*, nous avons créé une table intermédiaire contenant le nom de famille, le ou les prénoms et le nom complet en format long et abrégé (ex. Philippe Comtois et Comtois-P) de chacun des auteurs des articles de notre corpus. À partir de ces données, nous avons construit deux types d'initiales : celles qui incluent la première lettre du deuxième prénom et celles qui l'excluent. Par exemple, SMITH-J et SMITH-JJ. Les noms d'auteur ayant un seul caractère ont été supprimés de notre ensemble de données pour des raisons évidentes. En effet, les patronymes à une seule lettre sont plutôt rares; on en compte autant qu'il y a de lettres dans l'alphabet sur près de 2 millions de patronymes distincts dans le *Web of Science*.

Un script en SQL a été développé pour faire correspondre chaque nom tiré du *Web of Science* à chaque nom issu des libellés de contribution[17]. La concordance avec les contributeurs tirés des libellés de contributions et les auteurs recensés dans le *Web of Science* est principalement faite sur les initiales. Dans certains cas, par exemple pour les noms écrits au long, nous avons utilisé le nom de famille lorsqu'il était disponible dans la table Auteur. Tous les auteurs (N = 589 892) ont été liés à au moins une de leurs contributions pour 98% des articles (N = 85 260). Les correspondances manquantes s'expliquent par l'orthographe déficiente des noms d'auteurs dans les contributions ou carrément l'absence de ceux-ci. En effet, nous avons constaté l'absence de certains auteurs de l'ensemble des contributions, alors qu'ils sont

[17] Voir Annexe 6.

clairement dans la liste des auteurs des articles. Comme ce n'est pas une mince tâche d'identifier l'ensemble de ces cas et que cela n'ajoute que très peu à la présente étude, nous nous en tiendrons à l'aspect anecdotique du phénomène dans notre corpus. En effet, à peine 2% des articles de notre corpus seraient potentiellement touchés par la présence d'auteurs sans contribution explicite. Ainsi, notre corpus est maintenant composé des 85 260 articles dont la totalité des contributeurs ont été rattachés à leurs fiches correspondantes dans la base de données sur les publications de l'OST. Le lien avec les auteurs recensés dans la base de données de l'OST et le retrait des articles dont la correspondance était déficiente modifient légèrement la distribution des contributeurs et des articles dans la classification des contributions.

Le tableau 7 montre l'impact de ce choix. Les cinq principales contributions sont très peu touchées en termes de couples article-initiales supprimés par le retrait des 1 742 articles de notre corpus. La contribution *Other* est la plus affectée avec 15,1% de baisse du nombre de couples article-initiales respectivement.

Tableau 7. Nombre distinct de couples article-initiales (N et %) et nombre distinct d'article (N et %) par libellé de contribution avec et sans la correspondance des auteurs recensés dans le *Web of Science*

	Corpus sans la correspondance dans le Web of Science		Corpus avec la correspondance dans le Web of Science		Différence en %	
Contribution	**N article-initiales**	**N article**	**N article-initiales**	**N article**	**N article-initiales**	**N article**
Analyzed the data	320 080	85 900	306 592	84 221	4,2%	2,0%
Performed the experiments	311 679	82 811	297 893	81 183	4,4%	2,0%
Conceived and designed the experiments	288 765	85 406	277 302	83 734	4,0%	2,0%
Wrote the paper	287 796	86 517	274 615	84 789	4,6%	2,0%
Contributed reagents/materials/analysis tools	220 331	64 444	208 794	63 049	5,2%	2,2%
*Other**	*79 978*	*15 900*	*67 929*	*15 416*	15,1%	*3,0%*
Nombre distinct (N)	**632 799**	**87 002**	**589 892**	**85 260**	**6,8%**	**2,0%**

* Contient le libellé original de contribution *Other* ainsi que tous les autres libellés de contribution non-classés dans les cinq principales catégories.

Traduction des libellés

Comme notre classification est maintenant complétée, nous allons utiliser une traduction libre des libellés de contributions originaux fournis par l'auteur de correspondance de chacun des articles tirés des revues publiées par PLoS. Voici ce que nous proposons comme traduction (cf. tableau 8).

Tableau 8. Traduction des libellés de contributions

Originale	Traduction
Analyzed the data	Analyse des données
Conceived and designed the experiments	Conception et design des expériences
Performed the experiments	Réalisation des expériences
Contributed Reagents/Materials/Analysis Tools	Fourniture (Réactifs/Matériel/Outils d'analyse)
Wrote the paper	Rédaction de l'article

Nous n'emploierons pas la classe *Other* car elle ne procure aucun éclairage sur le type de contribution apporté à l'article.

Attribution d'une classe disciplinaire aux articles

Dans la version du *Web of science* de l'OST, les revues publiées par PLoS sont classées dans une seule spécialité de la classification disciplinaire de la National Science Foundation (NSF)[18] (cf. tableau 9).

Tableau 9. Nombre d'article selon la revue PLoS et la spécialité (NSF) attribuée à la revue

Revue	Spécialité (NSF)	N Article
PLoS ONE	Recherche biomédicale - général	75 546
PLoS Genetics	Génétique et hérédité	2 653
PLoS Pathogens	Parasitologie	2 095
PLoS Computational Biology	Mathématiques appliquées	1 950
PLoS Neglected Tropical Diseases	Médecine tropicale	1 756
PLoS Biology	Biologie - général	786
PLoS Medicine	Médecine générale	474
Total		**85 260**

Cependant, le nombre important d'articles publiés dans la revue *PLoS One* pose un certain problème sachant qu'il s'agit d'une revue multidisciplinaire qui publie des articles provenant pratiquement de tous les domaines de la recherche scientifique. Dans la mesure où nous désirons identifier les particularités de la division du travail dans les activités scientifiques en fonction des domaines de recherche, il est essentiel de pouvoir attribuer une discipline à chacun des articles. L'attribution d'une spécialité au niveau de l'article rend en effet possible la ventilation par discipline des contributions et du sexe des auteurs. Cet élément ajoute

[18] Voir : http://www.nsf.gov/statistics/seind14/content/chapter-5/at05-25.pdf.

assurément à la compréhension et permet de rendre compte, en partie, des particularités disciplinaires.

Waltman et van Eck (2012) ont développé une méthodologie qui, en regroupant les articles à l'aide des citations, permet le déploiement d'un système de classification des articles scientifiques à partir des mots-clés tirés des titres et des résumés. Nous nous sommes inspirés de cette méthodologie parce qu'elle a l'avantage d'être simple et facilement transposable à nos besoins. Cependant, contrairement à eux, nous ne reconstruirons pas un nouveau système de classification. En effet, la classification des revues de la NSF est amplement adéquate pour le regroupement des articles par spécialité. Nous avons donc utilisé les articles cités ou citant pour attribuer une spécialité NSF à chacun des articles. Précisément, nous avons calculé la fréquence des spécialités citées par chacun des articles publiés dans les revues de PLoS (N articles par spécialité) ainsi que la fréquence des spécialités citant les articles publiés dans les revues de PLoS. La somme des deux nombres permet d'identifier la spécialité la plus courante. Autrement dit, pour un article donné, la spécialité attribuée est celle dont la somme des fréquences des articles cités et citant était la plus élevée. À titre d'exemple, le tableau 10 présente les résultats de l'opération permettant l'attribution d'une spécialité à un article[19] donné.

[19] Petkova VI, Ehrsson HH (2008) If I Were You: Perceptual Illusion of Body Swapping. *PLoS ONE* 3(12): e3832. doi:10.1371/journal.pone.0003832.

Tableau 10. Exemple d'attribution d'une spécialité à un article donné

Spécialité	N article citant	% du total	N article cité	% du total	Total	% du total
Neurologie et neurochirurgie	20	29,9%	19	54,3%	39	38,2%
Recherche biomédicale - générale	20	29,9%	10	28,6%	30	29,4%
Psychologie expérimentale	8	11,9%	5	14,3%	13	12,7%
Psychologie - divers	7	10,4%	1	2,9%	8	7,8%
Informatique	3	4,5%	0	0,0%	3	2,9%
Psychologie - générale	2	3,0%	0	0,0%	2	2,0%
Biochimie et biologie moléculaire	2	3,0%	0	0,0%	2	2,0%
Chirurgie	1	1,5%	0	0,0%	1	1,0%
Psychiatrie	1	1,5%	0	0,0%	1	1,0%
Psychanalyse	1	1,5%	0	0,0%	1	1,0%
Psychologie du développement et des enfants	1	1,5%	0	0,0%	1	1,0%
Sciences sociales générales	1	1,5%	0	0,0%	1	1,0%
Total	**67**	**100,0%**	**35**	**100,0%**	**102**	**100,0%**

Dans l'exemple, les articles citant proviennent majoritairement de la Neurologie et neurochirurgie et de la Recherche biomédicale – générale avec 29,9% des articles chacune. Avec 19 (54,3%) documents cités par l'article donné en exemple, la spécialité Neurologie et chirurgie se retrouve au premier rang. Au total, 39 des 102 (38,2%) documents utilisés (cités et citant) se retrouvent dans la spécialité Neurologie et neurochirurgie.

Le tableau 11 montre la ventilation des articles par discipline telle qu'attribuée par les spécialités des articles cités et citant. Comme nous l'avions envisagé, les articles de la revue *PLoS One* se retrouvent dans plus d'une discipline. La médecine clinique (44,9%), la recherche biomédicale (41,8%) et la biologie (7,9%) sont les principales disciplines dans lesquelles sont classés les articles de *PLoS One.*

Tableau 11. Nombre d'articles (N et %) par discipline pour la revue *PLoS One*

Domaine - Discipline	N	%
Sciences naturelles et génie	**73 209**	**96,9%**
Recherche biomédicale	31 586	41,8%
Médecine clinique	33 893	44,9%
Biologie	5 954	7,9%
Sciences de la terre et de l'espace	634	0,8%
Physique	428	0,6%
Génie	320	0,4%
Chimie	269	0,4%
Mathématique	125	0,2%
Sciences sociales et humaines	**2 328**	**3,1%**
Psychologie	1 489	2,0%
Santé	408	0,5%
Sciences sociales	298	0,4%
Champs professionnels	127	0,2%
Humanités	5	0,0%
Arts	1	0,0%
Inconnu	**9**	**0,0%**
Total	**75 546**	**100,0%**

En examinant de plus près la classification à l'aide du tableau 12, nous constatons que, pour la plupart des revues, les articles se concentrent dans moins de quatre spécialités. Par exemple, 72,4% des articles de *PLoS Genetics* sont classés en Génétique et hérédité ainsi que dans la spécialité Biochimie et biologie moléculaire. Nous ne sommes pas surpris de voir la spécialité NSF attribuée manuellement à cette revue être parmi celles les plus attribuées de manière automatique.

Tableau 12. Principales spécialités (nombre d'articles) attribuées aux articles des revues PLoS

Revue (Spécialité originale)	Spécialité	Articles	% du total de la revue
PLoS Neglected Tropical Diseases (Médecine tropicale)	Médecine tropicale	645	36,7%
	Immunologie	340	19,4%
	Parasitologie	192	10,9%
	Microbiologie	90	5,1%
	Virologie	85	4,8%
	Biochimie et biologie moléculaire	80	4,6%
	Entomologie	59	3,4%
	Pharmacologie	55	3,1%
	Médecine générale	45	2,6%
	Recherche biomédicale - général	28	1,6%
PLoS ONE (Recherche biomédicale - général)	Biochimie et biologie moléculaire	15 865	21,0%
	Neurologie et neurochirurgie	8 854	11,7%
	Immunologie	6 528	8,6%
	Cancer	5 187	6,9%
	Recherche biomédicale - général	4 482	5,9%
	Génétique et hérédité	3 776	5,0%
	Microbiologie	3 364	4,5%
	Écologie	2 043	2,7%
	Système cardio-vasculaire	1 697	2,2%
	Médecine générale	1 682	2,2%
PLoS Pathogens (Parasitologie)	Virologie	604	28,8%
	Immunologie	532	25,4%
	Biochimie et biologie moléculaire	359	17,1%
	Microbiologie	310	14,8%
	Recherche biomédicale - général	125	6,0%
	Parasitologie	44	2,1%
	Botanique	29	1,4%
	Génétique et hérédité	17	0,8%
	Neurologie et neurochirurgie	15	0,7%
	Biologie cellulaire, cytologie et histologie	11	0,5%

Tableau 12. Principales spécialités attribuées aux articles des revues PLoS (suite)

Revue (Spécialité originale)	Spécialité	Articles	% du total de la revue
PLoS Biology (Biologie - général)	Biochimie et biologie moléculaire	294	37,4%
	Neurologie et neurochirurgie	141	17,9%
	Recherche biomédicale - général	89	11,3%
	Génétique et hérédité	81	10,3%
	Immunologie	42	5,3%
	Biologie cellulaire, cytologie et histologie	33	4,2%
	Microbiologie	26	3,3%
	Embryologie	20	2,5%
	Botanique	16	2,0%
	Écologie	11	1,4%
PLoS Computational Biology (Mathématiques appliquées)	Biochimie et biologie moléculaire	691	35,4%
	Neurologie et neurochirurgie	440	22,6%
	Recherche biomédicale - général	353	18,1%
	Génétique et hérédité	114	5,8%
	Immunologie	49	2,5%
	Microbiologie	42	2,2%
	Pharmacologie	30	1,5%
	Cancer	27	1,4%
	Biologie cellulaire, cytologie et histologie	24	1,2%
	Virologie	18	0,9%
	Physique - général	18	0,9%
PLoS Genetics (Génétique et hérédité)	Génétique et hérédité	1 040	39,2%
	Biochimie et biologie moléculaire	880	33,2%
	Recherche biomédicale - général	171	6,4%
	Microbiologie	148	5,6%
	Neurologie et neurochirurgie	95	3,6%
	Botanique	62	2,3%
	Biologie cellulaire, cytologie et histologie	59	2,2%
	Cancer	48	1,8%
	Embryologie	42	1,6%
	Immunologie	21	0,8%

Tableau 12. Principales spécialités attribuées aux articles des revues PLoS (suite)

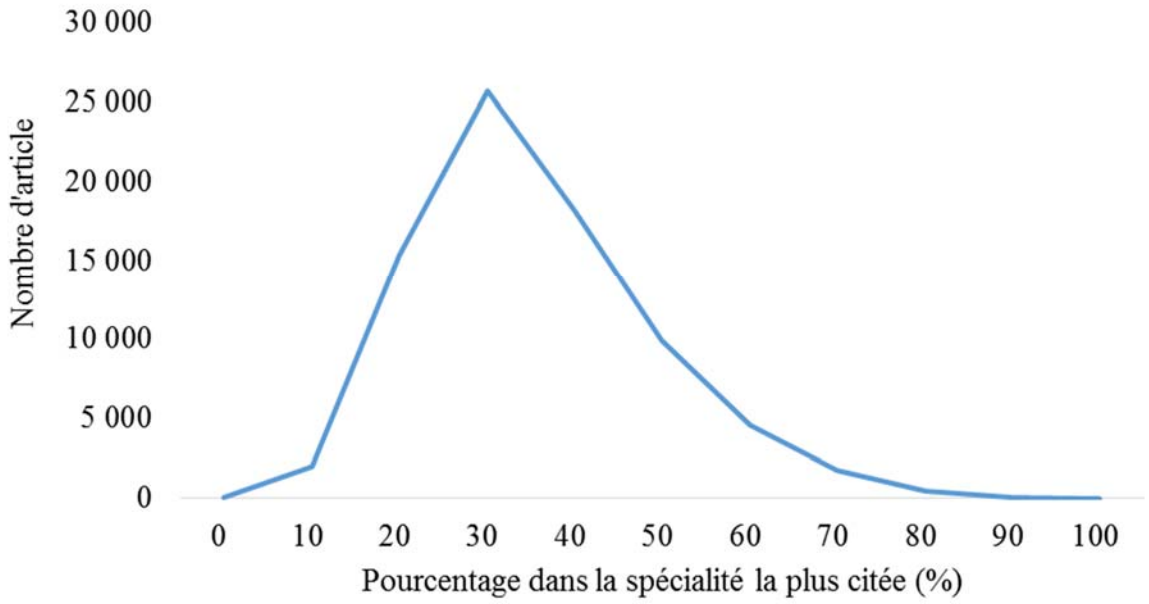
Revue (Spécialité originale)	Spécialité	Articles	% du total de la revue
PLoS Medicine (Médecine générale)	Médecine générale	184	38,8%
	Immunologie	81	17,1%
	Médecine tropicale	27	5,7%
	Psychiatrie	18	3,8%
	Santé publique	17	3,6%
	Cancer	17	3,6%
	Santé au travail et environnement	16	3,4%
	Neurologie et neurochirurgie	14	3,0%
	Système respiratoire	11	2,3%
	Endocrinologie	8	1,7%

En effet, le résultat de notre classement automatique tend à démontrer que cette méthode d'attribution, de par la proximité des spécialités les plus attribuées à chacun des articles et celle attribuée à la revue, propose une avenue fiable au classement disciplinaire des articles. Dans certains cas, la classification originale semble même faire défaut. Par exemple, la revue *PLoS Pathogens* est classée dans la spécialité Parasitologie alors que les articles semblent plutôt traiter de virologie et d'immunologie.

Afin de confirmer la validité de notre méthodologie, nous nous sommes interrogés sur la possibilité d'une attribution fautive en raison du faible nombre de références par spécialité pour un article donné. Par exemple, un article cite peu d'articles et ceux-ci sont distribués dans plusieurs spécialités. Nous avons donc calculé la part de chacune des spécialités en terme de documents cités et ce, pour l'ensemble des articles de notre corpus avec au moins une référence. Il s'avère que 77,8 % des références sont classées dans une spécialité représentant

30% et plus des références faites dans les articles de notre corpus (cf. figure 2). Cela nous apparaît fort acceptable d'autant que nous utilisons aussi les articles citant pour attribuer la spécialité.

Figure 2. Distribution des articles en fonction de la part de leurs références dans la spécialité la plus fréquemment citée



Identification du sexe

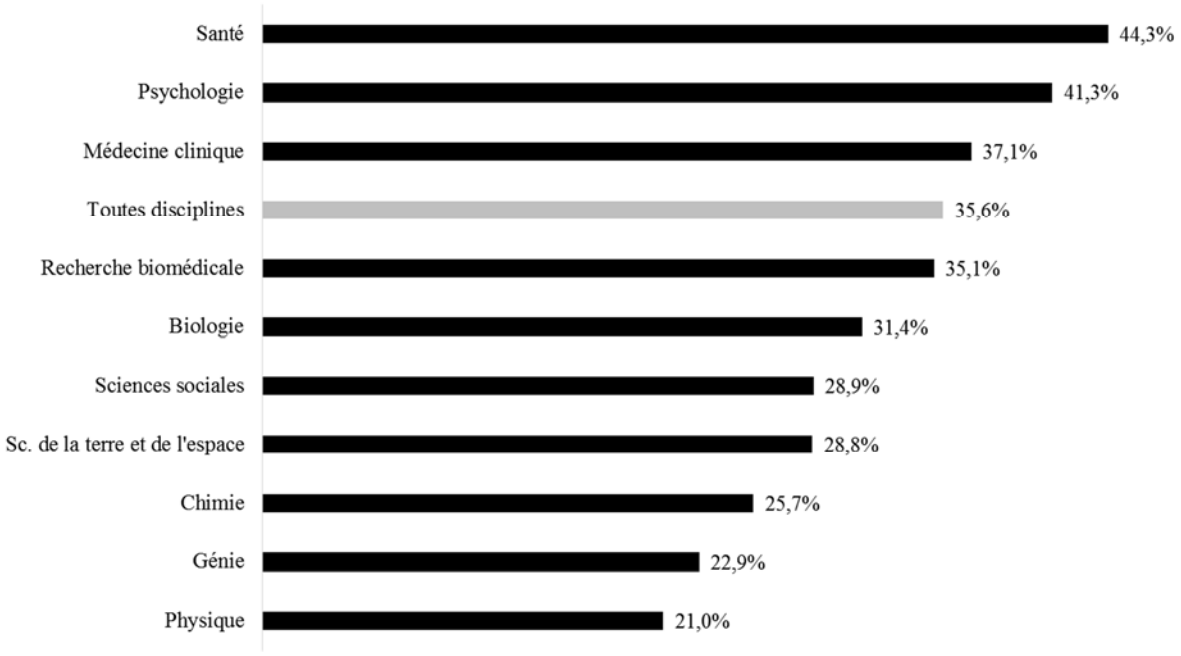
L'identification du sexe des auteurs a été faite à partir des tables de conversion développées dans le cadre de l'article Global Gender Disparities in Science (Larivière et al., 2013). Larivière, un des auteurs de l'article et directeur du présent mémoire, nous a gracieusement donné accès aux informations relatives à l'identification du sexe par le prénom, le nom et le pays d'attache des coauteurs. Pour les informations manquantes, nous avons employé la même méthodologie. Au total, il est possible de distinguer le sexe féminin ou masculin dans 88,1% des combinaisons article – auteur de notre corpus. Les femmes y sont présentes à hauteur de 32,5% et de 55,7% pour les hommes. Le reste est répartie dans les catégories INI et UNI pour les auteurs dont les initiales seulement sont recensées ou lorsque l'auteur a un prénom unisexe. Notre taux d'inconnus (8,9%) est légèrement supérieur à celui obtenu par Larivière et al. (8,4%)[20].

La figure 3 présente la part des auteurs féminins dans les articles de PLoS. La part des femmes sur l'ensemble des auteurs (H ou F) est en moyenne 35,6%[21]. C'est en santé que le pourcentage de femmes par article est le plus élevé avec 44,3%, suivent la psychologie (41,3%), la médecine clinique (37,1%), la recherche biomédicale (35,1%) et la biologie (31,4%). Les femmes représentent moins de 30% des auteurs dans les autres disciplines.

[20] Pour plus de détails : http://www.nature.com/polopoly_fs/7.14227.1386700530!/suppinfoFile/504211a_s1.pdf.

[21] Tel que calculé dans Larivière et al., 2013.

Figure 3. Pourcentage moyen de femmes auteures par article selon la discipline*



*10 principales disciplines en nombre d'auteur-article.

Résultats

Les contributions

Les résultats présentés dans cette section sont basés sur l'ensemble des articles (N = 85 131) comptant au moins une contribution parmi les cinq tâches les plus fréquentes telles que définies dans la méthodologie. La plupart des indicateurs sont construits sur la base de l'article. Par exemple, pour calculer le pourcentage moyen des auteurs par contribution selon la discipline présenté dans le tableau 13 nous avons d'abord calculé la proportion d'auteurs ayant effectué chacune des tâches au sein de chacun des articles et nous avons ensuite calculé la moyenne de ces proportions par discipline et par contribution.

Tableau 13. Pourcentage moyen des auteurs par contribution selon la discipline

	Analyse des données	Conception et design des expériences	Rédaction de l'article	Fourniture (réactifs/matériels/ outils d'analyses)	Réalisation des expériences
Toutes les disciplines	**58,8%**	**55,7%**	**56,6%**	**38,0%**	**52,8%**
Biologie	58,6%	64,0%	68,3%	59,3%	57,6%
Chimie	65,0%	61,1%	60,9%	54,4%	56,1%
Génie	63,7%	63,9%	68,9%	61,0%	51,9%
Médecine clinique	55,4%	53,9%	54,2%	48,6%	54,0%
Physique	71,7%	73,9%	80,4%	72,5%	59,2%
Psychologie	60,0%	76,9%	81,6%	64,1%	52,8%
Recherche biomédicale	63,1%	56,6%	55,4%	51,3%	56,2%
Santé	57,1%	66,7%	71,7%	58,3%	55,5%
Sciences de la terre et de l'espace	66,0%	64,5%	70,7%	65,1%	58,8%
Sciences sociales	66,9%	71,9%	80,5%	70,0%	61,4%

Pour l'ensemble des disciplines, la contribution des auteurs la plus fréquente est l'analyse des données avec une participation de 58,8% des chercheurs, alors que la fourniture de réactifs, de

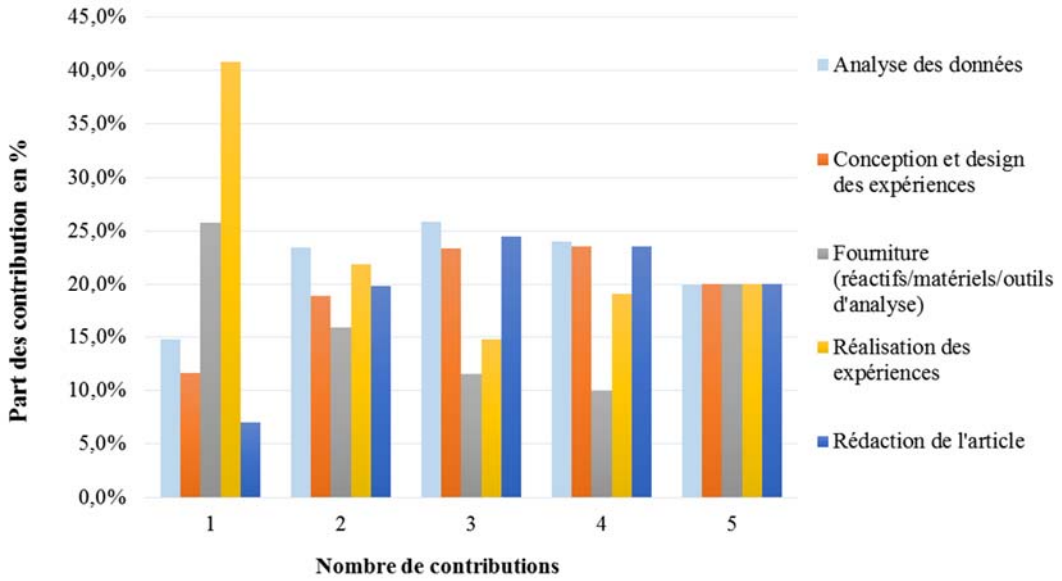
matériels ou d'outils d'analyse est, quant à elle, la moins fréquente avec 38% (cf. tableau 15). La rédaction de l'article occupe près de 57% des chercheurs.

Au niveau disciplinaire, les psychologues, les physiciens et des chercheurs en sciences sociales contribuent à la rédaction de l'article dans des proportions similaires avec 81,6%, 80,4% et 80,5% respectivement. Le pourcentage moyen des auteurs ayant contribués à la rédaction de l'article en médicine clinique (54,2%) et en recherche biomédicale (55,4%) se situe au-dessous de celui de l'ensemble des disciplines (56,6%).

La diversité des contributions

La figure 4 montre la répartition des contributions en pourcentage selon le nombre total de tâches réalisées par chaque chercheur.

Figure 4. Répartition des contributions (en %) selon le nombre de contributions réalisées par les auteurs



Pour chacun des auteurs, nous calculons d'abord le nombre total des contributions que nous répartissons ensuite selon leur nature pour calculer enfin la part qu'occupe chacune des contributions dans la participation globale de chaque auteur. Ainsi, un chercheur qui n'a participé qu'à la réalisation des expériences se voit attribuer une proportion de 100% pour ce type de contribution. Mais s'il a participé en plus à la rédaction de l'article, sa contribution sera calculée à 50% pour la réalisation des expériences et à 50% pour la rédaction. S'il a plutôt fourni trois types de contributions, chacune d'elle comptera pour 33% et ainsi de suite, jusqu'à cinq types de contribution. Cette façon de calculer ne tient évidemment pas compte du fait que l'effort requis par chaque contribution peut varier énormément de l'une à l'autre et, aussi, d'un chercheur à l'autre.

Le figure 4 montre ainsi que, lorsque les chercheurs ne réalisent qu'une seule tâche, la plus commune est la réalisation des expériences avec 40,7%, suivie de la fourniture des réactifs, du matériel ou des outils d'analyse avec près de 26%. La contribution la moins fréquente est la rédaction de l'article à seulement 7%. Autrement dit, lorsqu'un chercheur participe à la rédaction de l'article, cette tâche n'est que rarement la seule à laquelle il contribue. À trois contributions par chercheur, la conception et le design des expériences (23,4%), l'analyse des données (25,9%) et la rédaction de l'article (24,5) sont les activités les plus courantes. Le portrait est sensiblement le même lorsque les chercheur collaborent à quatre tâches. Évidemment, à cinq contributions, la répartition est uniforme à 20% pour chacune.

Ces résultats suggèrent que l'étendu de la collaboration est déterminée par le type de contribution attendu au sein de l'équipe. Par exemple, si un chercheur a participé à la

conception et au design des expériences, il y a de bonnes chances qu'il fasse également partie prenante de l'analyse et de la rédaction. *A contrario*, s'il n'a réalisé que les expériences ou fourni que les réactifs, il est peu probable qu'il participe aussi à la rédaction de l'article. Ces premiers résultats peuvent être précisés davantage en créant une matrice de co-contributions.

Les contributions conjointes

Le tableau 16 présente les co-contributions des auteurs. La matrice est construite sur la base des tâches effectuées par un même auteur. Elle montre que 43,6 % des chercheurs ayant contribués à la conception et au design des expériences ont aussi participés à l'écriture de l'article. Ces derniers sont également 41,6% à avoir à la fois analysé les données et rédigé l'article.

Tableau 14. Matrice du pourcentage moyen du nombre d'auteurs selon les co-contributions

	Analyse des données	Conception et design des expériences	Rédaction de l'article	Fourniture (réactifs/matériels/ outils d'analyse)	Réalisation des expériences
Analyse des données		39,0%	41,6%	22,2%	34,9%
Conception et design des expériences	39,0%		43,5%	22,3%	28,4%
Rédaction de l'article	41,6%	43,5%		22,1%	28,8%
Fourniture (réactifs/matériels /outils d'analyse)	22,2%	22,3%	22,1%		17,6%
Réalisation des expériences	34,9%	28,4%	28,8%	17,6%	

La conception des expériences et l'analyse vont aussi de pair. En effet, 39% des auteurs ont contribué à ces deux tâches au sein d'un même article. Il en va de même avec la réalisation des expériences et l'analyse des données (34,9%).

La réalisation des expériences et la fourniture de réactifs, de matériels ou d'outils d'analyse sont rarement l'œuvre d'un même chercheur. En effet, seulement 17,6 % des chercheurs ont contribué à ces deux tâches au sein d'un même article. Les chercheurs ayant réalisé les expériences sont associés à la rédaction de l'article dans moins de 30% des cas. Cela baisse à 22,1% pour ceux qui contribuent à la fourniture de réactifs, de matériels ou d'outils d'analyse.

Au regard de ces premiers résultats, l'existence de trois classes de contributions semblent apparaître. Celles liés directement à la conception et à la rédaction, celles liées aux tâches plus cléricales et, troisièmement, celles où la fourniture de fourniture de réactifs, de matériels ou d'outils d'analyse justifie à elle seule l'attribution du statut d'auteur.

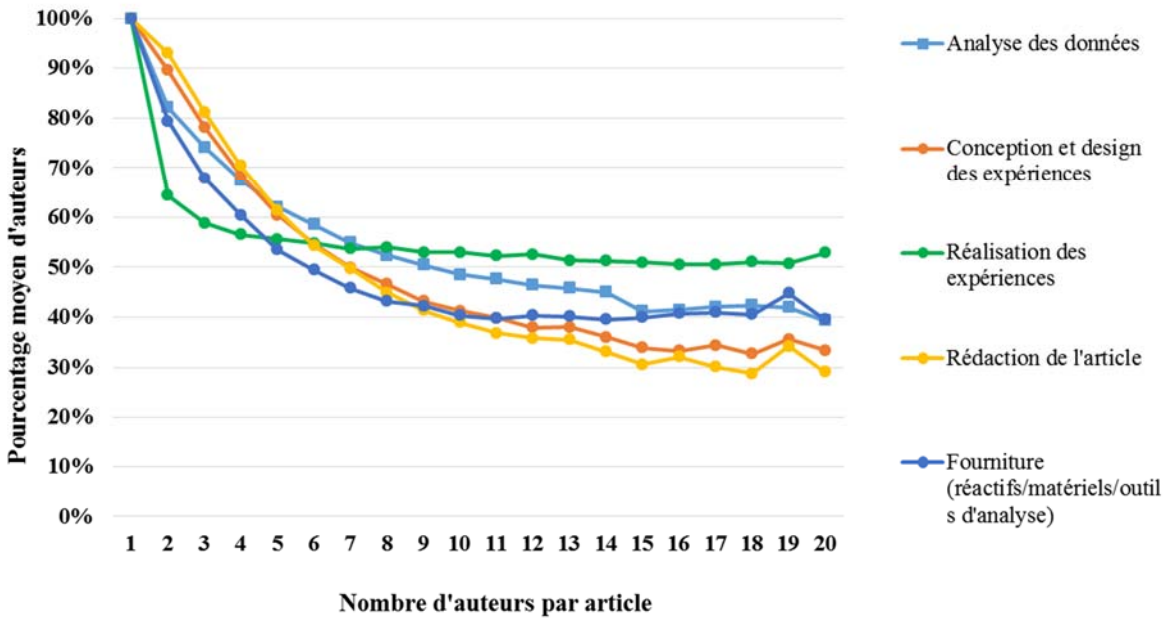
La taille de l'équipe de recherche

Comme nous l'avons vu précédemment, les équipes de recherche sont plus grandes que jamais et cela constitue une tendance lourde depuis plusieurs années. En ce sens, il est légitime de se poser la question : Est-ce que la taille de l'équipe affecte la distribution des tâches liées à l'activité scientifique? Pour y répondre nous avons calculé ci-dessous le pourcentage moyen d'auteurs par type de contribution en fonction du nombre total d'auteurs par article (cf. figure 5).

Pour un article donné, le nombre d'auteur ayant contribué à une tâche spécifique est divisé par le nombre total d'auteur. Par la suite, la proportion moyenne des auteurs affectés à chaque tâche est regroupée selon le type de contribution et le nombre total d'auteurs par article. Évidemment les auteurs solos réalisent 100 % de la tâche. Par exemple, si un auteur solo

affirme avoir réalisé trois tâches, il a réalisé 100% de celles-ci. Toutes les contributions ne sont pas présentes dans tous les articles (cf. tableau 6).

Figure 5. Pourcentage moyen des auteurs par contribution selon le nombre total d'auteurs par article



Le premier constat est que plus l'équipe grossit plus la proportion des auteurs affectés à chacune des tâches diminue. Autrement dit, la division du travail s'accentue, ce qui est tout à fait logique. Ainsi, la proportion d'auteurs qui contribuent à la rédaction passe de 82,2% dans une équipe de trois personnes, à 29,1 % dans une équipe de vingt personnes. À partir d'une équipe composée de cinq personnes, le pourcentage d'auteurs contribuant à la réalisation des expériences diminue moins rapidement que la part des auteurs collaborant à la conception et au design des expériences et à la rédaction de l'article. Lorsque les équipes deviennent assez grandes, à huit personnes ou plus, la réalisation des expériences se révèle en fait comme la

tâche à laquelle participe le plus grand nombre d'auteurs avec une proportion supérieure à 50%. La croissance des équipes accentue par ailleurs l'écart entre, d'une part, les contributions liées au travail plus clérical (analyse et expérience) ou à la fourniture de matériels et, d'autre part, les contributions normalement plus valorisées au sein des équipes de recherche comme la conception et le design des expériences et la rédaction de l'article qui deviennent assumées par une part des membres de l'équipe de moins en moins grande. Par exemple, lorsque l'équipe de comprend vingt personnes, moins de 35% d'entre elles participent à la conception de l'expérience ou à la rédaction de l'article, 40% participent à l'analyse des résultats ou fournissent du matériel, alors que plus de 50% réalisent les expériences.

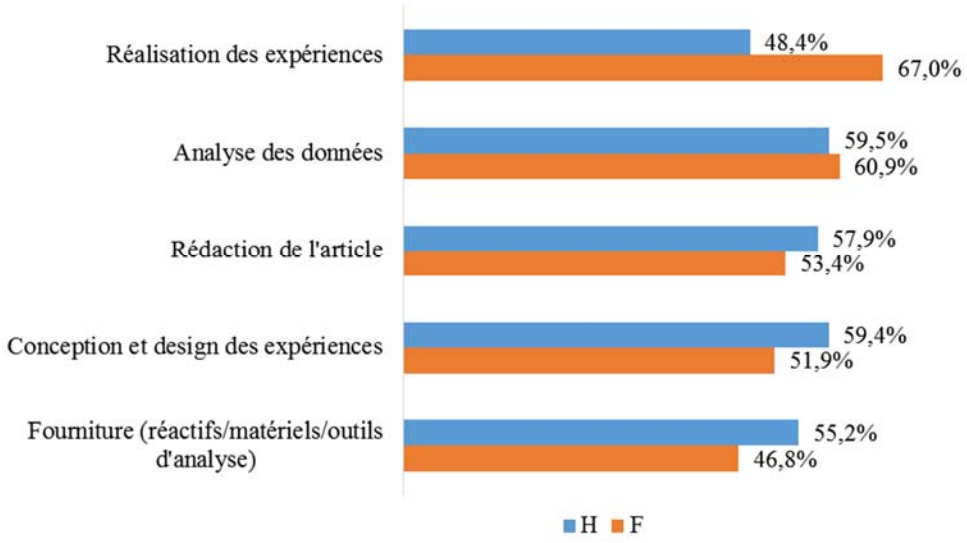
Les contributions selon le sexe

Les résultats présentés dans cette section sont issus de notre corpus comprenant l'ensemble des articles contenant au moins une contribution parmi les cinq tâches les plus fréquentes identifiées dans la méthodologie et au moins un auteur dont le sexe est attribué, homme ou femme (N = 84 141).

Ces données sont compilées sur la base de l'article et du sexe. Ainsi, tous les pourcentages sont calculés, non pas en fonction du nombre total de chercheurs, mais bien en fonction du nombre total d'auteurs de chacun des sexes. Par exemple, pour un article donné comptant cinq auteurs dont deux femmes et trois hommes, si l'une de ces deux des femmes rédige l'article et que les deux participent à la réalisation des expériences, on comptera donc 50% des femmes ayant collaboré à la rédaction et 100%, à la réalisation des expériences. La moyenne de ces

pourcentages est effectuée sur l'ensemble des articles et regroupés par différentes variables comme par exemple, le nombre total d'auteurs dans les articles ou encore, le sexe du premier auteur. Il est ainsi possible de calculer la distribution des divers types de contributions selon le sexe. Celle-ci révèle qu'au sein d'une équipe de recherche, 67,0% des femmes réalisent les expériences contre 48,4% des hommes (cf. figure 6). Cet écart entre les hommes et les femmes nous apparaît immense et fera l'objet d'une attention particulière dans les prochains paragraphes. En fait, ceci est sans contredit notre résultat le plus intéressant! Dans la même veine, nous remarquons aussi que l'analyse des données est accomplie par 60,9% des femmes et 59,5% des hommes.

Figure 6. Pourcentage moyen des hommes et des femmes par contribution



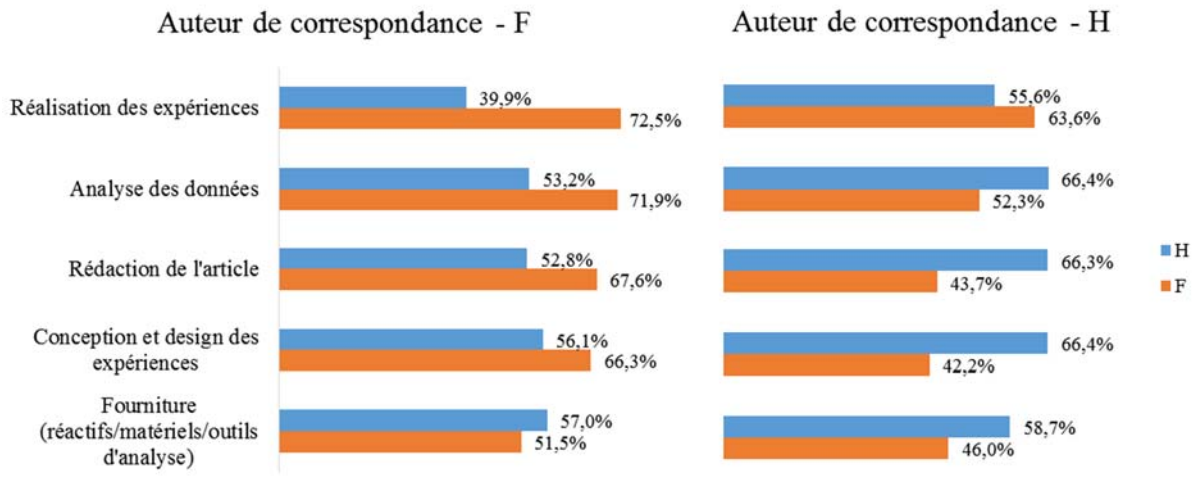
En revanche, nous voyons que les hommes sont proportionnellement plus nombreux à réaliser les contributions normalement plus valorisées au sein des équipes de recherche, soit la conception et le design des expériences, à 59,4% contre 51,9% pour les femmes; et la rédaction de l'article, à 57,9% contre 53,4% pour les femmes. Enfin, les hommes sont aussi plus nombreux à fournir des ressources matérielles, à 55,2% contre 46,8% pour les femmes.

Ces données tendent donc à démontrer que, globalement, les femmes sont davantage assignées à des tâches cléricales et qu'elles sont moins nombreuses à assumer les tâches liées à la conception et au design des expériences. Il est toutefois possible que de tels écarts dépendent moins de la discrimination dont les femmes pourraient être victimes que de leur âge moyen et des phases normales de progression des carrières scientifiques. Arrivées plus récemment en grand nombre dans des carrières scientifiques, les femmes seraient en moyenne plus jeunes que les hommes et cet écart d'âge pourrait expliquer à lui seul les différences constatées plus haut concernant leurs contributions moins prestigieuses à la réalisation des articles scientifiques. Si tel était bien le cas, l'écart entre les hommes et les femmes devrait donc être moins grand (voire inexistant) lorsque les femmes sont en position d'autorité ou lorsqu'elles peuvent être considérées comme des chercheurs seniors.

Les contributions selon le sexe de l'auteur de correspondance et du premier auteur

La figure 7 présente la proportion des femmes et des hommes par contribution en fonction du sexe de l'auteur correspondant, qui est habituellement l'auteur principal de l'article, celui avec qui les éditeurs de la revue et les lecteurs intéressés entreront en contact. Nous constatons ainsi que, lorsque l'auteur correspondant de l'article est une femme, la proportion de femmes qui exécutent l'expérimentation est de 72,5% contre 39,9% pour les hommes. Et en fait, sauf pour la fourniture de réactifs, de matériels ou d'outils d'analyse, les femmes sont proportionnellement plus nombreuses à réaliser l'ensemble des tâches, y compris celles qui sont les plus valorisées (écriture et conception). Ceci suggère que les équipes dirigées par des femmes font une plus grande place à la contribution des femmes que celles dirigées par les hommes.

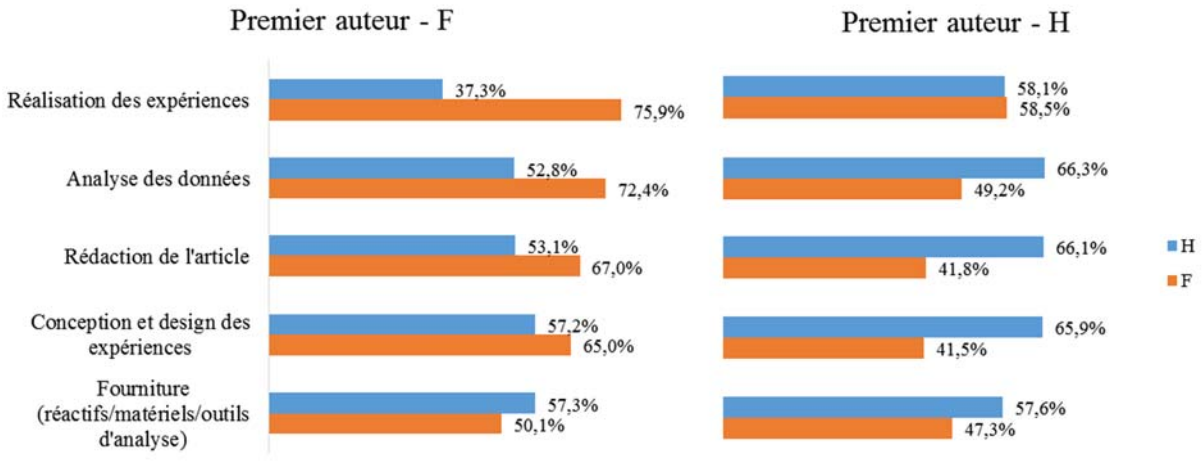
Figure 7. Pourcentage moyen des hommes et des femmes par contribution selon le sexe de l'auteur de correspondance



Réciproquement, lorsque l'auteur correspondant est un homme, nous constatons que les hommes contribuent proportionnellement davantage à toutes les tâches, sauf à la réalisation des expériences, à laquelle 63,6% des femmes contribuent contre 55,6% des hommes. La plus grande différence entre les rôles masculins et féminins dans les articles dirigés par des hommes est dans la conception et le design des expériences et la rédaction, où les hommes sont plus de 50% plus susceptibles d'effectuer ces tâches que les femmes.

Les résultats en fonction du sexe du premier auteur dévoilent les mêmes tendances. Les femmes jouent un rôle plus important dans les publications dont le premier auteur est une femme (cf. figure 8).

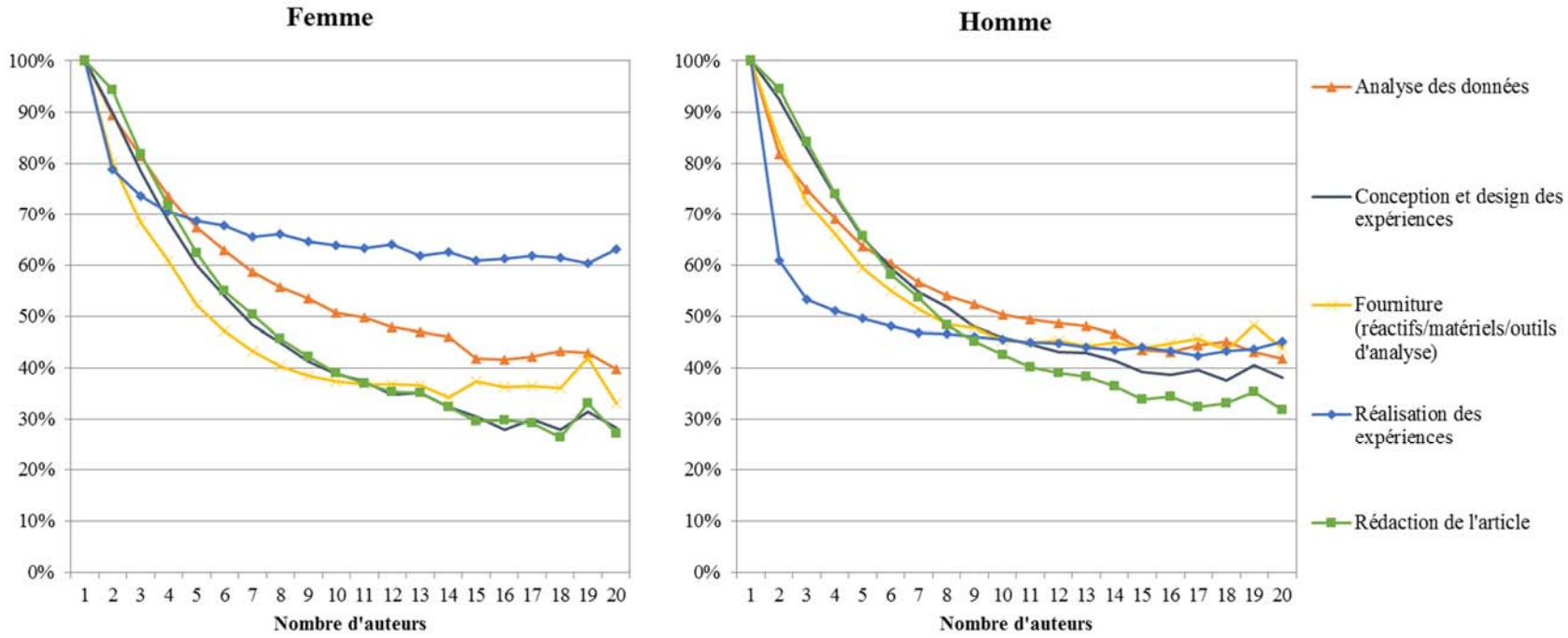
Figure 8. Pourcentage moyen des hommes et des femmes par contribution selon le sexe du premier auteur



La taille de l'équipe de recherche

Comme nous avons pu l'observer dans les résultats sur la relation entre contribution et nombre d'auteurs, la taille de l'équipe a une incidence sur la distribution des tâches et cela se manifeste aussi lorsque les contributions sont ventilées selon le sexe. La figure 9 (page suivante) montre, encore une fois, que les femmes sont proportionnellement plus nombreuses à réaliser les expériences. Dès que l'équipe atteint six personnes, au moins 64% d'entre elles réalisent les expériences et ce, peu importe la taille de l'équipe par la suite. Chez les hommes, à partir de cinq membres, le pourcentage moyen de ceux qui effectuent les expériences est à près de 50% et diminue à 45 % au sein d'une équipe de plus de dix personnes.

Figure 9. Pourcentage moyen des hommes et des femmes par contribution selon le nombre d'auteur par article

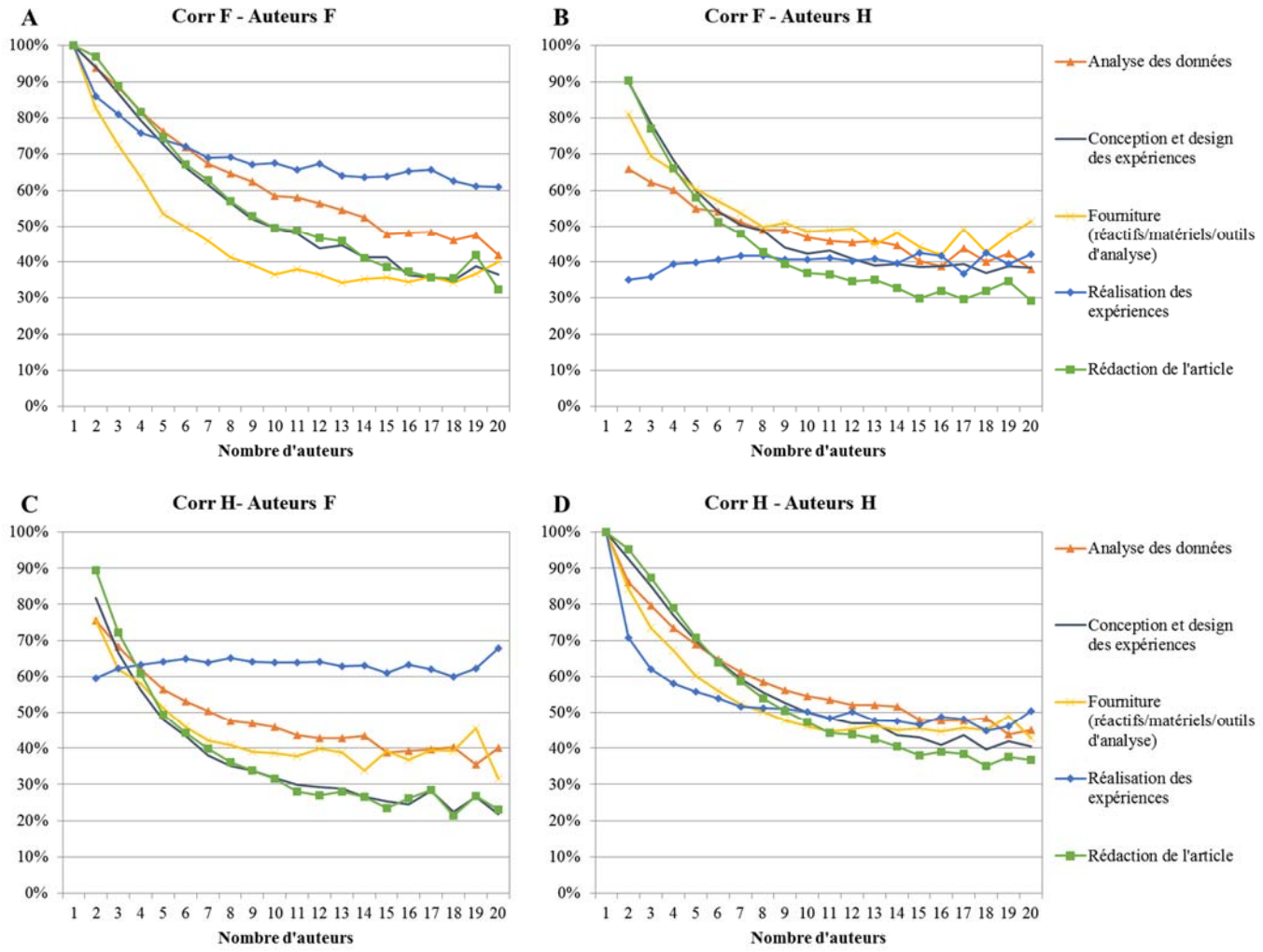


La figure 10 présente le pourcentage moyen d'hommes et de femmes par contribution selon le nombre d'auteurs et le sexe de l'auteur correspondant.

Les portions A et C montrent le pourcentage moyen de femmes par contribution quand l'auteur correspondant est une femme (A) ou un homme (C). Réciproquement, les portions B et D montrent le pourcentage moyen d'hommes par contribution quand l'auteur correspondant est une femme (B) ou un homme (D).

Encore une fois, ces données montrent que les femmes jouent proportionnellement un rôle plus important dans les équipes dirigées par des femmes (portion A) que dans celles dirigées par des hommes (portion C), notamment au niveau de l'écriture et de la conception des expériences. On constate également que les hommes sont proportionnellement moins nombreux à réaliser les expériences (autour de 40%) dans les équipes dirigées par les femmes (portion B) que ne le sont les femmes (60% et plus) dans les équipes dirigées par les hommes (portion C). Inversement, les hommes sont proportionnellement plus nombreux (au-delà de 35%) à participer à l'écriture et à la conception dans les grandes équipes dirigées par des hommes (portion D) que ne le sont les femmes (à moins de 30%, portion C).

Figure 10. Pourcentage moyen des hommes et des femmes par contribution selon le nombre d'auteurs et le sexe de l'auteur de correspondance

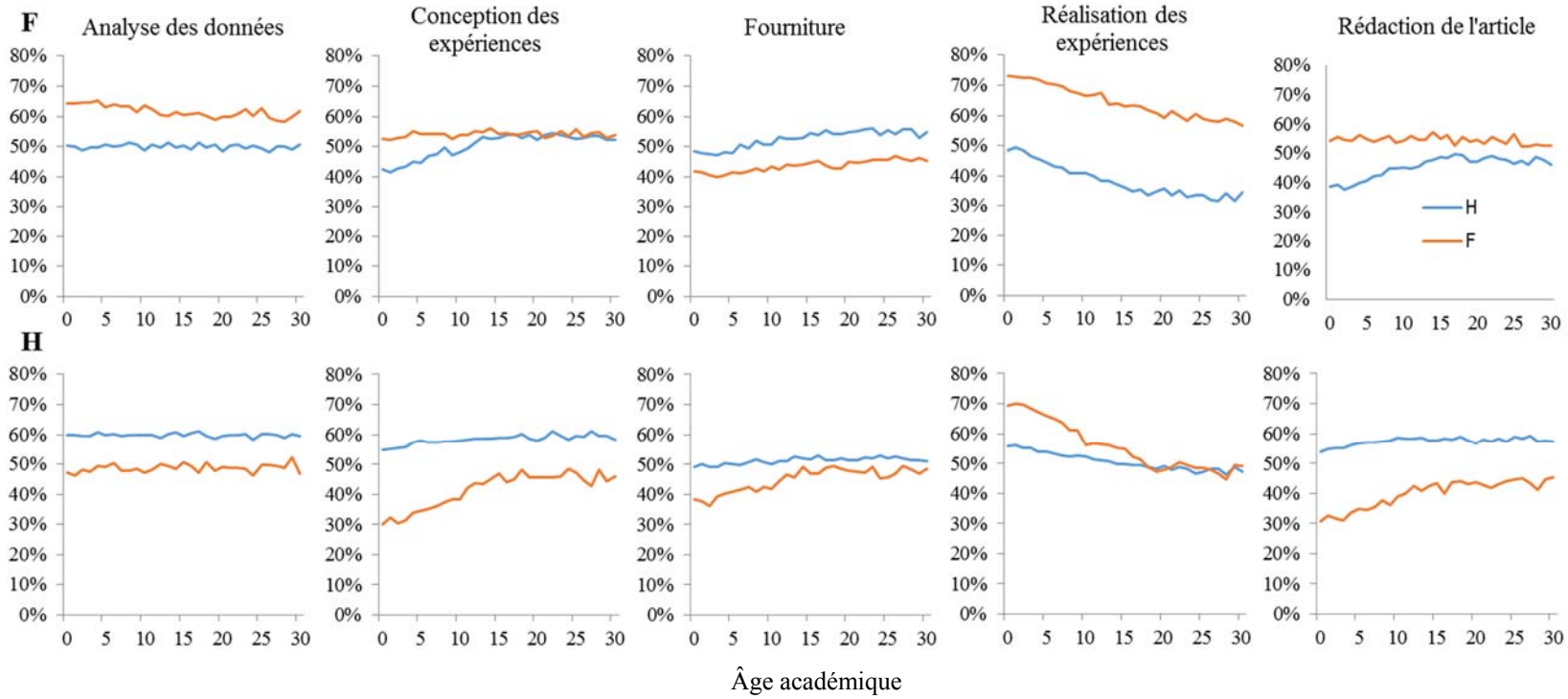


Il faut toutefois remarquer que les femmes participent toujours plus que les hommes à la réalisation des expériences dans les grandes équipes et ce, indépendamment du sexe de l'auteur correspondant. En effet, lorsque l'équipe compte dix membres ou plus, la proportion moyenne des femmes qui réalisent des expériences est toujours supérieure à 60% (portions A et C) alors que celle des hommes demeure inférieure à 50% (portions B et D). Il semble donc que, pour accéder au statut d'auteur, les femmes doivent plus souvent que les hommes se résoudre aux tâches moins valorisées liées à la réalisation des expériences. Cela dit, cet écart s'explique peut-être, du moins en partie, par le plus jeune âge moyen des femmes. Cette hypothèse est toutefois largement invalidée par les données de la section suivante.

Contributions en fonction de l'âge académique

La figure 11 montre le pourcentage moyen d'auteurs par contribution en fonction de l'âge académique et du sexe. L'âge académique est défini ici comme la différence entre l'année de publication du dernier et du premier article scientifique du chercheur. Sans surprise, les types de contributions varient selon l'âge. Par exemple, les auteurs sont un peu plus susceptibles de réaliser les expériences en début carrière que de contribuer en nature par des fournitures.

Figure 11. Pourcentage moyen des hommes et des femmes par contribution selon leur âge académique et le sexe de l'auteur correspondant (F = auteur de correspondance femme et H = auteur de correspondance homme)



De façon générale, la conception des expériences et la rédaction des articles deviennent aussi un peu plus fréquentes avec la progression dans la carrière. Reste que plusieurs écarts entre sexes persistent malgré la progression des carrières et que, dans certains cas, le sexe des responsables d'équipes (auteurs correspondants) semble jouer un rôle déterminant dans le maintien ou l'atténuation de ces écarts. On notera d'abord que la fourniture de matériel scientifique est plus fréquente chez les hommes et ce, indépendamment de l'âge académique des auteurs ou même du sexe de l'auteur de correspondance.

Inversement, ce sont les femmes qui participent le plus aux expériences et ce, dans les équipes dirigées tant par des hommes que par des femmes. Dans les équipes dirigées par les hommes toutefois, l'écart homme/femme à cet égard disparaît entre la 15^{e} et la 20^{e} année de la carrière. Dans les deux cas, c'est un peu moins de 50% des chercheurs plus âgés qui réalisent les expériences. En revanche, dans les équipes dirigées par les femmes, bien que les deux sexes voient leur engagement dans les expériences diminuer à mesure que progressent leurs carrières, l'écart homme/femmes persiste toujours. En effet, tandis qu'à 30 ans de carrière, la proportion des hommes engagés dans les expériences se situe à un peu moins de 40%, cette part s'élève à 60% chez les femmes.

L'analyse des données, quant à elle, est davantage pratiquée par les femmes dans les équipes dirigées par les femmes et pratiquée par des hommes dans les équipes dirigées par des hommes. En ce qui concerne enfin les tâches les plus valorisées de la conception des expériences et de la rédaction des articles, on constate encore une fois que le sexe de l'auteur de correspondance influence notablement les tendances. Alors que, dans les équipes dirigées

par des femmes, les hommes sont sensiblement moins engagés dans ces tâches que les femmes, l'inverse se produit dans les équipes dirigées par des hommes. Il faut remarquer par contre qu'avec la progression dans la carrière, l'écart entre les sexes se rétrécie davantage dans les équipes dirigées par des femmes que dans celles dirigées par des hommes. On constate en effet que la participation des hommes à la conception des expériences rejoint complètement celle des femmes à partir de la 15^{e} année de carrière dans les équipes dirigées par les femmes, alors que dans les équipes dirigées par des hommes, un écart important persiste tout au long des carrières (bien qu'il se rétrécisse). Dans la même veine, l'écart hommes/femmes dans la participation à la rédaction des articles diminue davantage dans les équipes dirigées par des femmes, à mesure que progressent les carrières académiques.

En somme, si la progression normale des carrières de chercheurs peut expliquer une partie des écarts hommes/femmes dans la réalisation des divers types de contributions aux articles scientifiques, elle est loin d'en rendre compte au complet. Quel que soit leur âge académique en effet, les femmes sont davantage assignées à la réalisation des expériences que les hommes et elles sont aussi moins susceptibles de participer à la conception des expériences et à la rédaction des articles, surtout lorsqu'elles évoluent dans des équipes dirigées par des hommes.

Discussion et conclusion

Notre recherche a démontré que l'utilisation des libellés de contributions aux articles scientifiques à des fins bibliométriques est possible et ce, même à grande échelle. L'analyse de contenu et la normalisation des contributions des auteurs nous ont permis de dégager cinq grandes catégories de contributions. Finalement, nous avons rempli notre objectif concernant la création d'un corpus intelligible pour l'étude de la division du travail en science, puisqu'elle ouvre à l'analyse bibliométrique plusieurs pistes de recherche encore inexplorées.

À travers deux études de cas portant sur la division du travail de recherche au sein des équipes de recherche, nous avons démontré le caractère opérationnel de notre méthode et sa capacité à livrer des résultats empiriques à la fois originaux et probants.

Beaucoup de mains font le travail léger[22]

De ce corpus, nous avons été en mesure de tirer l'information nécessaire à une première analyse des contributions de chacun des chercheurs dans le cadre d'activités menant à la publication d'un article scientifique. Nous avons montré l'existence de trois classes de contributions : celles qui sont liées à la conceptualisation et à la rédaction de l'article, celles qui sont liées au travail plus clérical de réalisation des expériences et des analyses et, enfin, celles qui donnent droit au statut d'auteur simplement grâce à la fourniture d'éléments matériels (réactifs, outils d'analyse, etc) utilisés dans le cadre de la recherche. Il va sans dire

[22] Traduction du proverbe *Many handis make light warke* de John Heywood in *The Proverbs and Epigrams of John Heywood*, 1562.

que l'univers des possibles est ici restreint à cinq contributions, mais cela n'empêche pas de constater la forte affinité de certaines contributions l'une pour l'autre et, inversement, l'incompatibilité apparente de certaines l'une pour l'autre.

Nous avons également pu montrer comment la division du travail s'accentue à mesure que la taille des équipes s'agrandit. Il est vrai que tous ne peuvent tenir le crayon, cependant l'augmentation de l'écart entre les contributions liées à un travail plus clérical, aux dons de matériels et celles habituellement plus valorisées au sein des équipes de recherche comme la conception et la rédaction supposent une structure hiérarchique établie.

L'utilisation d'une méthodologie éprouvée pour l'identification du sexe des auteurs nous a aussi permis de bonifier notre mémoire d'une deuxième étude de cas originale sur la contribution spécifique des auteurs en science selon leur sexe.

L'effet Cendrillon

Nous avons examiné la nature de la division du travail telle que mesurée par les libellés de contribution en fonction du sexe des auteurs. Plus précisément, nous avons comparé la nature et l'intensité de la contribution des co-auteurs en fonction de leur sexe et d'autres critères comme la taille des équipes de recherche, le sexe des chefs d'équipe et l'âge académique des auteurs. Nos résultats montrent que la collaboration ne semble pas conduire à une répartition équitable du travail. La prédominance des femmes dans la réalisation des expériences a été le résultat le plus surprenant. Il existe donc des inégalités dans la répartition du travail

scientifique; les femmes étant plus susceptibles d'être associées au travail clérical alors que les hommes sont plus susceptibles d'être associés aux contributions en matériels et au travail conceptuel. Cendrillon, dans le célèbre conte des frères Grimm, est réprimandée par sa sœur et renvoyée de la table avec l'avertissement : « *Away to the kitchen with her! And if she wants to eat, then she must earn it ».* Tout comme Cendrillon, les femmes en science doivent faire des efforts plus grands pour être créditées comme auteures.

La surreprésentation des femmes dans les rôles de production où le travail clérical prédomine constitue en soi un problème. Certains diront que cela est dû à la stratification de l'âge en science—avec plus de femmes dans les rangs juniors elles sont plus susceptibles d'être associées à des tâches subalternes. Cependant, comme nous l'avons vu, les femmes contribuent majoritairement à la réalisation des expériences et à l'analyse des résultats et ce, peu importe l'âge académique**,** leur statut ou encore la position dans la hiérarchie; comme première auteure ou auteure correspondant. Cela pourrait expliquer les disparités dans le taux de production scientifique entre les hommes et les femmes (Larivière et al., 2013) et ce, particulièrement dans les prestigieuses premières et dernières positions dans la liste des auteurs (West, Jacquet, King, Correll et Bergstrom, 2013).

Les femmes doivent publier au moins trois articles de plus que leurs homologues masculins dans des revues prestigieuses afin d'être considérées pour les postes de post-doctorant (Wenneras et Wold, 1997). Ces inégalités ont encore été démontrées dans une étude en double aveugle dans laquelle les chercheurs ont fourni des curriculum vitae identiques, avec des prénoms féminins et masculins assignés au hasard (Moss-Racusin, Dovidio, Brescoll, Graham

et Handelsman, 2012). À cet égard, les chercheurs masculins ont plus de chance de publier parce qu'ils reçoivent une grande partie des ressources. En effet, les femmes publient beaucoup moins d'articles dans les disciplines où les dépenses de recherche sont les plus élevées (Duch et al., 2012) ce qui pourrait expliquer le plus faible support reçu par les femmes de la part des institutions et des organismes subventionnaires (Ley et Hamilton, 2008).

D'autres ont aussi suggéré que les femmes ne postulent pas sur les postes de direction menant à la permanence et ce, en grande partie en raison de la famille et de leur mode de vie (Ceci et Williams, 2011). Il existe cependant des politiques de conciliations travail-familles[23] offert par les organismes subventionnaires et des bourses de maternités[24] proposées par certaines universités. Ces exemples de dispositions démontrent le souhait de ces institutions à maintenir ou faciliter l'intégration des femmes aux postes de professeurs.

Les prochaines questions ?

Les libellés de contribution permettent de répondre à d'autres questions sur l'activité scientifique non-traitées ici. Nous en avons déjà soumis quelques-unes précédemment et comme elles nous semblent intéressantes, les revoici : Comment la progression de la carrière des chercheurs affecte-t-elle l'évolution de leurs contributions aux articles qu'ils signent? Comment les diverses structures institutionnelles de la recherche affectent-elles la division du travail scientifique entre co-auteurs? Comment se manifeste la répartition géographique ou

[23] http://www.nserc-crsng.gc.ca/NSERC-CRSNG/policies-politiques/Wleave-Fconges_fra.asp.
[24] http://fesp.umontreal.ca/fileadmin/Documents/Soutien_financier/formulaire_maternite_Ete_2015.pdf.

socio-économique des contributions? Les recherches effectuées en collaboration entre auteurs des pays en voie de développement et auteurs des pays développés sont-elles équitables?

À la suite de nos résultats, l'analyse des libellés de contribution semble être une voie prometteuse pour apporter des éléments de réponse empiriques fiables et originaux à ces questions.

Bibliographie

Abramo, G., D'Angelo, C. A. et Murgia, G. (2013). Gender differences in research collaboration. Journal of Informetrics, 7(4), 811-822. doi : 10.1016/j.joi.2013.07.002

Aksnes, D. W., Rorstad, K., Piro, F. et Sivertsen, G. (2011). Are Female Researchers Less Cited? A Large-Scale Study of Norwegian Scientists. *Journal of the American Society for Information Science and Technology, 62*(4), 628-636. doi : 10.1002/Asi.21486

Atanassova, I. et Bertin, M. (2014). Faceted Semantic Search for Scientific Publications.

Barrios, M., Villarroya, A. et Borrego, A. (2013). Scientific production in psychology: a gender analysis. Scientometrics, 95(1), 15-23. doi : 10.1007/s11192-012-0816-4

Benos, D. J., Fabres, J., Farmer, J., Gutierrez, J. P., Hennessy, K., Kosek, D., Wang, K. (2005). Ethics and scientific publication. *Advances in Physiology Education, 29*(2), 59-74. doi : 10.1152/advan.00056.2004

Bertin, M., Atanassova, I., Gingras, Y. et Larivière, V. (à paraître). The invariant distribution of references in scientific articles. *Journal of the American Society for Information Science and Technology*.

Bertin, M. Atanassova., I; Larivière,V; Gingras,Y. (2014). The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. *14th International Society of Scientometrics and Informetrics Conference, Volume: 1*.

Biagioli, M. et Galison, P. (2003). *Scientific authorship : Credit and intellectual property in science*. New York, NY: Routledge.

Bird, K. S. (2011). Do women publish fewer journal articles than men? Sex differences in publication productivity in the social sciences. *British Journal of Sociology of Education, 32*(6), 921-937. doi : 10.1080/01425692.2011.596387

Birnholtz, J. P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science and Technology, 57*(13), 1758-1770. doi : 10.1002/Asi.20380

Bordons, M., Morillo, F., Fernández, M. T. et Gómez, I. (2003). One step further in the production of bibliometric indicators at the micro level: Differences by gender and

professional category of scientists. *Scientometrics, 57*(2), 159-173. doi : 10.1023/A:1024181400646

Borrego, A., Barrios, M., Villarroya, A. et Ollé, C. (2010). Scientific output and impact of postdoctoral scientists: a gender perspective. *Scientometrics, 83*(1), 93-101. doi : 10.1007/s11192-009-0025-y

Bošnjak, L. et Marušić, A. (2012). Prescribed practices of authorship: review of codes of ethics from professional bodies and journal guidelines across disciplines. *Scientometrics, 93*(3), 751-763. doi : 10.1007/s11192-012-0773-y

Braisher, T. L., Symonds, M. R. E. et Gemmell, N. J. (2005). Publication success in *Nature* and *Science* is not gender dependent. *Bioessays, 27*(8), 858-859. doi : 10.1002/Bies.20272

Bunz, U. (2005). Publish or perish: A limited author analysis of ICA and NCA journals. *Journal of Communication, 55*(4), 703-720. doi : 10.1111/j.1460-2466.2005.tb03018.x

Ceci, S. J. et Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences of the United States of America, 108*(8), 3157-3162. doi : 10.1073/pnas.1014871108

Claxton, L. D. (2005a). Scientific authorship Part 1. A window into scientific fraud? *Mutation Research-Reviews in Mutation Research, 589*(1), 17-30. doi : 10.1016/j.mrrev.2004.07.003

Claxton, L. D. (2005b). Scientific authorship Part 2. History, recurring issues, practices, and guidelines. *Mutation Research-Reviews in Mutation Research, 589*(1), 31-45. doi : 10.1016/j.mrrev.2004.07.002

Corley, E. A. (2005), How Do Career Strategies, Gender, and Work Environment Affect Faculty Productivity Levels in University-Based Science Centers?. *Review of Policy Research, 22*: 637-655. doi : 10.1111/j.1541-1338.2005.00161.x

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology, 52*(7), 558-569. doi : 10.1002/Asi.1097.Abs

Davis, P. J. et Gregerman, R. I. (1969). Parse analysis: a new method for the evaluation of investigators' bibliographies. *New England Journal of Medecine, 281*(18), 989-990. doi : 10.1056/NEJM196910302811805

Duch, J., Zeng, X. H. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K. et Amaral, L. A. N. (2012). The Possible Role of Resource Requirements and Academic Career-Choice Risk on Gender Differences in Publication Rate and Impact. *PLoS One, 7*(12). doi : 10.1371/journal.pone.0051332

Eggert, L. D. (2011). Best practices for allocating appropriate credit and responsibility to authors of multi-authored articles. *Frontiers in Psychology, 2*, 196. doi : 10.3389/fpsyg.2011.00196

Eloy, J. A., Svider, P., Chandrasekhar, S. S., Husain, Q., Mauro, K. M., Setzen, M. et Baredes, S. (2013). Gender Disparities in Scholarly Productivity within Academic Otolaryngology Departments. *Otolaryngology-Head and Neck Surgery, 148*(2), 215-222. doi : 10.1177/0194599812466055

Endersby, J. W. (1996). Collaborative research in the social sciences: Multiple authorship and publication credit. *Social Science Quarterly, 77*(2), 375-392.

Flanagin, A., Carey, L. A., Fontanarosa, P. B., Phillips, S. G., Pace, B. P., Lundberg, G. D. et Rennie, D. (1998). Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. *Jama-Journal of the American Medical Association, 280*(3), 222-224. doi : 10.1001/jama.280.3.222

Frandsen, T. F. et Nicolaisen, J. (2010). What is in a name? Credit assignment practices in different disciplines. *Journal of Informetrics, 4*(4), 608-617. doi : 10.1016/j.joi.2010.06.010

Gallivan, M. J. et Benbunan-Fich, R. (2007). Analyzing IS research productivity: an inclusive approach to global IS scholarship. European Journal of Information Systems, 16(1), 36-53. doi : 10.1057/palgrave.ejis.3000667

Garfield, E. (1979). Is Citation Analysis a Legitimate Evaluation Tool. *Scientometrics, 1*(4), 359-375. doi : 10.1007/Bf02019306

Gonzalez-Brambila, C. et Veloso, F. M. (2007). The determinants of research output and impact: A study of Mexican researchers. *Research Policy, 36*(7), 1035-1051. doi : 10.1016/j.respol.2007.03.005

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569-16572. doi : 10.1073/pnas.0507655102

Hoen, W. P., Walvoort, H. C. et Overbeke, A. J. P. M. (1998). What are the factors determining authorship and the order of the authors' names? : A study among authors of the Nederlands Tijdschrift voor Geneeskunde (Dutch Journal of Medicine). *Jama-Journal of the American Medical Association, 280*(3), 217-218. doi : 10.1001/jama.280.3.217

Horton, R. (1997). The signature of responsibility. *Lancet, 350*(9070), 5-6. doi : 10.1016/S0140-6736(05)66236-8

Hunt, R. (1991). Trying an authorship index. *Nature Photonics, 352*(6332), 187.

Hwang, S. S., Song, H. H., Baik, J. H., Jung, S. L., Park, S. H., Choi, K. H. et Park, Y. H. (2003). Researcher contributions and fulfillment of ICMJE authorship criteria: Analysis of author contribution lists in research articles with multiple authors published in *Radiology*. *Radiology, 226*(1), 16-23. doi : 10.1148/radiol.2261011255

ICMJE. (1988). Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *Annals of Internal Medicine, 108*(2), 258-265.

ICMJE. (2013). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. [en ligne] http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html (page consultée le 2015-07-03)

Keith, B., Layne, J. S., Babchuk, N. et Johnson, K. (2002). The context of scientific achievement: Sex status, organizational environments, and the timing of publication on scholarship outcomes. *Social Forces, 80*(4), 1253-1281. doi : 10.1353/sof.2002.0029

Larivière, V., Gingras, Y. et Archambault, É. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics, 68*(3), 519-533. doi : 10.1007/s11192-006-0127-8

Larivière, V. (2007). L'internationalisation de la recherche scientifique québécoise : comparaisons nationales, disciplinaires et effets de sexe, 1980-2005. *Compendium d'indicateurs de l'activité scientifique et technologique au Québec. L'internationalisation de la science et de la technologie* (p. 63-79): Institut de la statistique du Québec.

Lariviere, V., Vignola-Gagne, E., Villeneuve, C., Gelinas, P. et Gingras, Y. (2011). Sex differences in research funding, productivity and impact: an analysis of Quebec university professors. Scientometrics, 87(3), 483-498. doi : 10.1007/s11192-011-0369-y

Larivière, V., Ni, C., Gingras, Y., Cronin, B. et Sugimoto, C. R. (2013). Global gender disparities in science. *Nature, 504*(7479), 211-213. doi : 10.1038/504211a

Leahey, E. (2006). Gender differences in productivity: Research specialization as a missing link. *Gender & Society, 20*(6), 754-780. doi : 10.1177/0891243206293030

Lewison, G. (2001). The quantity and quality of female researchers: A bibliometric study of Iceland. *Scientometrics, 52*(1), 29-43. doi : 10.1023/A:1012794810883

Lewison, G. et Markusova, V. (2011). Female researchers in Russia: Have they become more visible? *Scientometrics, 89*(1), 139-152. doi : 10.1007/s11192-011-0435-5

Ley, T. J. et Hamilton, B. H. (2008). The Gender Gap in NIH Grant Applications. *Science, 322*(5907), 1472-1474. doi : 10.1126/science.1165878

Long, J. S. (1992). Measures of Sex-Differences in Scientific Productivity. *Social Forces, 71*(1), 159-178. doi : 10.2307/2579971

Marušić, A., Bošnjak, L. et Jerončić, A. (2011). A Systematic Review of Research on the Meaning, Ethics and Practices of Authorship across Scholarly Disciplines. *Plos One, 6*(9), e23477. doi : 10.1371/journal.pone.0023477

Maske, K. L., Durden, G. C. et Gaynor, P. E. (2003). Determinants of scholarly productivity among male and female economists. *Economic Inquiry, 41*(4), 555-564. doi : 10.1093/Ei/Cbg027

Mendlowicz, M. V., Coutinho, E. S. F., Laks, J., Fontenelle, L. F., Valença, A. M., Berger, W., . . . de Aguiar, G. A. (2011). Is there a 'gender gap' in authorship of the main Brazilian psychiatric journals at the beginning of the 21st century? *Scientometrics, 86*(1), 27-37. doi : 10.1007/s11192-010-0296-3

Merton, R. K. (1968). The Matthew Effect in Science. *Science, 159*(3810), 56-63. doi

Moed, H. F. (1996). Differences in the construction of SCI based bibliometric indicators among various producers: A first overview. *Scientometrics, 35*(2), 177-191. doi : 10.1007/Bf02018476

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. et Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America., 109*(41), 16474-16479. doi : 10.1073/pnas.1211286109

Moulopoulos, S. D., Sideris, D. A. et Georgilis, K. A. (1983). For debate . . . Individual contributions to multiauthor papers. *British Medical Journal, 287*(6405), 1608-1610. doi : 10.1136/bmj.287.6405.1608

Nature. (1999). Policy on papers' contributors. *Nature, 399*(6735), 393. doi : 10.1038/20743

Nature. (2009). Authorship policies. *Nature, 458*(7242), 1078. doi : 10.1038/4581078a

Nalimov, V.V., et Mulchenko, Z.M. (1969), Eshche raz k voprosu o kontseptsii eksponentsial'nogo rosta. [A word to add on the exponential growth concept.] *Nauchno-Tekhnicheskaya Informatsiya*. 8, 12-14. [Traduction anglaise in: *Automatic Documentation and Mathematical Linguistics*. 3 (1969) 37-40.]

Nourmohammadi, H. et Hodaei, F. (2014). Perspective of Iranian women's scientific production in high priority fields of science and technology. *Scientometrics, 98*(2), 1455-1471. doi : 10.1007/s11192-013-1098-1

Østby, G., Strand, H., Nordås, R. et Gleditsch, N. P. (2013). Gender Gap or Gender Bias in Peace Research? Publication Patterns and Citation Rates for *Journal of Peace Research*, 1983-2008. *International Studies Perspectives, 14*(4), 493-506. doi : 10.1111/Insp.12025

Penas, C. S. et Willett, P. (2006). Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science, 32*(5), 480-485. doi : 10.1177/0165551506066058

Pontille, D. (2004). *La signature scientifique : une sociologie pragmatique de l'attribution*. Paris: CNRS.

Price, D. J. D. (1965). Networks of Scientific Papers. *Science, 149*(3683), 510-515. doi : 10.1126/science.149.3683.510

Pritchar.A. (1969). Statistical Bibliography or Bibliometrics. *Journal of Documentation, 25*(4), 348-349.

Rennie, D. et Flanagin, A. (1994). Authorship - Authorship - Guests, Ghosts, Grafters, and the 2-Sided Coin. *Jama-Journal of the American Medical Association, 271*(6), 469-471. doi : 10.1001/jama.271.6.469

Rennie, D., Flanagin, A. et Yank, V. (2000). The contributions of authors. *Jama-Journal of the American Medical Association, 284*(1), 89-91. doi : 10.1001/jama.1997.03550070071041

Rennie, D., Yank, V. et Emanuel, L. (1997). When authorship fails - A proposal to make contributors accountable. *Jama-Journal of the American Medical Association, 278*(7), 579-585. doi : 10.1001/jama.278.7.579

Resnik, D. B. et Master, Z. (2011). Criteria for Authorship in Bioethics. *American Journal of Bioethics, 11*(10), 17-21. doi : 10.1080/15265161.2011.603795

Rigg, L. S., McCarragher, S. et Krmenec, A. (2012). Authorship, Collaboration, and Gender: Fifteen Years of Publication Productivity in Selected Geography Journals. *The Professional Geographer, 64*(4), 491-502. doi : 10.1080/00330124.2011.611434

Shapiro, D. W., Wenger, N. S. et Shapiro, M. F. (1994). The Contributions of Authors to Multiauthored Biomedical Research Papers. *Jama-Journal of the American Medical Association, 271*(6), 438-442. doi : 10.1001/jama.1994.03510300044036

Small, H. G. (1973). Co-Citation in Scientific Literature - New Measure of Relationship between two Documents. *Journal of the American Society for Information Science, 24*(4), 265-269. doi : 10.1002/asi.4630240406

Smith, E. et Williams-Jones, B. (2012). Authorship and Responsibility in Health Sciences Research: A Review of Procedures for Fairly Allocating Authorship in Multi-Author Studies. *Science and Engineering Ethics, 18*(2), 199-212. doi : 10.1007/s11948-011-9263-5

Smith, M. J., Jordan, C. J. et Walsby, J. C. (2011). Credit where credit's due: developing authorship strategies in the geosciences. *Proceedings of the Geologists Association, 122*(1), 2-6. doi : 10.1016/j.pgeola.2010.10.001

Smith, R. (1997). Authorship: time for a paradigm shift? *British Medical Journal, 314*(7086), 992. doi : 10.1136/bmj.314.7086.992

Sotudeh, H. et Khoshian, N. (2014). Gender differences in science: The case of scientific productivity in Nano Science & Technology during 2005-2007. *Scientometrics, 98*(1), 457-472. doi : 10.1007/s11192-013-1031-7

Stack, S. (2004). Gender, children and research productivity. *Research in Higher Education, 45*(8), 891-920. doi : 10.1007/s11162-004-5953-z

Testa, J. (2010). The Thomson Reuters Journal Selection Process. [en ligne] http://wokinfo.com/essays/journal-selection-process/ (Page accédée le 2015-07-03)

Waltman, L. et van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378-2392. doi : 10.1002/Asi.22748

Wenneras, C. et Wold, A. (1997). Nepotism and sexism in peer-review. *Nature, 387*(6631), 341-343.

West, J. D., Jacquet, J., King, M. M., Correll, S. J. et Bergstrom, C. T. (2013). The Role of Gender in Scholarly Authorship. *Plos One, 8*(7). doi : 10.1371/journal.pone.0066212

Xie, Y. et Shauman, K.A., 2003. *Women in Science: Career Processes and Outcomes*. Boston, MA : Harvard University Press.

Yank, V. et Rennie, D. (1999). Disclosure of researcher contributions: A study of original research articles in *The Lancet. Annals of Internal Medicine, 130*(8), 661-670. doi : 10.7326/0003-4819-130-8-199904200-00013

Zuckerman, H. A. (1968). Patterns of Name Ordering among Authors of Scientific Papers - Study of Social Symbolism and Its Ambiguity. *American Journal of Sociology, 74*(3), 276-291. doi : 10.1086/224641

Annexe 1. Variables PloS

Variable	Description
doi	Digital Object Identifier
publication_date	Publication Date
crossref	Citations - CrossRef
pubmed	Citations - PubMed Central
scopus	Citations - Scopus
counter	Total combined HTML + PDF + XML usage at the PLoS site (from day of publication)
counter_html	HTML usage at the PLoS site (from day of publication)
counter_pdf	PDF usage at the PLoS site (from day of publication)
counter_xml	XML usage at the PLoS site (from day of publication)
pmc	Total usage recorded at the PubMedCentral site
pmc_html	HTML usage recorded at the PubMedCentral site
pmc_pdf	PDF usage recorded at the PubMedCentral site
nature	Blog & Media - Nature Blogs
bloglines	Blog & Media - Bloglines
researchblogging	Blog & Media - ResearchBlogging.org
postgenomic	Blog & Media - Postgenomic (which no longer exists)
scienceseeker	Blog & Media - Science Seeker
wikipedia	Blog & Media - Wikipedia
trackbacks	Number of Trackbacks
citeulike	Social Network- CiteULike
twitter	Social Network- Twitter
facebook	Social Network - Facebook Likes
mendeley_readers	Social Network - Mendeley Readers
mendeley_groups	Social Network - Mendeley Groups
mendeley	Social Network - Mendeley Total
comments	Number of Comment threads
comment_replies	Number of replies to Comments
biod	na
figshare	na
f1000	na

Annexe 2. Code SQL de la création de l'URL pour le téléchargement automatisé

```
    SELECT DISTINCT idplos, SUBSTRING(DOI, CHARINDEX('/', doi) + 1, LEN(DOI))
AS doi, SUBSTRING(URL, 1, CHARINDEX('.org', url) + 4) +
'article/fetchObjectAttachment.action?uri=info%3Adoi%2F10.1371%2F' +
SUBSTRING(DOI,CHARINDEX('/',doi) + 1,LEN(DOI)) + '&representation=XML'AS url
    FROM bdben.dbo.plos
    WHERE idplos = ? /*parameter*/
```

Annexe 3. Code C# pour le téléchargement automatisé

```
using System;
using System.IO;
using System.Data;
using Microsoft.SqlServer.Dts.Pipeline.Wrapper;
using Microsoft.SqlServer.Dts.Runtime.Wrapper;
using System.Xml;


[Microsoft.SqlServer.Dts.Pipeline.SSISScriptComponentEntryPointAttribute]
public class ScriptMain : UserComponent
{
   string searchXMLURL;
   string doi;
   string filename;
   string webResource;


   public override void Input0_ProcessInputRow(Input0Buffer Row)
   {

      searchXMLURL = Row.url.ToString();
      doi = Row.doi.ToString();
      filename = Variables.LocalFolder + doi + ".xml";

      System.Net.WebClient myWebClient = new System.Net.WebClient();
      string webResource = searchXMLURL;


      myWebClient.DownloadFile(webResource, filename);


   }
   }
```

Annexe 4. Code C# pour la lecture des fichiers XML et l'extraction de la variable contribution

```
using System;
using System.IO;
using System.Xml;
using System.Data;
using Microsoft.SqlServer.Dts.Pipeline.Wrapper;
using Microsoft.SqlServer.Dts.Runtime.Wrapper;


[Microsoft.SqlServer.Dts.Pipeline.SSISScriptComponentEntryPointAttribute]


public class ScriptMain : UserComponent
{
   string fichiersource = string.Empty;
   string JournalTitle = string.Empty;
   string fntype = string.Empty;
   string contribution = string.Empty;

   public override void CreateNewOutputRows()
   {
      fichiersource = Variables.destination;
      XmlTextReader xmt = new XmlTextReader(fichiersource);


      while (xmt.Read())
      {

         if ((xmt.Name == "fn") && (xmt.HasAttributes))
         {
            xmt.MoveToAttribute("fn-type");
            fntype = xmt.GetAttribute("fn-type");

            if (fntype == "conflict")
            {
               xmt.MoveToAttribute("fn-type");
               fntype = xmt.GetAttribute("fn-type");
               if (fntype == "con")
               {
                  xmt.ReadToFollowing("p");
                  contribution = xmt.ReadInnerXml();

                  this.Output0Buffer.AddRow();
                  this.Output0Buffer.NomFichier = Variables.NomFichier;
                  this.Output0Buffer.DOI = "";
                  this.Output0Buffer.Contribution = contribution;
                  xmt.Close();
               }

            }
         }
         else if (fntype == "con")
```

```
        {
            xmt.ReadToFollowing("p");
            contribution = xmt.ReadInnerXml();

            this.Output0Buffer.AddRow();
            this.Output0Buffer.NomFichier = Variables.NomFichier;
            this.Output0Buffer.DOI = "";
            this.Output0Buffer.Contribution = contribution;
            xmt.Close();



        }

        this.Output0Buffer.EndOfRowset();

    }

  }
}
```

Annexe 5. Code SQL pour la distribution des contributions et des initiales

```
-- Corpus de base tiré du WoS
SELECT  COUNT(DISTINCT ar.id_art)
FROM    Pub_Expanded.dbo.Article AS ar
WHERE   Code_Revue BETWEEN 13071 AND 13078 AND Annee_Bibliographique BETWEEN 2008 AND 2013
AND Code_Document < 4

-- Documents (articles, notes, review) disponibles dans le WoS entre 2008 et 2013 pour avoir l'adresse de l'auteur avec un DOI

DROP TABLE #corpus
SELECT DISTINCT
      ar2.annee_bibliographique, ar.id_art, SUBSTRING(ar.doi, 5, LEN(ar.doi)) AS DOI
INTO    #corpus
FROM    Pub_Expanded.dbo.Identifiant AS ar
INNER JOIN Pub_Expanded.dbo.Article AS ar2 ON ar2.ID_Art = ar.Id_art
INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.id_art = ar.id_art
INNER JOIN Pub_Expanded.dbo.Adresse AS adr ON adr.id_art = ar.id_art AND adr.Ordre_tr = aa.Ordre_TR
INNER JOIN Pub_Expanded.dbo.Auteur AS au ON au.id_art = ar.id_art AND au.Ordre = aa.Ordre_Auteur
WHERE   Code_Revue BETWEEN 13071 AND 13078 AND Annee_Bibliographique BETWEEN 2008 AND 2013
AND DOI IS NOT NULL AND Code_Document < 4

-- Identification du corpus principal
-- Nombre de documents communs avec le download des contributions
SELECT  COUNT(DISTINCT id_art)
FROM    #corpus AS c
INNER JOIN plos.dbo.PlosArticle AS pa ON pa.doi = c.doi

-- Nombre de documents avec une contribution non null
SELECT  COUNT(DISTINCT c.id_art)
FROM    #corpus AS c
INNER JOIN plos.dbo.PlosArticle AS pa ON pa.doi = c.doi
INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = c.Id_art
WHERE   Contribution IS NOT NULL

-- Création de la table contenant les contributions originales pour l'extraction de la contribution et de l'auteur

DROP TABLE #traitement
SELECT DISTINCT
      c.id_art, c.doi, Contribution
INTO    #traitement
FROM    #corpus AS c
INNER JOIN plos.dbo.PlosArticle AS pa ON pa.doi = c.doi
INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = c.Id_art
WHERE   Code_Document < 4 AND Contribution IS NOT NULL


--- Table contenant les informations sur les auteurs du WoS
DROP TABLE #auteur
SELECT DISTINCT
      ar.id_art, SUBSTRING(i.DOI, 5, LEN(i.doi)) DOI,
```

```
      SUBSTRING(au.nom, CHARINDEX('-', nom) + 1, LEN(nom)) + SUBSTRING(au.nom, 1, 1) AS Initiales, Nom,
Nom_Famille,
      Prenom, Surnom, Ordre
INTO    #Auteur
FROM    Pub_Expanded.dbo.Article AS ar
INNER JOIN Pub_Expanded.dbo.auteur AS au ON au.id_art = ar.id_art
INNER JOIN Pub_Expanded.dbo.Identifiant AS i ON i.Id_art = ar.ID_Art
INNER JOIN #traitement AS t ON t.id_art = ar.id_art
WHERE   ar.code_revue BETWEEN 13071 AND 13078

-- ajout de la deuxième paire d'initiales

ALTER TABLE #Auteur
ADD initiales2 VARCHAR(50);

--Création des initiales des auteurs dans le WoS
BEGIN
   UPDATE  #Auteur
   SET     Initiales2 = SUBSTRING(au.nom, CHARINDEX('-', au.nom) + 1, LEN(au.nom)) + SUBSTRING(au.nom,
1, 1) + SUBSTRING(au.nom_famille,
                                                                  CHARINDEX('-',
                                                                  au.nom_famille) + 1,
                                                                  1)
   FROM    #auteur AS au
   WHERE   au.Nom_Famille LIKE '%-%'
END
BEGIN
   UPDATE  #Auteur
   SET     Initiales2 = UPPER(SUBSTRING(au.nom, CHARINDEX('-', au.nom) + 1, LEN(au.nom)) +
SUBSTRING(au.nom, 1, 1) + SUBSTRING(au.nom_famille,
                                                                  CHARINDEX(' ',
                                                                  au.nom_famille) + 1,
                                                                  1))
   FROM    #auteur AS au
   WHERE   au.Nom_Famille LIKE '% %'
END

-- Supprime les caractères spéciaux (TAB, RETOUR ET PARAGRAPHE) et les balises inutiles
BEGIN
   UPDATE  #traitement
   SET     contribution = RTRIM(LTRIM(REPLACE(REPLACE(REPLACE(UPPER(Contribution), CHAR(10), ' '),
CHAR(13), ''),
                                  CHAR(9), '')))
END

BEGIN
   UPDATE  #traitement
   SET     Contribution = LTRIM(RTRIM(REPLACE(Contribution,
                                  'The author(s) have made the following declarations about their contributions:',
                                  '')))
END

BEGIN
   UPDATE  #traitement
   SET     contribution = REPLACE(REPLACE(REPLACE(REPLACE(contribution, ', and', ','), '          ', ' '),
                                 '<italic>', ''), '</italic>', '')
END
```

```
BEGIN
   UPDATE  #traitement
   SET     contribution = REPLACE(contribution, 'org/1999/xlink">icmje</ext-link>', 'ICMJE')
END

BEGIN
   UPDATE  #traitement
   SET     contribution = REPLACE(contribution, '  ', ' ')
END

BEGIN
   UPDATE  #traitement
   SET     contribution = REPLACE(contribution,
                          '<EXT-LINK EXT-LINK-TYPE="URI" XLINK:HREF="HTTP://WWW.ICMJE.ORG/"
XLINK:TYPE="SIMPLE" XMLNS:XLINK="HTTP://WWW.W3.ICMJE ',
                          '')
   WHERE   contribution LIKE '<EXT%'
END

BEGIN
   UPDATE  #traitement
   SET     contribution = REPLACE(contribution,
                          '(<EXT-LINK EXT-LINK-TYPE="URI" XLINK:HREF="HTTP://CAMERA.CALIT2.NET/"
XLINK:TYPE="SIMPLE"
XMLNS:XLINK="HTTP://WWW.W3.ORG/1999/XLINK">HTTP://CAMERA.CALIT2.NET/</EXT-LINK>)',
                          ' ')
END



--- identification des cas problèmes (traitement utilisant les points et les deux points).

DROP TABLE #nb
SELECT  doi, LEN(contribution) - LEN(REPLACE(contribution, ':', '')) Nb,
       LEN(contribution) - LEN(REPLACE(contribution, '.', '')) Nbdot
INTO    #nb
FROM    #traitement AS t
WHERE   CHARINDEX(' ', Contribution) > 6
ORDER BY LEN(contribution) - LEN(REPLACE(contribution, ':', '')) DESC

-- parsing par contribution -  contenant les auteurs
DROP TABLE #parse

SELECT DISTINCT
       doi, contribution,
       DATALENGTH(LEFT('.' + y.Contribution + '.', N)) - DATALENGTH(REPLACE(LEFT('.' + y.Contribution + '.',
N), '.',
                                                    '')) Ordre,
       LTRIM(SUBSTRING('.' + y.Contribution + '.', N + 1, CHARINDEX('.', '.' + y.Contribution + '.', N + 1) - N - 1))
       AS ContributionAuteur
INTO    #parse
FROM    #traitement AS y ,
       bdben.dbo.nombres AS t
WHERE   t.N < LEN('.' + y.Contribution + '.') AND SUBSTRING('.' + y.Contribution + '.', N, 1) = '.' AND (
CHARINDEX(' ',
```

```
                                                    Contribution) > 6 OR contribution LIKE 'wrote %')
AND doi IN (
      SELECT DISTINCT
            doi
      FROM    #nb
      WHERE   nb = nbdot)
ORDER BY doi, ordre

-- Ajout de la colonne ContributionClean
ALTER TABLE #parse
ADD ContributionClean VARCHAR(750)

BEGIN
   UPDATE  #parse
   SET     ContributionClean = REPLACE(SUBSTRING(ContributionAuteur, 1, CHARINDEX(':',
ContributionAuteur)), ':', '')
   FROM    #parse
END

-- Ajout de la colonne auteur
ALTER TABLE #parse
ADD Auteur VARCHAR(2000)

BEGIN
   UPDATE  #parse
   SET     Auteur = LTRIM(REPLACE(SUBSTRING(ContributionAuteur, CHARINDEX(':', ContributionAuteur),
                        LEN(ContributionAuteur)), ':', ''))
   FROM    #parse
END

BEGIN
   UPDATE  #parse
   SET     Auteur = REPLACE(auteur, '-', '')
   FROM    #parse
END

-- Parsing des auteurs
DROP TABLE #parseAuteur
SELECT DISTINCT
      doi, contribution, REPLACE(REPLACE(contributionclean, '  ', ' '), ' ', ' ') ContributionClean,
      DATALENGTH(LEFT(' ' + y.Auteur + ' ', N)) - DATALENGTH(REPLACE(LEFT(' ' + y.Auteur + ' ', N), ' ', ''))
Ordre,
      LTRIM(SUBSTRING(' ' + y.Auteur + ' ', N + 1, CHARINDEX(' ', ' ' + y.Auteur + ' ', N + 1) - N - 1)) AS
AuteurClean
INTO    #parseAuteur
FROM    #parse AS y ,
      bdben.dbo.nombres AS t
WHERE   t.N < LEN(' ' + y.Auteur + ' ') AND SUBSTRING(' ' + y.Auteur + ' ', N, 1) = ' '
ORDER BY doi, ContributionClean, Ordre

-- tableau 3
SELECT  SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean))),
      COUNT(auteurclean) f
FROM    #parseauteur
GROUP BY SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean)))
```

```
ORDER BY f DESC

SELECT  SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean))),
      COUNT(auteurclean) f, COUNT(DISTINCT doi)
FROM    #parseauteur
GROUP BY SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean)))
ORDER BY f DESC

SELECT  COUNT(DISTINCT doi)
FROM    #parseauteur
WHERE   SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean))) NOT IN (
      SELECT DISTINCT TOP 10
            SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean)))
      FROM    #parseAuteur AS pa
      GROUP BY SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean)))
      HAVING  COUNT(DISTINCT doi) > 900)

SELECT  SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean))),
      COUNT(auteurclean) f, COUNT(DISTINCT doi)
FROM    #parseauteur
GROUP BY SUBSTRING(contributionclean, 1, 1) + LOWER(SUBSTRING(contributionclean, 2,
LEN(contributionclean)))
ORDER BY f DESC


--- Traitement résiduel avec 7 caractères et plus

DROP TABLE #residuel
SELECT DISTINCT
      t.doi, t.Contribution, CHARINDEX(' ', t.Contribution) long,
      SUBSTRING(t.Contribution, 1, CHARINDEX(' ', t.Contribution)) sub
INTO    #residuel
FROM    #traitement AS t
LEFT JOIN #parse AS pa ON pa.doi = t.doi
WHERE   pa.doi IS NULL AND CHARINDEX(' ', t.Contribution) > 6 OR t.Contribution LIKE 'WROTE %' OR
t.Contribution LIKE 'STUDY %' OR t.Contribution LIKE 'FINAL %' OR t.Contribution LIKE 'FIRST %' OR
t.Contribution LIKE 'ICMJE %' OR t.Contribution LIKE 'MODEL %' OR t.Contribution LIKE 'THIS %' OR
t.Contribution LIKE 'THE %' OR t.Contribution LIKE 'TOOK %' OR t.Contribution LIKE 'IDEA %' OR t.Contribution
LIKE 'FOR %' OR t.Contribution LIKE 'GAVE %' OR t.Contribution LIKE 'DATA %' OR t.Contribution LIKE
'BUILT %'


--SELECT distinct* FROM #residuel AS r

DROP TABLE #parserésiduel
SELECT DISTINCT
      doi, contribution,
      DATALENGTH(LEFT('.' + y.Contribution + '.', N)) - DATALENGTH(REPLACE(LEFT('.' + y.Contribution + '.',
N), '.',
                                                "")) Ordre,
      LTRIM(SUBSTRING('.' + y.Contribution + '.', N + 1, CHARINDEX('.', '.' + y.Contribution + '.', N + 1) - N - 1))
```

```
        AS ContributionAuteur
INTO    #parserésiduel
FROM    #residuel AS y ,
        bdben.dbo.nombres AS t
WHERE   t.N < LEN('.' + y.Contribution + '.') AND SUBSTRING('.' + y.Contribution + '.', N, 1) = '.'
ORDER BY doi, ordre

--- identification des libellés
-- ajout de la colonnne contributionclean
ALTER TABLE #parseresiduel
ADD ContributionClean VARCHAR(750)

BEGIN
   UPDATE  #parseresiduel
   SET     ContributionClean = REPLACE(SUBSTRING(ContributionAuteur, 1, CHARINDEX(':',
ContributionAuteur)), ':', '')
   FROM    #parseresiduel
END


-- ajouter un marqueur aux libellés
-- Déclarer les variables FETCH.
DECLARE @contribution VARCHAR(500) ,
   @nvcontribution VARCHAR(500) ,
   @doi VARCHAR(50) ,
   @SQL NVARCHAR(MAX)
DECLARE requete_cursor CURSOR
FOR
SELECT DISTINCT
        contributionclean, '|' + contributionclean AS nvContribution, doi
FROM    #parseresiduel
WHERE   contributionclean != ''
OPEN requete_cursor;

FETCH NEXT FROM requete_cursor
INTO @contribution, @nvcontribution, @doi
WHILE @@FETCH_STATUS = 0
   BEGIN

-- Peupler la requête SQL
        SET @SQL = 'BEGIN UPDATE #residuel SET Contribution = REPLACE(contribution, ''' + @contribution + ''','''
+ @nvcontribution + ''')
                                WHERE doi= ''' + @doi + '''END '

        EXECUTE(@SQL)

--PRINT(@SQL)

        FETCH NEXT FROM requete_cursor
INTO @contribution, @nvcontribution, @doi
   END

CLOSE requete_cursor;
DEALLOCATE requete_cursor;
GO

--- on refait le parsing avec les nouveaux marqueurs (suppression de la table qui nous a permis de les identifier)
```

```
DROP TABLE #parserésiduel
SELECT DISTINCT
      doi, contribution,
      DATALENGTH(LEFT('|' + y.Contribution + '|', N)) - DATALENGTH(REPLACE(LEFT('|' + y.Contribution + '|',
N), '|',
                                                      '')) Ordre,
      LTRIM(SUBSTRING('|' + y.Contribution + '|', N + 1, CHARINDEX('|', '|' + y.Contribution + '|', N + 1) - N - 1))
      AS ContributionAuteur
INTO   #parserésiduel
FROM   #residuel AS y ,
      bdben.dbo.nombres AS t
WHERE  t.N < LEN('|' + y.Contribution + '|') AND SUBSTRING('|' + y.Contribution + '|', N, 1) = '|'
ORDER BY doi, ordre


-- ajout de la colonnne contributionclean
ALTER TABLE #parseresiduel
ADD ContributionClean VARCHAR(750)

BEGIN
   UPDATE  #parseresiduel
   SET     ContributionClean = REPLACE(SUBSTRING(ContributionAuteur, 1, CHARINDEX(':',
ContributionAuteur)), ':', '')
   FROM    #parseresiduel
END

ALTER TABLE  #parseresiduel
ADD Auteur VARCHAR(2000)

BEGIN
   UPDATE  #parseresiduel
   SET     Auteur = LTRIM(REPLACE(SUBSTRING(ContributionAuteur, CHARINDEX(':', ContributionAuteur),
                                LEN(ContributionAuteur)), ':', ''))
   FROM    #parseresiduel
END

BEGIN
   UPDATE  #parseresiduel
   SET     Auteur = REPLACE(auteur, '-', '')
   FROM    #parseresiduel
END

BEGIN
   UPDATE  #parseresiduel
   SET     Auteur = REPLACE(auteur, '.', '')
   FROM    #parseresiduel
END


DROP TABLE #parseresiduelauteur

SELECT DISTINCT
      doi, contribution, ordre, contributionclean,
      DATALENGTH(LEFT(' ' + y.Auteur + ' ', N)) - DATALENGTH(REPLACE(LEFT(' ' + y.Auteur + ' ', N), ' ', ''))
Ordreauteur,
```

```
      LTRIM(SUBSTRING(' ' + y.Auteur + ' ', N + 1, CHARINDEX(' ', ' ' + y.Auteur + ' ', N + 1) - N - 1)) AS
AuteurClean
INTO    #parserésiduelAuteur
FROM    #parseresiduel AS y ,
      bdben.dbo.nombres AS t
WHERE   t.N < LEN(' ' + y.Auteur + ' ') AND SUBSTRING(' ' + y.Auteur + ' ', N, 1) = ' ' AND y.auteur NOT LIKE
'%,%'
ORDER BY doi, ordre


INSERT  INTO #parserésiduelAuteur
      SELECT DISTINCT
            doi, contribution, ordre, contributionclean,
            DATALENGTH(LEFT(',' + y.Auteur + ',', N)) - DATALENGTH(REPLACE(LEFT(',' + y.Auteur + ',', N), ',',
'')) Ordreauteur,
            LTRIM(SUBSTRING(',' + y.Auteur + ',', N + 1, CHARINDEX(',', ',' + y.Auteur + ',', N + 1) - N - 1)) AS
AuteurClean
      FROM    #parseresiduel AS y ,
            bdben.dbo.nombres AS t
      WHERE   t.N < LEN(',' + y.Auteur + ',') AND SUBSTRING(',' + y.Auteur + ',', N, 1) = ',' AND y.auteur LIKE
'%,%'
      ORDER BY doi, ordre

-- Résiduel 2
DROP TABLE #Résiduel2
SELECT DISTINCT
      doi, contribution
INTO    #Résiduel2
FROM    #traitement AS a
WHERE   doi NOT IN ( SELECT DISTINCT
                        doi
                  FROM   #parse AS c
                  UNION ALL
                  SELECT DISTINCT
                        doi
                  FROM   #parseresiduel)

UPDATE  #residuel2
SET     contribution = REPLACE(contribution, '-', '')


DROP TABLE #parserésiduel2
SELECT DISTINCT
      doi, contribution,
      DATALENGTH(LEFT('.' + y.Contribution + '.', N)) - DATALENGTH(REPLACE(LEFT('.' + y.Contribution + '.',
N), '.',
                                                      '')) Ordre,
      LTRIM(SUBSTRING('.' + y.Contribution + '.', N + 1, CHARINDEX('.', '.' + y.Contribution + '.', N + 1) - N - 1))
      AS ContributionAuteur
INTO    #parserésiduel2
FROM    #Résiduel2 AS y ,
      bdben.dbo.nombres AS t
WHERE   t.N < LEN('.' + y.Contribution + '.') AND SUBSTRING('.' + y.Contribution + '.', N, 1) = '.'
ORDER BY doi, ordre


-- ajout de la colonnne contributionclean
```

```
ALTER TABLE #parseresiduel2
ADD ContributionClean VARCHAR(750)

--SELECT DISTINCT contributionauteur FROM  #parseresiduel2

BEGIN
   UPDATE  #parseresiduel2
   SET     ContributionClean = 'PERFORMED THE EXPERIMENTS'
   FROM    #parseresiduel2
   WHERE   contributionauteur LIKE '%PERFORMED THE EXPERIMENTS'
END

BEGIN
   UPDATE  #parseresiduel2
   SET     ContributionClean = 'ANALYZED THE DATA'
   FROM    #parseresiduel2
   WHERE   contributionauteur LIKE '%ANALYZED THE DATA'
END


BEGIN
   UPDATE  #parseresiduel2
   SET     ContributionClean = 'CONCEIVED AND DESIGNED THE EXPERIMENTS'
   FROM    #parseresiduel2
   WHERE   contributionauteur LIKE '%CONCEIVED AND DESIGNED%EXPERIMENTS'
END

BEGIN
   UPDATE  #parseresiduel2
   SET     ContributionClean = 'CONTRIBUTED REAGENTS/MATERIALS/ANALYSIS TOOLS'
   FROM    #parseresiduel2
   WHERE   contributionauteur LIKE '%CONTRIBUTED REAGENTS%'
END

BEGIN
   UPDATE  #parseresiduel2
   SET     ContributionClean = 'WROTE THE PAPER'
   FROM    #parseresiduel2
   WHERE   contributionauteur LIKE '%WROTE THE PAPER%'
END

BEGIN
   UPDATE  #parseresiduel2
   SET     ContributionClean = 'WROTE THE PAPER'
   FROM    #parseresiduel2
   WHERE   contributionauteur LIKE '%WROTE%manus%'
END

-- Inconnu
SELECT DISTINCT
       DOI
FROM    #parserésiduel2 AS p
WHERE   Contributionclean IS NULL AND ContributionAuteur != ''

ALTER TABLE  #parseresiduel2
ADD Auteur VARCHAR(2000)
```

```
BEGIN
   UPDATE  #parseresiduel2
   SET     Auteur = contributionauteur
   FROM    #parseresiduel2
END


BEGIN
   UPDATE  #parseresiduel2
   SET     Auteur = REPLACE(auteur, '-', ' ')
   FROM    #parseresiduel2
END

BEGIN
   UPDATE  #parseresiduel2
   SET     Auteur = REPLACE(auteur, ',', ' ')
   FROM    #parseresiduel2
END


DROP TABLE #parseresiduelauteur2

SELECT DISTINCT
       doi, contribution, ordre, contributionclean,
       DATALENGTH(LEFT(' ' + y.Auteur + ' ', N)) - DATALENGTH(REPLACE(LEFT(' ' + y.Auteur + ' ', N), ' ', ''))
Ordreauteur,
       LTRIM(SUBSTRING(' ' + y.Auteur + ' ', N + 1, CHARINDEX(' ', ' ' + y.Auteur + ' ', N + 1) - N - 1)) AS
AuteurClean
INTO    #parserésiduelAuteur2
FROM    #parseresiduel2 AS y ,
       bdben.dbo.nombres AS t
WHERE   t.N < LEN(' ' + y.Auteur + ' ') AND SUBSTRING(' ' + y.Auteur + ' ', N, 1) = ' '
ORDER BY doi, ordre



-- Union des trois tables contenant les contributions associées aux auteurs
DROP TABLE #tempContribution
SELECT DISTINCT
       'parse1' AS Parse, doi, contribution, ordre, contributionclean, auteurclean
INTO    #tempContribution
FROM    #parseauteur
UNION ALL
SELECT DISTINCT
       'Parse2', doi, contribution, ordre, contributionclean, auteurclean
FROM    #parserésiduelAuteur AS pa
UNION ALL
SELECT DISTINCT
       'parse3', doi, contribution, ordre, contributionclean, auteurclean
FROM    #parserésiduelAuteur2 AS pa

DROP TABLE #ContributionAuteur
SELECT DISTINCT
       *, ROW_NUMBER() OVER ( ORDER BY Doi, ordre) idrow
INTO    #ContributionAuteur
FROM    #tempContribution AS tc
```

```
ORDER BY doi, ordre, idrow

SELECT  contribution, COUNT(nom), COUNT(DISTINCT doi)
FROM    Contributionauteur
GROUP BY contribution

UPDATE  #ContributionAuteur
SET     auteurclean = RTRIM(LTRIM(REPLACE(auteurclean, ' ', ' ')))

-- suppression des initiales < 2
DELETE  #contributionauteur
WHERE   LEN(auteurclean) < 2
```

Annexe 6. Création de la table de concordance sur la base des initiales

```
-- Création de la table de concordance sur la base des initiales

DROP TABLE #corpusauteur
SELECT DISTINCT
      0 AS step, a.doi, p.idrow, ar.nb_auteur AS NbAuteurWoS,
      p.contributionclean, UPPER(p.Auteurclean) InitialesPlos, a.ID_Art,
      a.Initiales, a.Nom, a.Nom_Famille, a.Prenom, a.Surnom, a.Ordre
INTO    #corpusauteur
FROM    #contributionauteur AS p
INNER JOIN #Auteur AS a ON a.doi = p.doi
                    AND ( a.Initiales = REPLACE(p.auteurclean, ',', '')
                        OR a.initiales2 = REPLACE(p.auteurclean, ',',
                                                  '') )
INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = a.ID_Art
INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.ID_Art = ar.ID_Art

--- ajout
INSERT  INTO #corpusauteur
      SELECT DISTINCT
            1, a.doi, p.idrow, ar.nb_auteur AS NbAuteurWoS,
            p.contributionclean, UPPER(p.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #contributionauteur AS p
      INNER JOIN #Auteur AS a ON a.doi = p.doi
                        AND REPLACE(p.auteurclean, ',', '') = SUBSTRING(prenom,
                                                                        1, 1)
                        + SUBSTRING(nom_famille, 1, 1)
      INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = a.ID_Art
      INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.ID_Art = ar.ID_Art
      LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                          AND c.ordre = a.ordre
      WHERE   c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND p.idrow NOT IN ( SELECT DISTINCT
                                 idrow
                        FROM    #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
      SELECT DISTINCT
            26, a.doi, p.idrow, ar.nb_auteur AS NbAuteurWoS,
            p.contributionclean, UPPER(p.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #contributionauteur AS p
      INNER JOIN #Auteur AS a ON a.doi = p.doi
                        AND REPLACE(p.auteurclean, ',', '') = prenom
                        + ' ' + nom_famille
      INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = a.ID_Art
      INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.ID_Art = ar.ID_Art
      LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                          AND c.ordre = a.ordre
```

```
    WHERE   c.id_art IS NULL
        AND c.ordre IS  NULL
        AND p.idrow NOT IN ( SELECT DISTINCT
                                idrow
                             FROM   #corpusauteur )



INSERT  INTO #corpusauteur
    SELECT DISTINCT
        2, a.doi, p.idrow, ar.nb_auteur AS NbAuteurWoS,
        p.contributionclean, UPPER(p.Auteurclean) InitialesPlos,
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #contributionauteur AS p
    INNER JOIN #Auteur AS a ON a.doi = p.doi
                    AND REPLACE(p.auteurclean, ',', '') = SUBSTRING(nom,
                                                            CHARINDEX('-',
                                                            nom) + 1, 1)
                    + SUBSTRING(nom, 1, 1)
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = a.ID_Art
    INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.ID_Art = ar.ID_Art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                        AND c.ordre = a.ordre
    WHERE   c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND p.idrow NOT IN ( SELECT DISTINCT
                                idrow
                             FROM   #corpusauteur )

-- ajout
INSERT  INTO #corpusauteur
    SELECT DISTINCT
        3, a.doi, p.idrow, ar.nb_auteur AS NbAuteurWoS,
        p.contributionclean, UPPER(p.Auteurclean) InitialesPlos,
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #contributionauteur AS p
    INNER JOIN #Auteur AS a ON a.doi = p.doi
                    AND LTRIM(REVERSE(SUBSTRING(REVERSE(p.auteurclean),
                                                1,
                                                CHARINDEX(' ',
                                                REVERSE(p.auteurclean))))) = a.nom_famille
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = a.ID_Art
    INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.ID_Art = ar.ID_Art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                        AND c.ordre = a.ordre
    WHERE   p.auteurclean LIKE '% %'
        AND c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND p.idrow NOT IN ( SELECT DISTINCT
                                idrow
                             FROM   #corpusauteur )



INSERT  INTO #corpusauteur
```

```
        SELECT DISTINCT
                4, a.doi, p.idrow, ar.nb_auteur AS NbAuteurWoS,
                p.contributionclean, UPPER(p.Auteurclean) InitialesPlos,
                a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
                a.Surnom, a.Ordre
        FROM    #contributionauteur AS p
        INNER JOIN #Auteur AS a ON a.doi = p.doi
                            AND SUBSTRING(auteurclean, 1, 1)
                            + SUBSTRING(auteurclean,
                                    CHARINDEX(' ', auteurclean) + 1,
                                    1) = SUBSTRING(a.prenom, 1, 1)
                            + SUBSTRING(a.nom_famille, 1, 1)
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = a.ID_Art
        INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.ID_Art = ar.ID_Art
        LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                                AND c.ordre = a.ordre
        WHERE   p.auteurclean LIKE '% %'
                AND p.doi = '10.1371/journal.pgen.1002584'
                AND c.id_art IS  NULL
                AND c.ordre IS  NULL
                AND p.idrow NOT IN ( SELECT DISTINCT
                                        idrow
                                FROM   #corpusauteur )


--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
                5, a.doi, p.idrow, ar.nb_auteur AS NbAuteurWoS,
                p.contributionclean, UPPER(p.Auteurclean) InitialesPlos,
                a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
                a.Surnom, a.Ordre
        FROM    #contributionauteur AS p
        INNER JOIN #Auteur AS a ON a.doi = p.doi
                            AND ( a.Initiales = REPLACE(p.auteurclean,
                                                ' ', '')
                                OR a.initiales2 = REPLACE(p.auteurclean,
                                                ' ', '') )
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.ID_Art = a.ID_Art
        INNER JOIN Pub_Expanded.dbo.Adresse_Auteur AS aa ON aa.ID_Art = ar.ID_Art
        LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                                AND c.ordre = a.ordre
        WHERE   c.id_art IS  NULL
                AND c.ordre IS  NULL
                AND p.idrow NOT IN ( SELECT DISTINCT
                                        idrow
                                FROM   #corpusauteur )



--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
                6, p.doi, a.idrow, ar.nb_auteur AS NbAuteurWoS,
                a.contributionclean, UPPER(a.Auteurclean) InitialesPlos,
                p.ID_Art, p.Initiales, p.Nom, p.Nom_Famille, p.Prenom,
                p.Surnom, p.Ordre
```

```
        FROM    #contributionauteur AS a
        INNER JOIN #auteur AS p ON p.doi = a.doi
                            AND ( a.auteurclean = p.nom_famille
                                OR REPLACE(SUBSTRING(a.AuteurClean,
                                                     CHARINDEX('.',
                                                     a.AuteurClean),
                                                     LEN(a.auteurclean)),
                                           '.', '') = REPLACE(SUBSTRING(p.nom,
                                                     1,
                                                     CHARINDEX('-',
                                                     p.nom)), '-', '')
                                OR p.nom_famille = REPLACE(a.auteurclean,
                                                     ',', '')
                                OR LTRIM(SUBSTRING(auteurclean,
                                                   CHARINDEX(' ',
                                                     auteurclean) + 1,
                                                   LEN(auteurclean)))
                                + '-' + SUBSTRING(auteurclean, 1, 1) = SUBSTRING(p.nom,
                                                     1,
                                                     CHARINDEX('-',
                                                     p.nom) + 1) )
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = p.id_art
        LEFT JOIN #corpusauteur AS c ON c.id_art = p.id_art
                                AND c.ordre = p.ordre
        WHERE   c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND a.idrow NOT IN ( SELECT DISTINCT
                                 idrow
                         FROM   #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
            7, p.doi, a.idrow, ar.nb_auteur AS NbAuteurWoS,
            a.contributionclean, UPPER(a.Auteurclean) InitialesPlos,
            p.ID_Art, p.Initiales, p.Nom, p.Nom_Famille, p.Prenom,
            p.Surnom, p.Ordre
        FROM    #contributionauteur AS a
        INNER JOIN #auteur AS p ON p.doi = a.doi
                            AND ( SUBSTRING(a.auteurclean, 1, 1)
                                + SUBSTRING(REVERSE(a.auteurclean), 1,
                                            1) = p.initiales )
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = p.id_art
        LEFT JOIN #corpusauteur AS c ON c.id_art = p.id_art
                                AND c.ordre = p.ordre
        WHERE   c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND LEN(a.auteurclean) = 3
            AND a.idrow NOT IN ( SELECT DISTINCT
                                 idrow
                         FROM   #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
            8, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
            ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
```

```
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #auteur AS a
    INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                    AND c.ordre = a.ordre
    WHERE   LEN(a.nom_famille) - LEN(REPLACE(a.nom_famille, ' ', '')) = 2
        AND SUBSTRING(a.nom, CHARINDEX('-', a.nom) + 1, LEN(a.nom))
        + SUBSTRING(a.nom_famille, 1, 1) + SUBSTRING(a.nom_famille,
                                    CHARINDEX(' ',
                                     a.nom_famille)
                                    + 1, 1)
        + SUBSTRING(SUBSTRING(a.nom_famille,
                    CHARINDEX(' ', a.nom_famille) + 1,
                    LEN(a.nom_famille)),
                CHARINDEX(' ',
                    SUBSTRING(a.nom_famille,
                        CHARINDEX(' ', a.nom_famille)
                        + 1, LEN(a.nom_famille))) + 1,
                1) = ca.auteurclean
        AND c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND ca.idrow NOT IN ( SELECT DISTINCT
                    idrow
                FROM      #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
    SELECT DISTINCT
        9, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
        ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #auteur AS a
    INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                    AND c.ordre = a.ordre
    WHERE   LEN(a.nom_famille) - LEN(REPLACE(a.nom_famille, ' ', '')) = 2
        AND SUBSTRING(a.prenom, CHARINDEX(' ', a.prenom) + 1, 1)
        + SUBSTRING(SUBSTRING(a.prenom, CHARINDEX(' ', a.prenom) + 1,
                    LEN(a.prenom)),
                CHARINDEX(' ',
                    SUBSTRING(a.prenom,
                        CHARINDEX(' ', a.prenom) + 1,
                        LEN(a.prenom))), 1)
        + SUBSTRING(a.nom_famille, 1, 1) + SUBSTRING(a.nom_famille,
                                    CHARINDEX(' ',
                                     a.nom_famille)
                                    + 1, 1)
        + SUBSTRING(SUBSTRING(a.nom_famille,
                    CHARINDEX(' ', a.nom_famille) + 1,
                    LEN(a.nom_famille)),
                CHARINDEX(' ',
                    SUBSTRING(a.nom_famille,
                        CHARINDEX(' ', a.nom_famille)
```

```
                                + 1, LEN(a.nom_famille))) + 1,
                    1) = ca.auteurclean
            AND c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND ca.idrow NOT IN ( SELECT DISTINCT
                                  idrow
                          FROM    #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
                10, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
                ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
                a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
                a.Surnom, a.Ordre
        FROM    #auteur AS a
        INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
        LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                                AND c.ordre = a.ordre
        WHERE   LEN(a.nom_famille) - LEN(REPLACE(a.nom_famille, ' ', '')) = 2
                AND SUBSTRING(a.prenom, 1, 1) + SUBSTRING(SUBSTRING(a.prenom,
                                          1, LEN(a.prenom)),
                                        CHARINDEX(' ',
                                          SUBSTRING(a.prenom,
                                          1, LEN(a.prenom)))
                                        + 1, 1)
                + SUBSTRING(a.nom_famille, 1, 1) + SUBSTRING(a.nom_famille,
                                          CHARINDEX(' ',
                                          a.nom_famille)
                                          + 1, 1)
                + SUBSTRING(SUBSTRING(a.nom_famille,
                            CHARINDEX(' ', a.nom_famille) + 1,
                            LEN(a.nom_famille)),
                        CHARINDEX(' ',
                            SUBSTRING(a.nom_famille,
                                CHARINDEX(' ', a.nom_famille)
                                + 1, LEN(a.nom_famille))) + 1,
                        1) = ca.auteurclean
                AND c.id_art IS  NULL
                AND c.ordre IS  NULL
                AND a.prenom IS NOT NULL
                AND ca.idrow NOT IN ( SELECT DISTINCT
                                      idrow
                              FROM    #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
                11, a.doi, b.idrow, ar.nb_auteur AS NbAuteurWoS,
                b.contributionclean, UPPER(b.Auteurclean) InitialesPlos,
                a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
                a.Surnom, a.Ordre
        FROM    #auteur AS a
        INNER JOIN #contributionauteur AS b ON a.doi = b.doi
                                AND LTRIM(REPLACE(REPLACE(SUBSTRING(a.nom_famille,
                                          CHARINDEX(' ',
```

```
                                        a.nom_famille),
                                        LEN(a.nom_famille)),
                                        '-', ''), "",
                                        '')) = b.auteurclean
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
        LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                            AND c.ordre = a.ordre
        WHERE   LEN(a.nom_famille) - LEN(REPLACE(a.nom_famille, ' ', '')) = 1
            AND c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND b.idrow NOT IN ( SELECT DISTINCT
                                 idrow
                         FROM   #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
            12, a.doi, b.idrow, ar.nb_auteur AS NbAuteurWoS,
            b.contributionclean, UPPER(b.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
        FROM    #auteur AS a
        INNER JOIN #contributionauteur AS b ON a.doi = b.doi
                                    AND SUBSTRING(a.prenom, 1, 1)
                                    + SUBSTRING(a.prenom,
                                            CHARINDEX(' ',
                                              a.prenom) + 1, 1)
                                    + LTRIM(REPLACE(REPLACE(SUBSTRING(a.nom_famille,
                                              CHARINDEX(' ',
                                              a.nom_famille)
                                              + 1, 1), '-', ''),
                                              "", '')) = b.auteurclean
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
        LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                            AND c.ordre = a.ordre
        WHERE   c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND b.idrow NOT IN ( SELECT DISTINCT
                                 idrow
                         FROM   #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
        SELECT DISTINCT
            13, a.doi, b.idrow, ar.nb_auteur AS NbAuteurWoS,
            b.contributionclean, UPPER(b.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
        FROM    #auteur AS a
        INNER JOIN #contributionauteur AS b ON a.doi = b.doi
                                    AND a.initiales
                                    + SUBSTRING(a.surnom, 1, 1) = b.auteurclean
        INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
        LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                            AND c.ordre = a.ordre
        WHERE   c.id_art IS  NULL
            AND c.ordre IS  NULL
```

```
            AND a.surnom IS NOT NULL
            AND b.idrow NOT IN ( SELECT DISTINCT
                                 idrow
                        FROM   #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
      SELECT DISTINCT
            14, a.doi, b.idrow, ar.nb_auteur AS NbAuteurWoS,
            b.contributionclean, UPPER(b.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #auteur AS a
      INNER JOIN #contributionauteur AS b ON a.doi = b.doi /*AND SUBSTRING(a.initiales,1,1)+
                  SUBSTRING(REVERSE(a.initiales),1,1) =
SUBSTRING(b.auteurclean,1,1)+SUBSTRING(REVERSE(b.auteurclean),1,1)*/
      INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
      LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                              AND c.ordre = a.ordre
      WHERE   c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND SUBSTRING(a.initiales, 1, 1)
            + SUBSTRING(REVERSE(a.initiales), 1, 1) = SUBSTRING(b.auteurclean,
                                          1, 1)
            + SUBSTRING(REVERSE(b.auteurclean), 1, 1)
            AND LEN(a.prenom) - LEN(REPLACE(a.prenom, ' ', '')) > 0
            AND LEN(b.auteurclean) < 5
            AND b.idrow NOT IN ( SELECT DISTINCT
                                 idrow
                        FROM   #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
      SELECT DISTINCT
            15, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
            ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #auteur AS a
      INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
      INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
      LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                              AND c.ordre = a.ordre
      WHERE   LEN(a.nom_famille) - LEN(REPLACE(a.nom_famille, '-', '')) > 0
            AND SUBSTRING(a.prenom, CHARINDEX(' ', a.prenom) + 1, 1)
            + SUBSTRING(SUBSTRING(a.prenom, CHARINDEX(' ', a.prenom) + 1,
                        LEN(a.prenom)),
                  CHARINDEX(' ',
                        SUBSTRING(a.prenom,
                              CHARINDEX(' ', a.prenom) + 1,
                              LEN(a.prenom))), 1)
            + SUBSTRING(a.nom_famille, 1, 1) + SUBSTRING(a.nom_famille,
                                    CHARINDEX(' ',
                                     a.nom_famille)
                                    + 1, 1)
            + SUBSTRING(SUBSTRING(a.nom_famille,
                        CHARINDEX('-', a.nom_famille) + 1,
```

```
                    LEN(a.nom_famille)),
                CHARINDEX('-',
                    SUBSTRING(a.nom_famille,
                        CHARINDEX('-', a.nom_famille)
                        + 1, LEN(a.nom_famille))) + 1,
                1) = ca.auteurclean
        AND c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND ca.idrow NOT IN ( SELECT DISTINCT
                        idrow
                FROM    #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
    SELECT DISTINCT
        16, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
        ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #auteur AS a
    INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                    AND c.ordre = a.ordre
    WHERE   LEN(a.nom_famille) - LEN(REPLACE(a.nom_famille, '-', '')) > 0
        AND SUBSTRING(a.nom, CHARINDEX('-', a.nom) + 1, LEN(a.nom))
        + SUBSTRING(a.nom_famille, 1, 1) + SUBSTRING(a.nom_famille,
                                CHARINDEX('-',
                                 a.nom_famille)
                                + 1, 1)
        + SUBSTRING(SUBSTRING(a.nom_famille,
                    CHARINDEX('-', a.nom_famille) + 1,
                    LEN(a.nom_famille)),
                CHARINDEX('-',
                    SUBSTRING(a.nom_famille,
                        CHARINDEX('-', a.nom_famille)
                        + 1, LEN(a.nom_famille))) + 1,
                1) = ca.auteurclean
        AND c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND ca.idrow NOT IN ( SELECT DISTINCT
                        idrow
                FROM    #corpusauteur )

--- ajout
INSERT  INTO #corpusauteur
    SELECT DISTINCT
        17, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
        ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #auteur AS a
    INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                    AND c.ordre = a.ordre
    WHERE   SUBSTRING(a.nom, 1, 1) + SUBSTRING(a.nom,
```

```
                                    CHARINDEX('-', a.nom) + 1,
                                    LEN(a.nom)) = ca.auteurclean
            AND c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND ca.idrow NOT IN ( SELECT DISTINCT
                                  idrow
                          FROM    #corpusauteur )
-- ajout

INSERT  INTO #corpusauteur
      SELECT DISTINCT
            18, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
            ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #auteur AS a
      INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
      INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
      LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                              AND c.ordre = a.ordre
      WHERE   SUBSTRING(a.initiales, 2, LEN(a.initiales)) = ca.auteurclean
            AND c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND a.prenom LIKE '[A-Z].%'
            AND ca.idrow NOT IN ( SELECT DISTINCT
                                  idrow
                          FROM    #corpusauteur )
 -- ajout

INSERT  INTO #corpusauteur
      SELECT DISTINCT
            19, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
            ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #auteur AS a
      INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
      INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
      LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                              AND c.ordre = a.ordre
      WHERE   SUBSTRING(REVERSE(a.initiales), 1, 3) = SUBSTRING(REVERSE(ca.auteurclean),
                                          1, 3)
            AND c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND LEN(ca.auteurclean) > 3
            AND ca.idrow NOT IN ( SELECT DISTINCT
                                  idrow
                          FROM    #corpusauteur )
 -- ajout

INSERT  INTO #corpusauteur
      SELECT DISTINCT
            20, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
            ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #auteur AS a
```

```
    INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                        AND c.ordre = a.ordre
    WHERE   SUBSTRING(a.initiales, 1, 3) = SUBSTRING(ca.auteurclean, 1, 3)
        AND c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND LEN(ca.auteurclean) > 3
        AND ca.idrow NOT IN ( SELECT DISTINCT
                              idrow
                    FROM      #corpusauteur )

 -- ajout

INSERT  INTO #corpusauteur
    SELECT DISTINCT
        21, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
        ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #auteur AS a
    INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                        AND c.ordre = a.ordre
    WHERE   SUBSTRING(a.initiales, 1, 2) = SUBSTRING(ca.auteurclean, 1, 2)
        AND LEN(ca.auteurclean) > 3
        AND initialesplos NOT LIKE '% %'
        AND c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND ca.idrow NOT IN ( SELECT DISTINCT
                              idrow
                    FROM      #corpusauteur )

 -- ajout

INSERT  INTO #corpusauteur
    SELECT DISTINCT
        22, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
        ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
        a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
        a.Surnom, a.Ordre
    FROM    #auteur AS a
    INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
    LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                        AND c.ordre = a.ordre
    WHERE   SUBSTRING(a.prenom, 1, 1) + REPLACE(SUBSTRING(a.nom, 1,
                                    CHARINDEX('-',
                                    a.nom)), '-', '') = ca.auteurclean
        AND c.id_art IS  NULL
        AND c.ordre IS  NULL
        AND LEN(ca.auteurclean) > 3
        AND ca.idrow NOT IN ( SELECT DISTINCT
                              idrow
                    FROM      #corpusauteur )
```

```
 -- ajout

INSERT  INTO #corpusauteur
      SELECT DISTINCT
            23, ca.doi, ca.idrow, ar.nb_auteur AS NbAuteurWoS,
            ca.contributionclean, UPPER(ca.Auteurclean) InitialesPlos,
            a.ID_Art, a.Initiales, a.Nom, a.Nom_Famille, a.Prenom,
            a.Surnom, a.Ordre
      FROM    #auteur AS a
      INNER JOIN #ContributionAuteur AS ca ON a.doi = ca.doi
      INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = a.id_art
      LEFT JOIN #corpusauteur AS c ON c.id_art = a.id_art
                              AND c.ordre = a.ordre
      WHERE   REPLACE(SUBSTRING(a.nom_famille, 1,
                        CHARINDEX('-', a.nom_famille)), '-', '') = ca.auteurclean
            AND c.id_art IS  NULL
            AND c.ordre IS  NULL
            AND LEN(ca.auteurclean) > 3
            AND ca.idrow NOT IN ( SELECT DISTINCT
                                    idrow
                                  FROM      #corpusauteur )


-- ajout

--un seul qui manque
DROP TABLE #tempauteur
SELECT DISTINCT
      a.*
INTO    #tempauteur
FROM    #auteur AS a
LEFT JOIN #corpusauteur AS b ON a.doi = b.doi
                        AND a.ordre = b.ordre
WHERE   a.doi IN ( SELECT DISTINCT
                        doi
               FROM     #corpusauteur
               GROUP BY doi, nbauteurwos
               HAVING   nbauteurwos - COUNT(DISTINCT ordre) = 1 )
      AND b.ordre IS NULL

--un seul qui reste
DROP TABLE #tempcon
SELECT DISTINCT
      *
INTO    #tempcon
FROM    #contributionauteur
WHERE   doi IN ( SELECT DISTINCT
                    a.doi
               FROM   #contributionauteur AS a
               LEFT JOIN #corpusauteur AS b ON a.idrow = b.idrow
               WHERE  b.idrow IS NULL
                    AND a.doi IN ( SELECT DISTINCT
                                        doi
                                   FROM     #seul )
               GROUP BY a.doi
               HAVING COUNT(DISTINCT auteurclean) = 1 )
      AND idrow NOT IN ( SELECT DISTINCT
```

```
                    idrow
                FROM    #corpusauteur )


INSERT  INTO #corpusauteur
    SELECT DISTINCT
        24, t.doi, b.idrow, ar.Nb_Auteur, b.contributionclean,
        UPPER(b.Auteurclean) InitialesPlos, t.ID_Art, t.Initiales,
        t.Nom, t.Nom_Famille, t.Prenom, t.Surnom, t.Ordre
    FROM    #tempauteur AS t
    INNER JOIN #tempcon AS b ON b.doi = t.doi
    INNER JOIN Pub_Expanded.dbo.Article AS ar ON ar.id_art = t.id_art
                                AND b.idrow NOT IN (
                                SELECT DISTINCT
                                    idrow
                                FROM   #corpusauteur )
```