Preprint de l'article publié dans I2D - Information, données & documents n°3/2016

Impact de l'Open Access sur les citations : une étude de cas

Frédérique Bordignon[1], Mathieu Andro[2]

1 Direction de la Documentation, Pôle IST, Ecole des Ponts ParisTech, 77 455 Champs-sur-Marne, France

2 DV IST, INRA, 78026 Versailles, France

L'auto-archivage des publications scientifiques a été initié dans les années 1970 par des scientifiques et a connu un grand développement depuis 1991 et le lancement de l'archive des physiciens, arXiv. Il a pris, par la suite, le nom d'« Open Access » en 2001, suite à la Budapest Open Access Initiative (BOAI). Aujourd'hui, on distingue l'auto-archivage traditionnel des archives ouvertes ou Green Open Access du modèle Gold Open Access qui connaît un développement important ces dernières années. Avec ce dernier modèle, ce sont désormais les auteurs et leurs institutions qui financent les frais de publications afin qu'elles soient accessibles en Open Access et, à la différence du modèle Green, directement sur le site de l'éditeur, sans période d'embargo et dans la mise en forme finale.

Des études de grande envergure ont montré une forte augmentation du volume de publications accessibles en Open Access ces 20 dernières années avec des différences en fonction de la nature de la mise à disposition (Green ou Gold) et le champ disciplinaire [1, 2]. 53.9% des papiers parus entre 2008 et 2013 peuvent être téléchargés gratuitement, avec un taux encore plus élevé en Europe (58.6%) [2].
Cette croissance de l'Open Access devrait avoir des conséquences logiques sur la vitesse de dissémination des résultats scientifiques qu'on peut repérer par les citations.

De nombreuses études ont déjà été publiées afin de mesurer l'impact de l'Open Access sur la quantité de citations des publications. Les résultats de ces études, parfois contradictoires, alimentent les débats sur l'Open Access. Alors pourquoi publier une énième étude sur le sujet ? Tout d'abord, force est de constater que peu d'études en langue française ont été produites à ce jour à propos d'un sujet qui préoccupe sans doute autant les professionnels nationaux que les chercheurs internationaux. Ensuite, l'originalité de notre démarche provient du fait que nous avons cherché à collecter les dates de "libération" de chaque article (autrement dit la date à partir de laquelle un document est en accès libre) et sommes ainsi parvenus à mesurer la vitesse de citation avant et après "libération" des articles. Cela permet de mieux comparer l'effet de l'Open Access sur le taux de citation des articles dans la mesure où cela permet de comparer le taux de citations avant et après "libération" et d'éviter ainsi la plupart des biais mentionnés dans la littérature. En effet, la plupart des biais signalés proviennent de la comparaison d'un corpus Open Access avec un corpus non Open Access. Le corpus Open Access pourrait être plus cité car les publications que les auteurs diffusent en Open Access seraient celles dont ils seraient le plus fiers, donc celles susceptibles d'être finalement le plus citées. Mais, si nous comparons un même corpus d'articles avant et après "libération" Open Access, ce biais majeur est évité.

Etat de l'art

Avant de se pencher directement sur une éventuelle relation entre Open Access et citations, il ne faut pas négliger la littérature relatant d'autres facteurs influençant le taux de citations, y compris en dehors de la qualité intrinsèque du travail présenté.

Si certaines études [3] relatent une forte corrélation entre les citations et le nombre de téléchargements (phénomène qu'il est facile de rattacher à une éventuelle diffusion en Open Access), d'autres identifient des facteurs plus surprenants liés aux caractéristiques formelles des articles. Par exemple, une étude ne portant que sur des revues Open Access [4] révèle que les articles dont les titres sont courts reçoivent davantage de vues et de citations que ceux qui ont des titres longs. Les titres qui contiennent un point d'interrogation, qui contiennent la référence à une région géographique spécifique, et ceux qui ont des deux-points ou un tiret sont associés à un plus faible nombre de citations. Les articles dont le titre décrit des résultats sont cités plus souvent que ceux qui décrivent une méthode. Les publications issues d'un financement sponsorisé sont plus citées [5]. Par ailleurs les citations reçues par les papiers peuvent être influencées par le domaine disciplinaire, le genre, l'ancienneté et le statut des auteurs, le prestige de l'institution, la revue, le pays de résidence, par la longueur des papiers (spécialement les *reviews* et ceux qui citent eux-mêmes beaucoup de références, semblent rassembler plus de citations dans certains domaines), par le nombre d'auteurs [6] et même par le nombre d'équations présentes dans le texte [7].

Une multitude d'études prétendent que l'Open Access a un effet positif sur le taux de citations des articles. Le projet OpCit[1] (The Open Citation project) a maintenu pendant 10 ans environ (2004-2013) une bibliographie dédiée aux publications scientifiques portant sur l'impact de l'Open Access sur le taux de citations. C'est désormais le site internet de Sparc Europe[2] qui recense ces travaux et maintient la liste, qu'on trouve aussi dans les mises à jour régulières réalisées par Ben A. Wagner, un bibliothécaire de l'Université de Buffalo [8]. Cela permet d'identifier rapidement une tendance dans les résultats obtenus avec, au jour de la rédaction de ce document, 70 études recensées, 46 démontrant un avantage de l'Open Access sur les citations, 17 ne montrant pas d'avantage de l'Open Access sur les citations et 7 qui ne tirent aucune conclusion tranchée sur la question.
Ainsi, dès 2001, un article publié dans la célèbre revue *Nature* [9] semblait mettre en évidence un accroissement de 157% des citations de publications en Open Access par rapport aux autres à partir d'un échantillon de 119 924 articles de conférences dans le domaine des sciences informatiques. Cette étude fut suivie d'une multitude d'autres qui parvinrent également à des conclusions similaires dans d'autres domaines [8, 10].
La comparaison des résultats est difficile puisqu'elles portent sur des corpus définis par des périmètres et des sources différents ; elles portent en effet :

- sur un domaine disciplinaire ou plusieurs domaines comparés,
- sur une même revue avec le taux de citations comparé entre les articles Open Access et non-Open Access
- sur les documents identifiés dans une archive institutionnelle ou disciplinaire, dans le Web of Science, Scopus, arXiv, Google Scholar...
- sur le taux de citations comparé selon la modalité de publication en Open Access (Green, Gold, delayed Open Access, etc)
- sur une région géographique ou des revues d'une même langue
- sur les revues d'un même éditeur
- sur un ou plusieurs types de documents (différentes versions d'une publication, *working papers*, ouvrages ou chapitres d'ouvrages)
- sur des volumes différents répartis sur des périodes différentes

[1] http://opcit.eprints.org
[2] http://sparceurope.org/oaca

- sur des revues aux indicateurs de notoriété (facteurs d'impact, etc.) différents (parfois volontairement comparés)

La plupart de ces études partent du postulat selon lequel si l'Open Access améliore la visibilité et l'accessibilité de publications, il doit nécessairement et mécaniquement en améliorer le taux de citation. Des réserves ont néanmoins parfois été émises concernant de possibles biais et la difficulté à isoler l'impact du facteur Open Access seul sur le taux de citation des articles. Pour résumer les explications possibles à un impact positif de l'Open Access sur le taux de citations, nous reprenons la typologie ("anatomy") proposée par Stevan Harnad *et al* [11] entre autres, identifiée pour tout ou partie également par Alma Swan et Ben A. Wagner [10, 12]:

- *Early Advantage* ou avantage de précocité. Une recherche rendue publique plus tôt peut logiquement commencer à être utilisée plus tôt ; l'auto-archivage de preprints avant évaluation et publication augmenterait donc l'impact de la recherche grâce à la possibilité pour eux d'être étudiés, réutilisés et donc cités plus rapidement. Cet effet est notamment attendu pour les travaux déposés dans arXiv dans leur toute première forme en raison d'une "fenêtre d'exposition" plus longue. En fait, il n'y a pas un avantage brut de l'Open Access sur les citations d'articles déposés dans arXiv mais en revanche la citation est accélérée. Cette accélération est due au fait qu'arXiv rende les papiers accessibles plus tôt, plutôt qu'au fait qu'arXiv les distribuent gratuitement [12]. Et c'est bien l'objectif premier d'une archive de preprints que d'accélérer la dissémination de la recherche. Mais il semblerait que la publication d'un article en Open Access via l'option proposée par un éditeur soit elle aussi un facteur d'accélération de la citation [13].
- *Quality Advantage* ou avantage de qualité. L'auto-archivage de postprints, après acceptation donc de la publication, augmenterait l'impact sur la recherche puisque ceux-ci sont de meilleure qualité (notamment grâce à la discussion des preprints et les améliorations qui s'ensuivent [13]) et accessibles librement. Ils sont sélectionnés par les auteurs citants pour leur qualité indépendamment des contraintes éventuelles d'accès à cette littérature [14].
- *Quality Bias* ou avantage lié à l'auto-sélection. Un biais lié à la qualité apparaîtrait quand les auteurs sélectionnent leurs meilleurs papiers pour les auto-archiver. Et ce serait la différence de qualité des articles qui mènerait à une différence dans le nombre de citations [15, 16].
- *Usage Advantage* ou avantage d'usage ou de téléchargements. L'auto-archivage augmenterait les téléchargements et donc l'impact sur la recherche. les téléchargements sont des indicateurs précoces du nombre de citations à venir [17].
- *Competitive advantage* ou avantage concurrentiel. Les articles en Open Access sont en compétition avec ceux qui ne le sont pas et dans la mesure où ils sont plus facilement accessibles que les autres, ils seraient cités davantage.

Pour le Quality Advantage et le Quality Bias, le dilemme est de déterminer le lien de causalité entre citations et qualité, autrement dit de savoir si les publications sont librement accessibles parce qu'elles sont aussi les plus citées, ou si elles sont les plus citées parce qu'elles sont librement accessibles. Les 2 facteurs jouent en faveur de l'Open Access puisque les articles les meilleurs trouvent un bénéfice à être accessibles et que les meilleurs papiers sont aussi susceptibles d'être plus auto-archivés [18]. L'hypothèse selon laquelle l'Open Access ne serait pas une condition suffisante mais bien nécessaire est confirmée par Koler-Povh *et al* [19] qui voient dans leur corpus un impact plus fort de l'Open Access sur les articles des revues aux plus forts facteurs d'impact du domaine "civil engineering". Même conclusion pour McCabe *et al* [20] qui en concluent alors qu'il peut y avoir des gagnants et des perdants à l'Open Access.

Dans notre étude nous souhaitons bien sûr vérifier s'il existe un avantage concurrentiel des articles en Open Access par rapport à ceux qui ne le sont pas mais nous souhaitons aussi mesurer l'impact d'une "libération"

précoce des travaux sur leur taux de citations, cette “libération” n’étant pas forcément assurée par l’auteur lui-même, mais parfois par l’éditeur, cela nous éloigne du simple concept d’auto-archivage et nous oblige à décrire avec précision notre corpus pour le positionner par rapport aux définitions de l’Open Access. Car en effet, une revue de littérature montre qu’il n’y a pas qu’une seule définition consensuelle de l’Open Access mais bien plusieurs allant de l’Open Access dit “idéal” à une mise en ligne dont il faut se contenter faute de mieux.

Les papiers accessibles gratuitement, publiés dans des revues à comité de lecture peuvent avoir une grande variété de formes (preprint, html, pdf), être diffusés via une grande variété de media (dépôts institutionnels, sites des éditeurs, agrégateurs de contenus) et avoir différents niveaux de disponibilité (immédiat, différé, transitoire) et les définitions basées sur la distinction entre Open Access Green et Gold ne suffisent plus à décrire la situation et placer l’immédiateté de la mise à disposition comme critère pour recevoir le label Open Access semble désormais trop restrictif [2]. Une étude a d’ailleurs été consacrée [21] à ces revues à Open Access différé, c’est-à-dire aux revues payantes dont les contenus sont “libérés” sur le web par l’éditeur à la fin d’une période d’embargo. Ces revues seraient a priori exclues du périmètre de l’Open Access (idéal) puisque l’embargo est une barrière et que la définition de la BOAI prévoit un accès sans barrière. Cette étude montre que le volume total d’articles libérés après un délai (111 000) est 10 fois supérieur à celui des articles libérés un à un dans des revues hybrides (12 000). Par ailleurs, elle montre que ces revues à Open Access différé reçoivent 2 fois plus de citations que celles sur abonnement et 3 fois plus que celles en Open Access immédiat.

Archambault *et al* ne conçoivent pas non plus leur analyse sur la distinction entre Gratis (gratuit mais pas forcément sans copyright) et Libre (gratuit et avec des restrictions allégées quant à l’utilisation et le copyright) Open Access [2]. Nous allons nous aussi adopter une définition large de l’Open Access pour notre étude et inclure tous les articles auxquels il est possible d’accéder gratuitement, par la voie Green ou Gold, avant ou après leur publication, ou immédiatement à leur parution, n’importe où sur le web via n’importe quel type de sites (éditeurs, archives institutionnelles ou disciplinaires, agrégateurs de contenus, pages personnelles et même réseaux sociaux commerciaux) à l’exception des sites illégaux.

Données, constitution du corpus

Nous avons identifié 347 publications dans des revues à comité de lecture, signées en 2010 de chercheurs de l'École des Ponts. Nous avons utilisé le Web of Science et Scopus pour identifier la majorité de ces publications et nous en avons complété la liste à partir du rapport d’activités de l’établissement pour ne pas négliger les publications en Sciences Humaines et Sociales, très souvent absentes de ces grandes bases bibliographiques.

Pour chacune de ces publications, nous avons cherché à savoir s’il existait une version disponible en ligne via une simple requête dans Google portant généralement simplement sur son titre et nous avons identifié le site “libérateur” et la date de mise en ligne que nous appelons “date de libération”. Cette date a été identifiée directement sur les sites ou avec l’aide des fonctionnalités avancées de date du moteur de recherche Google. Lorsque nous avons trouvé plusieurs fichiers avec des dates différentes pour une même publication, nous avons conservé celui qui avait été mis en ligne le plus tôt.

Nous avons aussi collecté le nombre de citations uniques signalées par le Web of Science et Scopus. Pour les publications absentes de ces bases, nous avons recueilli celles remontées par Google Scholar en les vérifiant toutes une à une, pour être sûrs qu’il s’agissait bien du bon document cité et pour éliminer les citations par des thèses, des sites ou encore des blogs scientifiques (qui n’existent pas dans le WoS et Scopus). L’idéal aurait sans doute été de collecter les citations comptabilisées par les 3 bases pour toutes les références, mais ce type d’extraction étant pratiquement impossible à automatiser depuis Google Scholar, nous l’avons limité

aux références absentes du Web of Science et de Scopus. Cela concerne 40 références qui sont toutes en SHS, il aurait été dommage selon nous de nous priver de l'analyse de ce domaine.
Nous avons ensuite réalisé un laborieux travail de collecte pour dater chaque citation : nous avons exploité en priorité l'API d'Elsevier et des formules Excel, puis nous avons complété manuellement. Nous n'avons évidemment pas conservé la date de publication "officielle" mais bien celle de la publication/mise en ligne effective par l'éditeur, incluant donc les cas de "early-view".

Méthodologie pour l'analyse

Pour analyser nos données, nous avons identifié 4 groupes de publications :

- **Groupe 1 (non-Open Access) :** aucune version électronique n'est disponible gratuitement sur Internet.
- **Groupe 2 ("libération" précoce) :** la date de "libération" de l'article précède la date de publication officielle.
- **Groupe 3 ("libération" tardive) :** la "libération" de la publication est intervenue après sa publication.
- **Groupe 4 (Gold Open Access) :** la publication a été faite directement en Open Access. Nous n'avons que 4 publications dans ce groupe, la pratique de payer pour une mise en Open Access immédiate n'est pas répandue dans l'établissement. C'est pourquoi nous excluons ce groupe de la plupart de nos analyses.
- **Groupe 5 :** le contenu de la publication est bien disponible mais il nous est impossible d'en déterminer la date de mise en ligne. Cela ne concerne que 3 publications dont 2 n'ont jamais reçu de citations. C'est pourquoi nous excluons ce groupe de la plupart de nos analyses.

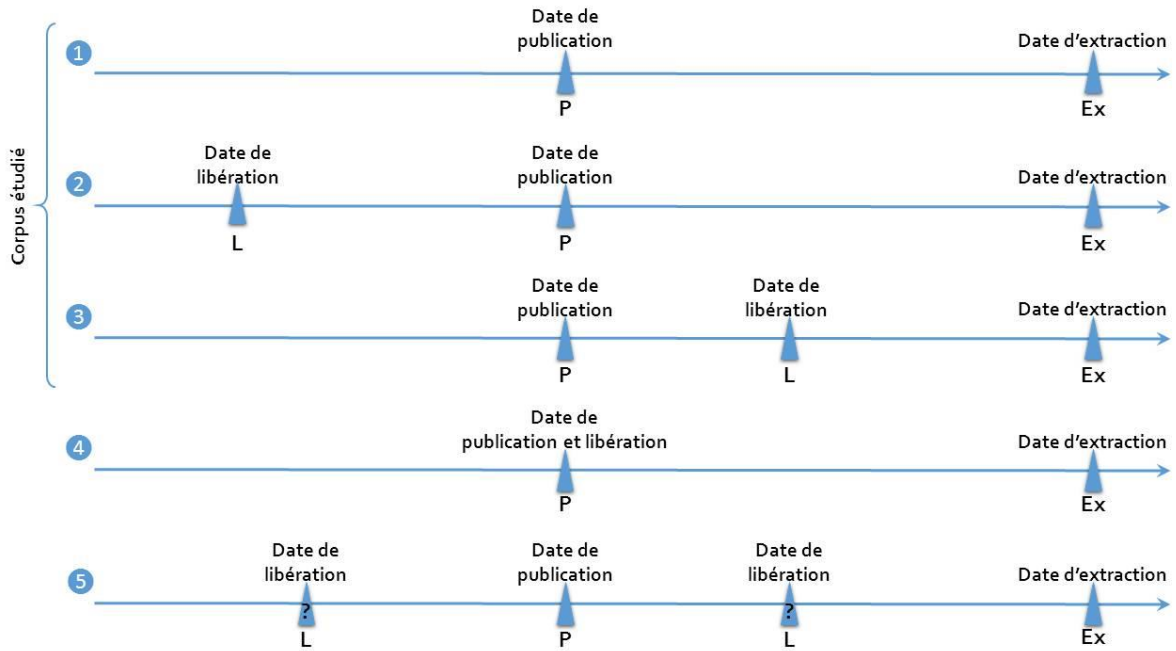

Figure 1 : Caractérisation schématique des 5 groupes du corpus

	GR 1 (non-OA)		GR2 (libération précoce)		GR3 (libération tardive)	
	Total	% du total	Total	% du total	Total	% du total
Publications	118	35%	92	27%	130	38%
Citations	538	18%	1167	39%	1283	43%

Tableau 1 : Répartition des données selon les 3 groupes retenus pour analyse

	GR1		GR2		GR3		
Domaines	Nbre de publis	%	Nbre de publis	%	Nbre de publis	%	Total
Engineering	53	**52%**	26	25%	23	23%	102
Mathematics	20	22%	47	**52%**	24	26%	91
Social Sciences	18	28%	4	6%	42	**66%**	64
Computer Science	18	30%	25	**41%**	18	30%	61
Environmental Science	24	**45%**	7	13%	22	42%	53
Earth and Planetary Sciences	21	41%	8	16%	22	**43%**	51
Physics and Astronomy	12	25%	20	**42%**	16	33%	48
Materials Science	22	**50%**	8	18%	14	32%	44
Biochem., Genetics and Molecular Biology	8	**42%**	5	26%	6	32%	19
Chemistry	2	13%	6	38%	8	**50%**	16
Agricultural and Biological Sciences	6	**55%**	0	0%	5	45%	11
Chemical Engineering	6	**67%**	0	0%	3	33%	9
Economics, Econometrics and Finance	2	25%	2	25%	4	**50%**	8
Business, Management and Accounting	3	**43%**	2	29%	2	29%	7
Decision Sciences	1	14%	6	**86%**	0	0%	7
Arts and Humanities	4	**67%**	0	0%	2	33%	6
Energy	2	33%	1	17%	3	**50%**	6
Medicine	1	25%	2	**50%**	1	25%	4
Immunology and Microbiology	0	0%	0	0%	3	**100%**	3
Nursing	1	**100%**	0	0%	0	0%	1
Psychology	1	**100%**	0	0%	0	0%	1

Tableau 2 : Répartition par domaine des publications des 3 groupes retenus pour analyse

La caractérisation du corpus en 3 groupes est très instructive à l'examen des données par domaine ; le **Tableau 2** montre en effet que 4-5 ans après publication, il « reste » des articles non-accessibles, mais que, contrairement aux idées reçues, les sciences sociales en présentent un taux parmi les plus bas (28%) grâce notamment à des libérations tardives (66%). Par ailleurs, dans notre corpus, en se cantonnant aux domaines les plus représentés (n > 30), on retrouve sans surprise les domaines des Mathématiques, de l'Informatique et de la Physique avec des proportions importantes de publications dans le GR2, conformément aux habitudes de partage précoce que l'on connaît de ces communautés. Enfin, les domaines pour lesquels le GR1 est le plus représenté sont l'Ingénierie, les Sciences environnementales et les Sciences des matériaux.

A ce stade, il est déjà très intéressant de décrypter les habitudes et choix des chercheurs pour la première mise à disposition de leur production. Dans le **Tableau 3**, nous distinguons HAL et arXiv des autres archives ouvertes uniquement pour en montrer l'usage respectif. Nous appelons « agrégateurs » les sites du type de Cairn ou Revues.org. Enfin, nous comptons une publication comme étant diffusée par un site institutionnel

ou personnel lorsqu'elle est effectivement stockée sur ce site et qu'elle n'est pas juste relayée par un lien vers un autre site.

HAL	35%
Sites institutionnels	18%
Agrégateurs	14%
arXiv	10%
Sites éditeurs	8%
Researchgate	8%
Autres archives ouvertes	3%
Sites personnels	2%
Divers	1%

Tableau 3 : Distribution des sites "libérateurs" des publications des Groupes 2 et 3

Plus de la moitié de la production est distribuée via HAL et des sites institutionnels (**Tableau 3**). Et en examinant la répartition des Groupes 2 et 3 en fonction du type du site "libérateur" (**Figure 2**), on constate que les "libérations" précoces se font sur HAL, les sites institutionnels et naturellement arXiv. Les "libérations" tardives sont davantage le fait des éditeurs ou agrégateurs de revues par des fins d'embargos mécaniques. Et sans doute que lorsque la "libération" n'est pas programmée par cette fin d'embargo, c'est plutôt vers HAL et aussi Researchgate que le chercheur se tourne pour diffuser son travail.

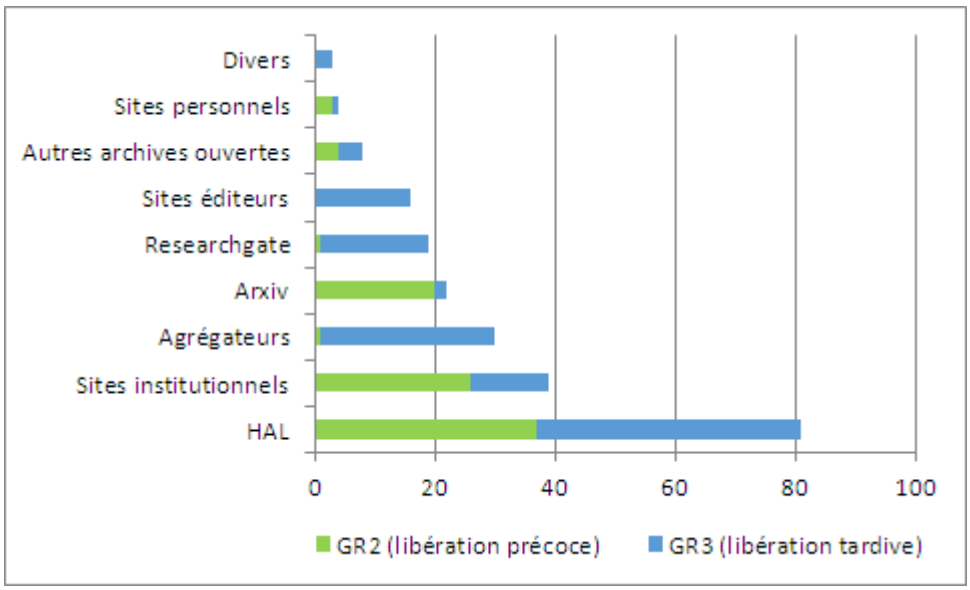

Figure 2 : Répartition des publications des Groupes 2 et 3 en fonction du type de site "libérateur"

A partir de ce corpus, nous cherchons à vérifier les hypothèses suivantes :

1. Est-ce que les publications du Groupe 1 (non-Open Access) sont moins citées que celles du Groupe 3 ("libération" tardive) et encore moins que celles du Groupe 2 ("libération" précoce) ?
2. Est-ce que les publications du Groupe 3 sont plus citées après leur "libération" qu'avant ?
3. Est-ce qu'il est vrai que plus une publication est libérée tôt plus elle est citée et plus elle est citée tôt ?

Comme notre corpus est très hétérogène puisqu'il couvre 21 domaines différents (selon la nomenclature SCImago/JournalMetrics), il est difficile d'avoir une approche globale pour l'ensemble du corpus pour ce qui concerne le nombre de citations. Les habitudes de citations sont en effet différentes d'un champ disciplinaire

à un autre et cette hétérogénéité des pratiques ne se reflète pas de manière identique dans les 3 groupes. Autrement dit, la sur-représentation d'un domaine dans l'un des groupes pourrait constituer un biais. Par ailleurs, indépendamment ou non des domaines de publications, les revues dont sont issues les publications de notre corpus ont des indicateurs de notoriété différents, ou n'en ont pas. Là aussi il est difficile de traiter de la même manière une citation d'une revue au facteur d'impact de 0 avec une citation d'une revue au facteur d'impact de 12 par exemple. C'est pourquoi nous avons procédé à une normalisation des données de citations en utilisant le SJR (SCImago Journal Rank) moyen des 21 domaines concernés.
Comme une revue peut relever de plusieurs domaines, plutôt que de chercher à attribuer arbitrairement un domaine unique à chaque publication (en fonction du champ disciplinaire détecté par le titre par exemple), nous avons préféré procéder à une normalisation à partir du SJR le plus élevé et du moins élevé parmi les domaines concernés ce qui nous conduit à obtenir une fourchette de données suffisamment explicites pour mener à bien notre étude.
Pour chaque publication, nous avons donc :

- c : le nombre de citations à la date d'extraction
- SJR+ : le SJR moyen du domaine pour lequel il est le plus haut
- SJR- : le SJR moyen du domaine pour lequel il est le plus bas
- c+ : le nombre de citations normalisé en fonction de SJR+ (c+=c/SJR+)
- c- : le nombre de citations normalisé en fonction de SJR- (c-=c/SJR-)

Pour calculer le taux de citations mensuel par groupe, nous procédons de la façon suivante :

- Groupe 1 (non-Open Access) = c+ ou c- / nombre de mois entre P et Ex
- Groupe 2 ("libération" précoce) = c+ ou c- / nombre de mois entre L et Ex
- Groupe 3 ("libération" tardive) = c+ ou c- / nombre de mois entre L et Ex
  vs c+ ou c- / nombre de mois entre P et L

Résultats

Le taux de citation selon les groupes

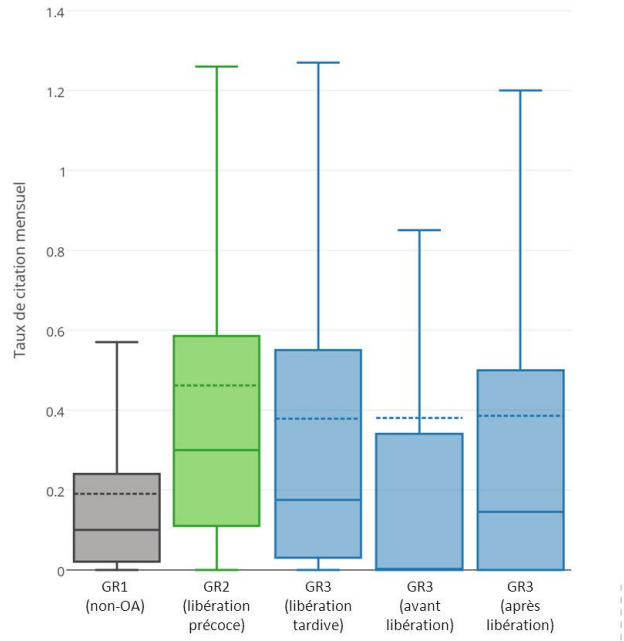

Figure 3 : Taux de citation mensuel selon les groupes avec une normalisation par le SJR le plus bas (moyenne en pointillés, valeurs extrêmes retirées, p < 0,001)

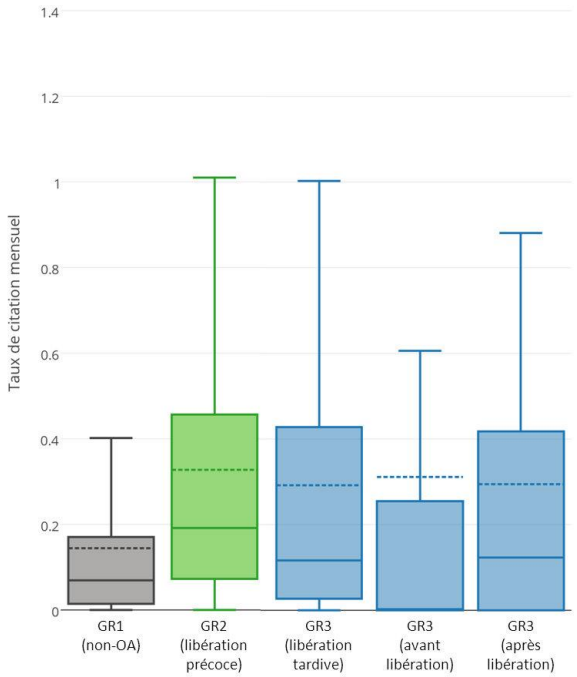

Figure 4 : Taux de citation mensuel selon les groupes avec une normalisation par le SJR le plus haut (moyenne en pointillés, valeurs extrêmes retirées, p < 0,01)

La **Figure 3** et la **Figure 4** sont connues en statistique pour être des « boîtes à moustaches » qui permettent la représentation schématique de la distribution d'une variable, ici le taux de citation mensuel. Ainsi le nombre

total de publications de chacun des groupes (GR1, GR2 et GR3) et sous-groupes (GR3 avant libération, GR3 après libération) est-il réparti en 4 ensembles de même nombre. Le positionnement et la taille des boîtes permet alors de distinguer aisément sur quels taux est répartie la moitié des publications dans chacun des groupes. Les « moustaches » donnent une indication sur l'amplitude de la répartition.
Dans les deux configurations (c+ et c-), on a des résultats statistiquement très significatifs et, en examinant le taux de citation mensuel par publication dans chaque groupe, on peut dire :

- que la "libération" d'une publication, qu'elle soit précoce ou tardive, est nettement plus favorable que la publication non Open Access ;
- que la "libération" précoce est souvent plus favorable qu'une "libération" tardive, avec un taux de citation mensuel médian (trait plein) nettement supérieur pour les publications du Groupe 2 par rapport à celles du Groupe 3 ;
- que la "libération" des publications du Groupe 3 est bien un facteur déclenchant des citations, le taux de citation mensuel médian étant à 0 avant "libération". La moyenne est cependant stable.

Notons enfin que le recours à l'indicateur le plus haut ou le plus bas pour normaliser nos données ne modifie pas en profondeur le profil de nos groupes. D'ailleurs cette distinction ne s'avère pas statistiquement significative ($p > 0,5$).

Le taux de citation selon les domaines

Si nos 2 premières hypothèses semblent ici vérifiées, il faut y apporter une nuance dans la mesure où le **Tableau 4** montre qu'il existe des différences d'un domaine à l'autre et que les sciences sociales se démarquent avec des données qui ne démontrent pas un avantage de l'Open Access sur les citations.

	GR1 vs GR2 La “libération” précoce est-elle favorable ?		GR1 vs GR3 La “libération” tardive est-elle favorable ?		GR2 vs GR3 La “libération” précoce est-elle plus favorable qu'une “libération” tardive ?		Nombre de références		
	c+	c-	c+	c-	c+	c-	c+	c-	Total[3]
Computer Science	oui	oui	oui	oui	oui	non	*11*	*63*	*74*
Earth and Planetary Sciences	oui	oui	oui	oui	oui	non	*48*	*27*	*75*
Engineering	oui	oui	oui	oui	non	oui	*30*	*80*	*110*
Environmental Science	oui	oui	oui	oui	non	non	*32*	*29*	*61*
Mathematics	oui	oui	non	non	oui	oui	*55*	*46*	*101*
Physics and Astronomy	oui	oui	non	oui	oui	oui	*34*	*12*	*46*
Social Sciences	non	non	non	non	non	non	*49*	*57*	*106*

Tableau 4 : Comparaison du taux de citation mensuel moyen par publication selon les groupes et les domaines (p < 0,3 pour c+ et p < 0,4 pour c-)

Il convient évidemment d’appréhender ces résultats avec prudence, les tests montrant qu’ils ne sont pas statistiquement significatifs. Le **Tableau 4** ne présente que les domaines pour lesquels nous avons au moins 30 références dans au moins une des deux configurations de normalisation.

[3] Lorsqu’une référence n’est présente que dans un seul domaine, elle est comptée une fois pour chaque configuration ce qui explique les légères variations de total par rapport au **Tableau 2**.

L'évolution du taux de citation selon les groupes

Afin de vérifier notre dernière hypothèse concernant le rythme de citations en fonction des groupes, nous utilisons les dates de publication des articles citant ceux de notre corpus.

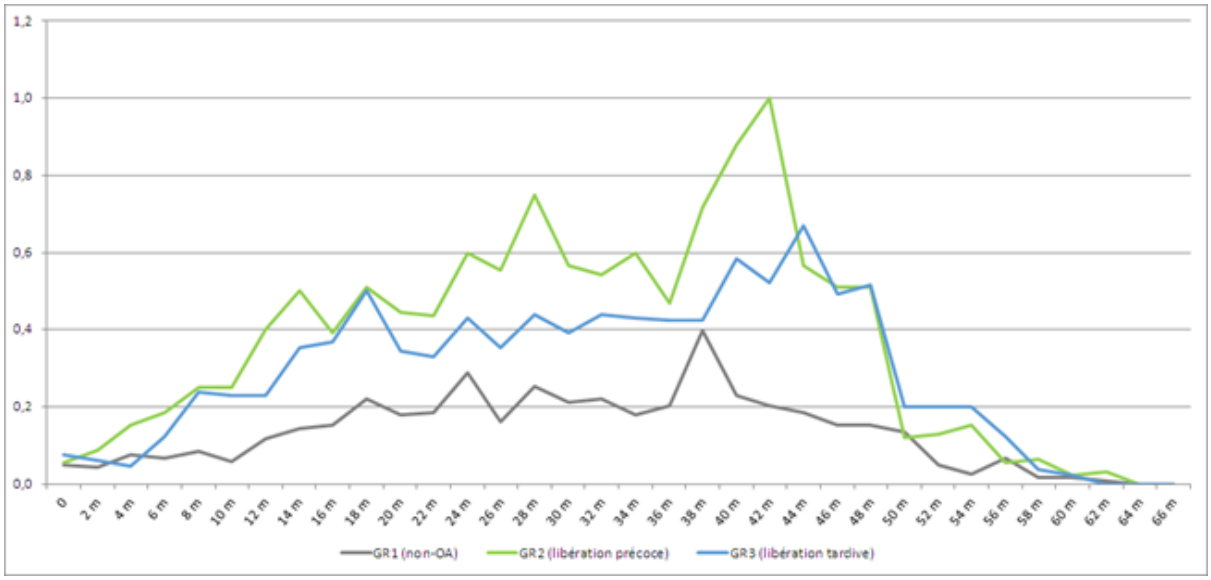

Figure 5 : Évolution du taux de citation mensuel selon les groupes

En abscisses, le temps par tranche de 2 mois, en ordonnées, le nombre total de citations ramené au nombre total de publications par groupe, pour pouvoir les comparer. On voit, de nouveau, clairement que le Groupe 1 reçoit moins de citations que le 3, qui en reçoit moins que le 2. On observe aussi des pics de citations dans chacun des groupes, ils sont d'envergure variable et apparaissent à des moments différents :

- GR1 : pic de faible amplitude apparaissant aux alentours de 38 mois après publication ; demi-vie de citations à 29,9 mois
- GR2 : pic de grande amplitude apparaissant aux alentours de 42 mois après publication ; demi-vie de citations à 30,2 mois
- GR3 : pic de moyenne amplitude apparaissant aux alentours de 44 mois après publication ; demi-vie de citations à 31,9 mois

La demi-vie de citations[4] est autour des 30 mois pour les 3 groupes avec des différences qui ne paraissent pas significatives.

Enfin, pour finir notre analyse du corpus, si on s'intéresse aux articles qui ne sont pas cités du tout, on constate des taux similaires pour les Groupes 1 et 3, en revanche, le Groupe 2 détient un faible pourcentage d'articles non cités.

[4] Nombre de mois entre la publication et la date à laquelle la moitié des citations est atteinte.

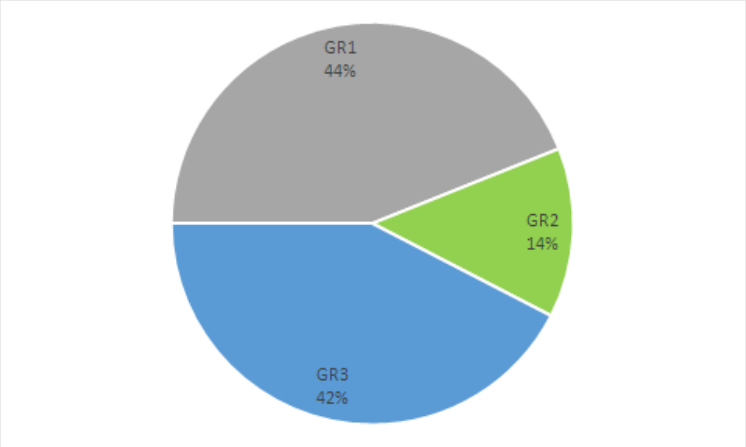

Figure 6 : Répartition des articles non-cités en fonction des groupes

Discussion et perspectives

Une partie des questions soulevées par notre travail porte sur la nature de nos données mais il faut souligner le caractère exploratoire de cette étude et son périmètre volontairement restreint à la production d'un établissement. Il conviendrait aujourd'hui de confirmer nos résultats sur un corpus plus important pour vérifier si les tendances que nous révélons se confirment, et notamment les différences détectées d'une discipline à une autre.

Ce premier travail a cependant permis de montrer combien la collecte des données nécessaires à ce type d'analyses était difficile à automatiser complètement ; il pourrait être intéressant de recourir à du crowdsourcing rémunéré. Il s'agirait de demander à des crowdsourcers recrutés sur des plateformes de vérifier si chaque article est accessible gratuitement et, le cas échéant, à quelle date il a été "libéré", puis de confronter les informations collectées concernant le nombre de citations par au moins 3 travailleurs différents, afin d'en contrôler la qualité, de vérifier leur justesse, et de s'assurer d'une bonne fiabilité des données recueillies. Nous avons envisagé le recours à cette méthode dans un premier temps avant de renoncer, faute de financements et aussi parce que le corpus à traiter était relativement modeste. Néanmoins, cette possibilité pourrait être mise en œuvre dans le cadre d'une étude élargie à d'autres institutions ou portant sur un périmètre thématique ou géographique plus vaste.

Concernant la collecte des références et les sites sources, nous n'avons rien exclu *a priori*, à l'instar d'un chercheur en quête du texte intégral d'un article via Google. Nous avons ainsi considéré, de manière pragmatique, qu'une publication était réputée être en Open Access qu'elle soit diffusée par une archive ouverte, le site de l'éditeur, un réseau social scientifique ou le site personnel de l'auteur. Certaines publications nous ont peut-être échappé parce que mal référencées. Elles ont ainsi peu de chances de trouver lecteur et peuvent difficilement être considérées comme étant en libre accès.

Concernant les résultats, si nos constats s'inscrivent dans la tendance de la majorité des études avec un taux de citation meilleur pour les références en Open Access, la **Figure 3** et la **Figure 4** montrent qu'une libération précoce est souvent plus favorable qu'une libération tardive. Ceci contredit l'étude de Laakso & Björk [21] évoquée plus haut qui concluait à un plus fort taux de citations pour les articles mis en Open Access par les éditeurs après une période d'embargo. Mais peut-être que cette différence est due au fait que nous avons mesuré ici le taux de citation de tous les articles quel que soit le site "libérateur" et qu'il existe peut-être un avantage lorsque la "libération" est effectuée par l'éditeur et pas par l'auteur.

L'examen de la situation par domaine met au jour des différences qui devront être confirmées par des études futures. Si nous n'expliquons pas pourquoi une libération précoce n'est pas plus favorable qu'une libération tardive en Sciences de l'environnement, la situation inverse constatée en Mathématiques correspond bien à la pratique connue des mathématiciens d'auto-archiver leurs preprints dans ArXiv. On s'attendait alors à un profil similaire en Physique mais les résultats semblent plus nuancés. Le **Tableau 4** révèle en revanche très clairement un contexte spécifique des Sciences sociales avec un impact nul de l'Open Access sur les citations. La spécificité des SHS en matière de publication est déjà connue et a abouti à des recommandations qui invitent à tenir compte de ces différences dans la mise en œuvre des politiques d'Open Access [22], le débat portant souvent sur la longueur de l'embargo imposé par l'éditeur avant d'autoriser l'auto-archivage par l'auteur ou sur la durée de la barrière mobile [23–25]. Dans le rapport IPP consacré aux revues SHS [23], les auteurs ont procédé à un suivi temporel des vues (et non des citations) et concluent qu'une barrière mobile pénalise la revue en termes d'audience. Cela rejoint nos conclusions dans ce domaine avec l'absence d'effet positif d'une libération (trop) tardive même si vues et citations sont à considérer différemment, une augmentation de vues n'entraînant pas forcément une augmentation de citations [26].

La **Figure 5** met en lumière un léger décalage pour le Groupe 2 qui reçoit plus tôt et plus fortement des citations. On constate aussi un démarrage plus rapide des citations dès les tout premiers mois de parution, comme si les articles prenaient moins de temps à se faire connaître grâce à leur libération précoce.
Enfin, la diminution des citations est progressive pour le Groupe 1 et brutale pour les Groupes 2 et 3 aux alentours du 50ème mois. Cela nous conduit à dire qu'une "libération" tardive n'ouvre pas une fenêtre plus tardive d'exposition aux citations et que dans tous les cas, passés les 4 ans, les citations sont rares. D'ailleurs, la demi-vie de citation est atteinte quasi simultanément pour les 3 groupes. Nous ne trouvons pas dans la littérature de point de comparaison pour ces données ; cet indicateur est réputé varier d'un domaine à un autre, d'une revue à une autre [27], et mène à des conclusions controversées [28, 29] mais c'est manifestement une proposition originale que nous faisons ici de le croiser avec le mode de diffusion (Open Access ou non). Si nos résultats venaient à être confirmés, cela permettrait de conclure que la libération de la production scientifique n'aurait pas d'incidence sur sa longévité.

Conclusions

L'originalité de cette étude, qui n'est pas la première sur le sujet, tient au fait que nous avons cherché à mesurer les citations mensuelles avant et après la "libération" des articles en collectant les dates de citations. Nous avons ainsi évité le principal biais reproché à de nombreuses études qui est de procéder à la comparaison d'articles issus de corpus différents.
En plus de confirmer comme beaucoup d'autres l'ont fait auparavant un avantage net de l'Open Access sur le taux de citations, nous constatons aussi qu'une "libération" précoce peut avoir un impact plus favorable qu'une "libération" tardive dans certains champs disciplinaires. Cette première étude nous permet dès à présent d'utiliser ces arguments pour inciter les chercheurs de l'établissement à déposer leur production dans une archive ouverte et à le faire le plus tôt possible.

Références

[1] LAAKSO Mikael, BJÖRK Bo-Christer. « Anatomy of open access publishing: a study of longitudinal development and internal structure. ». *BMC medicine*, vol n°10, n°1, 2012, p. 124

[2] ARCHAMBAULT Eric, AMYOT Didier, DESCHAMPS Philippe, NICOL Aurore, PROVENCHER Françoise, REBOUT Lise, et al. « Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013 ». 2014

[3] BRODY Tim, HARNAD Stevan, CARR Leslie. « Earlier Web usage statistics as predictors of later citation impact ». *Journal of the American Society for Information Science and Technology*, vol n°57, n°8, 2006, p. 1060-72

[4] PAIVA Carlos Eduardo, DA SILVEIRA NOGUEIRA LIMA João Paulo, PAIVA Bianca Sakamoto Ribeiro. « Articles with short titles describing the results are cited more often. ». *Clinics (São Paulo, Brazil)*, vol n°67, n°5, 2012, p. 509-13

[5] WANG Jue, SHAPIRA Philip. « Is There a Relationship between Research Sponsorship and Publication Impact? An Analysis of Funding Acknowledgments in Nanotechnology Papers. ». *PloS One*, vol n°10, n°2, 2015, p. e0117727

[6] ROBSON J Barbara, MOUSQUÈS Aurélie. « Predicting citation counts of environmental modelling papers ». In: Ames, D. P., Quinn, N.W.T., Rizzoli AE, éditeur. International Environmental Modelling and Software Society (iEMSs) 7th Intl Congress on Env Modelling and Software, San Diego, CA, USA 2014

[7] FAWCETT Tim W, HIGGINSON Andrew D. « Heavy use of equations impedes communication among biologists. ». *Proceedings of the National Academy of Sciences of the United States of America*, vol n°109, n°29, 2012, p. 11735-9

[8] WAGNER Ben A. « Citation Impact Advantages of Open Access (OA) Articles over non-OA Articles – Updated 12/22/2014 ».

[9] LAWRENCE Steve. « Free online availability substantially increases a paper's impact ». *Nature*, vol n°411, 2001, p. 521

[10] SWAN Alma. « The Open Access citation advantage - Studies and results to date - Technical report ». 2010

[11] HARNAD Stevan, CARR Les, SWAN Alma, SALE Arthur, BOSC Hélène. « Maximizing and Measuring Research Impact Through University and Research-Funder Open-Access Self-Archiving Mandates. ». *Wissenschaftsmanagement*, vol n°15, n°4, 2009, p. 36-41

[12] MOED Henk F. « The effect of « Open Access » upon citation impact: An analysis of ArXiv's Condensed Matter Section ». *ArXiv ID:cs/0611060*, 2006

[13] EYSENBACH Gunther. « Citation advantage of open access articles. ». *PLoS biology*, vol n°4, n°5, 2006, p. e157

[14] GARGOURI Yassine, HAJJEM Chawki, LARIVIÈRE Vincent, GINGRAS Yves, CARR Les, BRODY Tim, et al. « Self-selected or mandated, open access increases citation impact for higher quality research. ». *PloS One*, vol n°5, n°10, 2010, p. e13636

[15] DAVIS Philip M, FROMERTH Michael J. « Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? ». *Scientometrics*, vol n°71, n°2, 2007, p. 203-15

[16] KURTZ Michael J, EICHHORN Guenther, ACCOMAZZI Alberto, GRANT Carolyn, DEMLEITNER Markus, HENNEKEN Edwin, et al. « The effect of use and access on citations ». *Information Processing and Management*, vol n°41, n°6, 2005, p. 1395-402

[17] BRODY Tim, CARR Les, GINGRAS Yves, HAJJEM Chawki, HARNAD Stevan, SWAN Alma. « Incentivizing the Open Access Research Web: Publication-Archiving, Data-Archiving and Scientometrics ». *CTWatch Quarterly*, vol n°3, n°3, 2007

[18] HAJJEM Chawki, HARNAD Stevan. « The Open Access Citation Advantage: Quality Advantage Or Quality Bias? ». *ArXiv ID:cs/0701137*, 2007

[19] KOLER-POVH Teja, JUŽNIČ Primož, TURK Goran. « Impact of open access on citation of scholarly publications in the field of civil engineering ». *Scientometrics*, vol n°98, 2014, p. 1033-45

[20] MCCABE Mark J, SNYDER Christopher M. « The Rich Get Richer and the Poor Get Poorer: The Effect of Open Access on Cites to Science Journals Across the Quality Spectrum ». *SSRN Electronic Journal*, 2013

[21] LAAKSO Mikael, BJÖRK Bo Christer. « Delayed open access: An overlooked high-impact category of openly available scientific literature ». *Journal of the American Society for Information Science and Technology*, vol n°64, n°7, 2013, p. 1323-9

[22] CHARTRON Ghislaine. « Open access et SHS : Controverses ». *Revue européenne des sciences sociales. European Journal of Social Sciences*, n°52-1, 2014, p. 37-63

[23] BACACHE-BEAUVALLET Maya, BENHAMOU Françoise, BOURREAU Marc. « Rapport IPP n°11 – Les revues de sciences humaines et sociales en France : libre accès et audience ». Institut des Politiques Publiques; 2015. 92 p

[24] DARLEY Rebecca, REYNOLDS Daniel, WICKHAM Chris, éditeurs. « Open access journals in humanities and social science ». The British Academy; 2014. 108 p

[25] VINCENT Nigel, WICKHAM Chris, éditeurs. « Debating open access ». The British Academy; 2013. 128 p

[26] DAVIS Philip M. « Open access, readership, citations: a randomized controlled trial of scientific journal publishing. ». *The FASEB journal*, vol n°25, n°7, 2011, p. 2129-34

[27] DAVIS Philip M. « Journal Usage Half-Life ». *Association of American Publishers*, 2013

[28] LARIVIÈRE Vincent, GINGRAS Yves, ARCHAMBAULT Éric. « The decline in the concentration of citations, 1900-2007 ». *Journal of the American Society for Information Science and Technology*, vol n°60, n°4, 2009, p. 858-62

[29] PAROLO Pietro Della Briotta, PAN Raj Kumar, GHOSH Rumi, HUBERMAN Bernardo a, KASKI Kimmo, FORTUNATO Santo. « Attention decay in science ». *ArXiv ID:1503.01881*, 2015