Publications françaises Open Access 2010-2014 - Étude bibliométrique

Valérie Bonvallot, Simone Chrétien, Anne-Marie Badolato

dépasser les frontières

Institut de l'information scientifique et technique

facilite l'accès aux résultats issus des différents champs de la recherche mondiale

Publications françaises Open Access 2010-2014
Remarques méthodologiques et préconisations

Septembre 2016

Introduction

Afin de disposer d'éléments concrets pour éclairer la décision publique, le groupe de travail BSN4 a demandé à l'Inist la réalisation d'une étude bibliométrique sur les publications françaises en Open Access (OA) en science technique médecine (STM) sur une période de 5 ans, de 2010 à 2014, afin d'en connaître le poids dans la production scientifique française et tenter d'estimer le montant des APC.

Ce document présente des remarques et préconisations méthodologiques suite à la réalisation de cette étude. L'étude fait l'objet de deux rapports, l'un traitant de la méthodologie, l'autre des résultats :

- **OpenAccess_BSN4_Méthodologie_30septembre2016.docx**
- **OpenAccess_BSN4_EtudeBibliométrique_30septembre2016.docx**

Un certain nombre de résultats sont consultables via le tableau de bord dynamique créé à l'aide de l'outil libre ezVIS[1], dont l'accès est réservé.

Les limites de chaque étape de la méthode sont évoquées et des préconisations émises :

- Source pour la constitution du corpus : justification du choix du WoS comme la source unique de collecte des publications,
- Sources pour le marquage des périodiques : examen des listes de périodiques utilisées (DOAJ, JCR, ROAD, liste des « Grands Editeurs ») et évaluation des risques d'erreur liés à l'utilisation de ces listes (avec un calcul de taux d'erreur) et à la méthode d'alignement,
- Source pour le repérage des publications OA issues de périodiques non OA : explicitation des difficultés rencontrées avec notamment les limites de CrossRef,
- Sources pour le montant des APCs : défauts de son attribution,
- Répartition du montant des APCs par organisme : limites de la répartition des publications pour 5 organismes et des critères retenus pour déterminer l'organisme payeur.

Source constitutive du corpus : le WoS

Web of Science

La base produite par Thomson Reuters, *Web of Science* (WoS) *Core Collection*, a été retenue pour cette étude pour les raisons suivantes :

- source multidisciplinaire, connue par le plus grand nombre, globalement représentative de la production scientifique mondiale et couvrant la majorité du domaine STM retenu dans l'étude,
- source maîtrisée par le prestataire.

Autres sources envisageables

L'exploitation d'autres sources est envisageable soit pour compléter le repérage des publications françaises en science technique médecine (STM), soit pour initialiser une étude concernant les sciences humaines et sociales (SHS) :

- Scopus : avant son utilisation, il est utile d'étudier son contenu et ses métadonnées. Un rapprochement avec les résultats de l'étude menée - dans le cadre de Conditor - d'une comparaison Scopus-WoS est donc souhaitable.
- PubMed Central (PMC) : si cette base présente l'avantage de fournir une information concernant le statut de l'article, elle concerne essentiellement le domaine biomédical. Son utilisation risque donc de biaiser les résultats des autres domaines. Elle peut être utilisée pour

[1] https://github.com/madec-project/ezvis/

des études spécifiques concernant des organismes ou des thématiques en lien avec cette source.

- Conditor : cette future ressource doit recenser de manière plus exhaustive la production scientifique française. Sa constitution peut prévoir des métadonnées exploitables pour répondre plus facilement à des questions concernant l'Open Access.

Préconisations

Pour le champ des STM, le WoS source multidisciplinaire est satisfaisant d'un point de vue couverture. D'autres sources devront être utilisées pour étendre l'étude aux SHS. A terme, Conditor semble la ressource incontournable qui pourrait d'ailleurs proposer les informations liées à l'Open Access et aux APCs.

Sources pour repérer et marquer les périodiques OA et « Hybride »

DOAJ (Directory of Open Access Journals)

Dans le cadre de cette étude, le fichier DOAJ, source d'information reconnue qui propose un inventaire des périodiques OA, est exploité dans sa totalité, soit plus de 8 800 titres au 30 mai 2016. D'autres approches plus restrictives auraient pu être envisagées.

Périodiques avec un sceau de certification

Ne retenir que les périodiques auxquels le DOAJ attribue un sceau de certification d'OA est trop limitatif puisque l'on arrive à 352 titres. https://doaj.org/faq#seal

Périodiques selon les critères 2014

Ne conserver que les périodiques qui ont été retenus suite à la création de nouveaux critères de sélection en 2014 est envisageable. On passe alors de plus de 8 800 à plus de 4 000 titres. Ces nouveaux critères sont décrits à l'adresse suivante : https://doaj.org/publishers#advice.

Périodiques avec licences « CC by »

Ne retenir que les titres pour lesquels une mention de licence *CC by* est attribuée est également envisageable. On passe alors de plus de 8 800 à plus de 5 900 titres.

Voici les résultats obtenus en fonction des critères retenus :

	Périodiques OA à écarter	Publications à écarter	Part des publications (Corpus 34 660)
DOAJ - sceau de certification	Solution non retenue		
DOAJ - critères 2014	472	10 504	30%
DOAJ - licences « CC by »	212	5 450	15,7%

Tableau 1 : Part des publications à écarter en fonction de sélection de différents critères à partir du DOAJ

Préconisations

Le DOAJ marque plus de 80% du corpus (785 périodiques, 28 214 publications).
9% du corpus est marqué uniquement par lui (173 périodiques, 3 097 publications).
L'utilisation de la liste complète est préconisée pour ne pas diminuer ce pourcentage de 16 à 30% selon les critères retenus.

Le DOAJ propose depuis fin 2015 une API. Elle serait à explorer pour récupérer de l'information au niveau de l'article grâce au DOI et non plus seulement au niveau du périodique.

JCR – ROAD – « Grands Editeurs »

Qualité des sources utilisées : taux d'erreur

Au-delà du DOAJ, plusieurs autres sources sont utilisées pour repérer les périodiques OA ou « Hybride ». Le JCR (Journal Citation Reports) est choisi pour son lien avec le WoS, les listes des « Grands éditeurs » pour tenter de couvrir au maximum le corpus, ROAD (répertoire des ressources scientifiques et universitaires en libre accès) pour tester ce nouvel inventaire.
Pour les listes ROAD et « Grands Editeurs », la date à laquelle le périodique est devenu OA n'est pas mentionnée. Si ce point est important pour l'étude longitudinale 2010-2014, il n'est pas un inconvénient en cas de mises à jour annuelles ultérieures.
De plus, l'absence de nombreux ISSNs dans les listes des « Grands Editeurs » a pu empêcher de repérer des périodiques OA ou « Hybride » (*voir plus loin*).

Pour estimer un taux d'erreur pour l'ensemble du corpus France de 34 660 publications OA, on s'intéresse aux périodiques marqués par une seule des sources, respectivement : le JCR, ROAD, les listes des « Grands Editeurs ».

Pour chacune des sources, un échantillon comprenant les périodiques les plus productifs est constitué (56% des publications pour le JCR, 72% pour ROAD et les « Grands Editeurs ») et leur statut est validé manuellement. Un taux d'erreur est calculé à partir de cet échantillon.

	Périodiques OA*	Publis OA**	Part des publis (Corpus 34 660)	Publis de l'échantillon	Publis non OA de l'échantillon	Taux d'erreur[2]
JCR	220	2 441	7%	1 372	214[3]	1%
ROAD	76	1 836	5%	1 314	230[4]	1%
« Grands Editeurs »	28	1 113	3%	805	0	0%

Tableau 2 : Taux d'erreur pour les listes JCR – ROAD – « Grands Editeurs »

* Périodiques OA : périodiques OA marqués uniquement par la source concernée

** Publis OA : publications OA marquées uniquement par la source concernée

Préconisations

Si la seule liste du DOAJ est utilisée, le corpus final est constitué de 15% (+/-2%) de publications en moins. Ceci étant, la multiplication des sources complexifie les traitements.
Conserver le JCR est nécessaire car il est lié au WoS, source de collecte, et qu'il propose un statut de périodique par année.
La liste ROAD est à surveiller car elle est encore en phase beta et ne propose pas de statut de périodique par année.
Quant aux listes des « Grands Editeurs », la qualité des données accessibles sur leurs sites n'est pas toujours satisfaisante.
La solution serait la constitution d'un réservoir unique d'une liste de périodiques auxquels seraient associées leurs différentes caractéristiques (titres, ISSNs, statut et montant des APC par année, éditeur …) en fonction des années.

Méthode d'alignement des titres et ISSNs

Présentation de la méthode

Le marquage des périodiques est effectué à l'aide d'une comparaison de chaînes de caractères d'une part sur les titres (après une homogénéisation minimale) et d'autre part sur les ISSNs, de manière successive. Un alignement combiné, et non pas successif, Titre + ISSN, risque en effet d'être trop restrictif pour deux raisons :

- Hétérogénéité des titres entre les différentes listes.
- Absence de la donnée ISSN dans certaines listes de « Grands Editeurs » trouvées sur leurs sites.

Risques d'erreur

Les risques d'erreur peuvent être liés à :

- la date de constitution des listes (ROAD et « Grands Editeurs ») : un périodique OA en 2014 sera considéré comme OA quelle que soit l'année traitée (sauf si ce périodique est présent dans la liste du DOAJ qui précise la date à laquelle le périodique est devenu OA).

[2] Taux d'erreur = (Publis OA x Publis non OA de l'échantillon / Publis de l'échantillon) / 34 660

[3] Ces publications concernent : *Santé Publique* qui n'est pas OA, mais dont les articles concernant la période sont accessibles sur internet (à la date du 18 août 2016), et *Inra Productions Animales* (40 publications sans DOI) dont le statut n'est pas clair. Aucun DOI n'étant attribué, nous n'avons pas vérifié l'accessibilité de ces documents. La question est de savoir pourquoi ils sont considérés comme OA par Thomson Reuters.

[4] Ces publications concernent le *Bulletin De L Academie Nationale De Medecine* qui n'est effectivement pas un périodique OA mais « Hybride ». Aucun DOI n'étant attribué, nous n'avons pas vérifié l'accessibilité de ces 230 documents.

- l'absence d'ISSNs dans la liste des « Grands Editeurs » : l'alignement est plus difficile sur le titre seul même si, comme on l'a vu, certains traitements d'homogénéisation sont réalisés.
- des alignements sur des titres identiques mais des ISSNs différents ne représentant pas le même périodique (cf. Tableau 3).
- des ISSN différents, ISSN print dans la notice Wos et ISSN électronique dans les listes de périodiques, ou inversement un même ISSN avec des titres présentant des graphies différentes (cf. Tableau 3).

Bien que ces types d'erreur possibles aient été identifiés, ils sont difficilement quantifiables.

WoS	DOAJ
Nucleus 1949-1034	Nucleus 1678-6602 1982-2278

Tableau 3 : Exemple de titres identiques correspondant à 2 périodiques distincts

Titre WoS	ISSN WoS	Titre DOAJ ou ROAD ou Grands Editeurs	ISSN DOAJ ou DOAJ ou ROAD ou « Grands Editeurs »
POLIMEROS-CIENCIA E TECNOLOGIA	1806-9282	Polímeros	0104-4230

Tableau 4 : Exemple envisageable d'impossibilité d'alignement non rencontré dans l'étude

Préconisations

Utiliser le fichier d'issn.org pour enrichir les listes où les ISSNs sont absents. L'alignement sur les titres conforte la nécessité d'une source « référentielle » pour éviter d'être confronté aux multiples formes d'écriture et pour avoir les différents ISSNs associés à un titre.

Bilan du marquage des périodiques

Taux d'erreur

Si, comme on l'a vu plus haut, le taux d'erreur, pour les sources hors DOAJ, est faible (cf Tableau 2), le taux d'erreur lié à la méthode d'alignement et aux difficultés inhérentes aux titres et aux ISSNs ne peut pas être estimé.

Indépendamment du taux d'erreur, la première limite de l'étude est le taux élevé de périodiques « SansStatut ».

Périodiques « SansStatut »

Sur la période 2010-2014 il y a près de **36% des périodiques dont le statut n'est pas défini** (« SansStatut » : périodique qui n'appartient pas à un « Grand Editeur » et est absent des listes JCR, DOAJ et ROAD).

Pour de futures études, le taux de marquage pourrait être augmenté en effectuant des listes pour des éditeurs supplémentaires. En effet, s'il s'avère qu'il y a plus de 1 300 éditeurs, 12 d'entre eux produisent la moitié des publications issues de périodiques « SansStatut »[5]**.** Pour chacun de ces 12 éditeurs, le statut des périodiques représentant plus de 4% pourrait être obtenu manuellement : une soixantaine de titres seraient alors à considérer.

[5] Amer Physical Soc ; Iop Publishing Ltd ; Royal Soc Chemistry ; Oxford Univ Press ; Lippincott Williams & Wilkins ; Edp Sciences S A ; Masson Editeur ; Spie-Int Soc Optical Engineering ; Cambridge Univ Press ; Amer Inst Physics ; Amer Geophysical Union ; Amer Soc Microbiology

Pour réduire ces traitements manuels fastidieux, d'autres sources pourraient également être exploitées.

Autres sources envisageables

SHERPA/RoMEO – http://www.sherpa.ac.uk/romeo/index.php
SHERPA/RoMEO, qui recense les politiques des éditeurs de revues scientifiques en matière de copyright et d'auto-archivage, propose une liste de périodiques qui fournit l'information sur le statut OA ou non du périodique. Ce site est alimenté par le DOAJ, le service Zetoc de la British Library, et la liste « Entrez » hébergée par le NCBI (*National Center for Biotechnology Information*).
La liste de périodiques n'étant pas déchargeable et les données récupérables présentes sur le site ne présentant pas la mention d'OA (http://www.sherpa.ac.uk/romeo/journalbrowse.php?la=en&fIDnum=&mode=simple), il serait intéressant d'étudier l'API proposée : http://www.sherpa.ac.uk/romeo/apimanual.php?la=en&fIDnum=|&mode=simple.

PubMed Central (PMC)
Le fichier de PubMed Central (PMC) qui liste les périodiques présents dans la base avec la mention OA (immédiat ou suite à un embargo) ou « Hybride » (http://www.ncbi.nlm.nih.gov/pmc/journals/#csvfile) peut être exploité. Même si les périodiques biomédicaux sont favorisés, c'est une piste envisageable.

Source pour repérer les publications OA issues de périodiques non OA

Si compléter le marquage des périodiques OA et « Hybride » et celui des publications OA issues de périodiques OA est réalisable, la difficulté demeure de repérer les publications OA au sein de périodiques « Hybride ». Même s'il progresse, le taux de ces documents est extrêmement faible sur la période avec moins de 2% (584 publications). Ces documents sont repérés grâce à l'application CrossRef.

CrossRef

Sur près de 300 000 DOIs des publications issues de périodiques non OA interrogés via l'API de CrossRef, 584 documents ont répondu aux critères définis dans la méthodologie (0,2% de réponses « positives ») et sont donc considérés comme OA, soit moins de 2% des 34 660 publications du corpus France OA.

Selon un représentant de CrossRef, lors de l'utilisation de la ressource, la mention des licences est renseignée pour environ 5% des DOI. En attendant l'amélioration de l'alimentation de ce réservoir et afin de mieux repérer les publications OA issues de périodiques « Hybride », de nouvelles sources peuvent être exploitées.

Autres sources envisageables

PubMed Central

En interrogeant PMC à partir des DOIs présents dans le WoS, on peut récupérer l'information concernant le statut de l'article. On combine alors cette donnée au statut du périodique repéré à l'aide du fichier de PubMed Central mentionné ci-dessus[6]. De plus, PubMed Central offre un certain nombre

[6] http://www.ncbi.nlm.nih.gov/pmc/journals/#csvfile

de filtres qui permettent de repérer les articles hybrides (ACS Author Choice, Elsevier Sponsored Documents, etc…)

L'étude des APIs (http://www.ncbi.nlm.nih.gov/pmc/tools/developers/) pourrait être intéressante.

Comme déjà évoqué plus haut, les publications biomédicales seraient favorisées. Reste alors la distinction à effectuer entre l'article « purement » OA et l'article accessible après une période d'embargo.

doai.io

Cette application (http://doai.io/), développée par une association française[7] permet de retrouver, à partir d'un DOI, un article OA grâce essentiellement aux données du moteur Bielefeld Academic Search Engine (BASE[8]). Des tests à partir d'une dizaine de DOIs ont été effectués par l'équipe Inist. On peut mentionner que d'une part des articles non OA ont été retrouvés[9] et que d'autre part, des articles OA n'ont pas été repérés. Mais le faible nombre de tests ne permet pas de conclure sur l'efficacité ou non de l'outil. La difficulté est également de connaître la raison de l'accessibilité de l'article : article « réellement » OA ou article devenu OA après une période d'embargo. L'étude de cet outil peut se poursuivre. Cependant l'absence d'API oblige à faire des tests DOI par DOI, ce qui semble difficile à réaliser lorsque, pour une année, plus de 75 000 articles français sont publiés.

Etude de M. Laakso et BC Björk

Une étude longitudinale sur les articles OA issus de périodiques hybrides a été publiée : « Laakso, M., & Björk, B.-C. Hybrid open access—A longitudinal study. Journal of Informetrics (2016), http://dx.doi.org/10.1016/j.joi.2016.08.002 ». Les auteurs pourraient être sollicités pour une recherche, dans leur base, des DOIs du corpus France pour repérer d'éventuelles publications OA.

Sources pour le montant des APCs

Une source centralisée d'information sur les APCs n'existant pas, les données APCs sont récupérées à partir de la liste du DOAJ et d'une liste de périodiques fournie par l'Inserm. Cette dernière complète celle du DOAJ où près des trois quarts des périodiques du corpus France OA n'ont pas de donnée APC. Comme le document concernant la méthodologie le précise, quand un titre de périodique a un APC dans les 2 listes, le montant de la liste proposée par l'Inserm est retenu même s'il y a des écarts conséquents entre les 2 montants proposés.

Malgré ces 2 listes, 237 périodiques (2 605 publications, 7,5%, dont 1 200 concernent un périodique : *Journal Of Physics Conference Series*) restent sans donnée APC. Ces chiffres diminuent si l'on s'intéresse uniquement aux publications pour lesquelles la France (métropole et outre-mer) est mentionnée dans le champ « *Reprint Author* » : 2 089 publications sans donnée APC (6%), 4 898 publications avec 0 pour APC (14%). La seule solution pour améliorer ces 2 points est de traiter individuellement les périodiques concernés.

Il est à noter qu'un seul montant APC pour un périodique ayant été appliqué pour toutes les années, une progression annuelle ne peut être mise en évidence dans cette étude. Elle peut l'être si des mises à jour sont faites ultérieurement.

[7] Association CAPSH (Comité pour l'Accessibilité aux Publications en Sciences et Humanités), responsable du projet Dissemin : http://association.dissem.in/

[8] https://www.base-search.net/

[9] 10.1021/ar200327w ; 10.1021/ar200214k ; 10.1021/cs300538z …

Préconisations

La solution est la même que celle envisagée pour repérer et marquer les périodiques, à savoir : la constitution d'un réservoir unique d'une liste de périodiques auxquels seraient associées leurs différentes caractéristiques (titres, ISSNs, statut et montant des APC par année, éditeur …) pour chaque année.

Répartition des publications OA et du montant des APCs par organisme

Comme le document sur la méthodologie le précise, un point de la demande initiale du groupe BSN4 concerne le calcul de la part des publications OA et de leur coût par organisme. Des traitements, au départ non prévus, ont été réalisés à partir des champs « Affiliations » et « *Reprint Author* » car l'OST n'a pu fournir les marquages qu'il récupère lors des activités de repérages effectués par les organismes de l'ESR.

Répartition des publications OA par organisme

Le choix a priori de 5 organismes (CEA, CNRS, Inra, Inserm, ParisTech) et la recherche sur le seul sigle de l'organisme ne sont pas satisfaisants. Mais, cette première approche a permis de montrer que la méthode de repérage est pertinente car les chiffres concernant les publications OA de l'Inra et de l'Inserm correspondent à ce que ces organismes ont calculé de leur côté.
Pour élargir la répartition aux autres organismes français et affiner les résultats, il est indispensable que l'OST communique les marquages obtenus grâce au travail des différentes institutions ou de récupérer, dans le futur réservoir de Conditor, les marquages créés automatiquement.

Répartition du montant des APCs par organisme

Si la répartition des publications OA par organisme est satisfaisante, il n'en va pas de même pour la répartition des coûts pour les organismes retenus. En effet, l'Inra et l'Inserm ne retrouvent pas les chiffres escomptés.
Une comparaison avec les coûts réels payés par l'Inserm à BioMed Central pour l'année 2014 met en évidence un surcoût de 165 000€ : alors que l'étude attribue un montant d'APCs de près de 250 000€ pour l'Inserm, le coût réel est de près de 85 000€ soit presque le tiers (Pour plus de détails voir l'Annexe 1).

A cela plusieurs raisons possibles, le champ « *Reprint Author* » utilisé pour connaître l'organisme payeur de l'APC :

- n'est peut-être pas l'information la plus appropriée pour connaître l'organisme payeur. Dans ce cas, quelle donnée exploiter ? Nous avons fait une expérimentation sur le champ consacré au mail pour l'année 2014. Elle n'a pas été appliquée à la période. En effet, ce champ :
    - peut donner une autre information organisationnelle autre que celle du champ « *Reprint Author* ». Quelle donnée retenir dans ce cas ?
    - peut également présenter plusieurs organismes,
    - n'est pas toujours renseigné,
    - propose pour les 2/3 des adresses mails des données inexploitables car elles n'ont pas l'extension « .fr »,
    - propose des adresses avec une extension « .fr » mais dont le préfixe ne définit pas l'organisme (« yahoo.fr » par exemple),
    - peut présenter une adresse qui ne reflète pas l'organisme payeur mais l'organisme « hébergeur » : les laboratoires CNRS ont souvent une adresse universitaire.

- peut contenir plusieurs organismes. Dans ce cas, le montant est attribué à chaque organisme, ce qui signifie qu'il est comptabilisé plusieurs fois.

Si le montant des APCs est fortement surestimé, il faut rappeler qu'en plus les publications OA issues de périodiques « hybride » n'ont pas été repérées de manière adéquate.

La réflexion sur la répartition des coûts est donc à poursuivre. Une piste pourrait être l'exploitation des données comptables des organismes s'ils détiennent une information du type DOI ou identifiant d'article (clé UT pour le WoS, PMID *(PubMed identifier)* pour PubMed …) associé à un montant d'APC. Elle serait à l'image de l'initiative allemande OPEN@PC qui propose de recenser les dépenses APC des organismes de recherche (http://treemaps.intact-project.org/)

Préconisations

En plus de la constitution d'un réservoir unique d'une liste de périodiques auxquels seraient associées leurs différentes caractéristiques (titres, ISSNs, statut et montant des APC par année, éditeur …) pour chaque année, exploiter les données comptables car l'utilisation du champ « RP » n'est pas pertinente.

Conclusions et perspectives

Malgré la complexité de l'étude liée à la manipulation de nombreuses informations, éparses, partielles, provenant de différents organismes, présentées sous divers formats, il s'avère qu'en l'état actuel des choses on peut valider globalement la méthodologie pour le repérage des publications OA puisque l'Inra et l'Inserm retrouvent les données qu'ils obtiennent par ailleurs et qu'un taux d'erreur (dans le choix des sources retenues en dehors du DOAJ) estimé à 2% est acceptable.

Ceci dit, avant la répercussion du marquage des périodiques sur les publications, 36% des périodiques sont considérés comme « SansStatut »[10] et 30% comme « Hybride ». Moins de 2% de ces publications ont été repérées par CrossRef comme étant des publications OA, ce qui est vraiment très peu même si ces données recoupent les résultats d'autres études.

Si une mise à jour annuelle est envisagée, les listes des périodiques OA et « Hybride » initialement utilisées seront actualisées et enrichies. Le fichier d'issn.org pourra également être exploité pour améliorer les alignements des titres et des ISSNs.

La proposition de rechercher manuellement le statut d'une soixantaine de périodiques est une solution envisageable pour réduire la part des périodiques « SansStatut ».

La proposition d'explorer d'autres applications (PubMed Central, SHERPA/RoMEO, doai.io) ne garantit pas nécessairement l'obtention de meilleurs résultats conséquents :

- PubMed Central privilégie le biomédical,
- SHERPA/RoMEO s'appuie beaucoup sur le DOAJ et reste au niveau du périodique,

[10] Un périodique « Sans statut » est un périodique dont le statut n'est pas défini car il :
- n'appartient à aucun des 7 « Grands Editeurs » retenus
  ET
- est absent des listes JCR, DOAJ et ROAD.

- doai.io fournit actuellement des résultats insatisfaisants et le mode de recherche sans API est inenvisageable eu égard au volume des données à traiter.

Actuellement, rien n'existe pour repérer de manière satisfaisante les publications OA issues de périodiques « Hybride ». Une alimentation régulière et fiable de CrossRef concernant les mentions de licence par article serait donc un plus non négligeable. Il serait souhaitable de présenter aux éditeurs l'intérêt d'un tel outil pour qu'ils le renseignent systématiquement.
Un contact avec les auteurs finlandais de l'étude longitudinale dédiée aux publications OA issues de périodiques « Hybride » est à envisager[11].
Quant à l'utilisation de PubMed Central, elle entrainera des biais disciplinaires.

Une autre limite de l'étude actuelle est la répartition des publications par organisme qui s'est restreinte à 5 organismes. Cette étape serait indéniablement facilitée par la récupération des marquages de l'OST. Sans l'aide de ce dernier, de nouveaux traitements sont à envisager et pourquoi pas, à moyen terme (2-3 ans) la récupération des marquages automatiques effectués dans la plate-forme Conditor.

Concernant le calcul du montant des APCs actuel, les résultats ne sont pas satisfaisants :

- une forte surestimation des montants dans la répartition des coûts par organisme,
- la limite de l'utilisation du seul champ *« Reprint Author »*,
- des données APC disparates qui entraînent le besoin d'une source d'information officielle pour une mise à jour des données. Le fichier de l'Inserm a permis de combler les manques du DOAJ et des listes des « Grands Editeurs » constituées pour l'occasion,

Même si la proposition de prendre en compte les données comptables des organismes semble difficile à mettre en œuvre (information peu précise, parcellaire, nécessitant des identifiants périodique ou article …), elle est à approfondir.

Dans le cadre d'une mise à jour de l'étude, si la méthodologie pour repérer l'OA peut être conservée, l'estimation des montants APC nécessite vraisemblablement l'exploitation de données comptables.

[11] http://dx.doi.org/10.1016/j.joi.2016.08.002

Annexe 1 : Comparaison des coûts attribués à l'Inserm et des coûts réels pour BioMedCentral en 2014

2014	Attribué à l'Inserm[12]	Payés par l'Inserm	Différence
Articles	122		
Coût	€ 249 792		
Aricles communs	40	40	
Coût	€ 80 815	€ 63 310	€ 17 505
Articles manquants		15	
Coût		€ 21 485	
Articles en trop	82		
Coût	€ 168 977		
Articles		55	
Coût		€ 84 795	€ 164 997

Tableau 5 : Comparaison des coûts estimés pour l'Inserm et des coûts réels pour BioMedCentral en 2014

Pour l'année 2014, alors que l'étude attribue un montant d'APCs de près de 250 000€ pour l'Inserm pour BMC, le coût réel est de près de 85 000€ soit presque le tiers.
Sur les 122 publications OA RP chez BMC attribuées à l'Inserm dans l'étude, 40 ont été effectivement payées par l'organisme. Un premier constat est déjà un écart de 17 500€ de coût pour les mêmes publications identifiées.
15 publications payées par l'Inserm n'ont pas été repérées dans l'étude. A cela différentes raisons :

- 4 sont absentes du WoS,
- 1 vient des bases SSCI et A&H,
- 5 ne mentionnent pas Inserm dans le RP,
- 1 affiche « Inst Natl Sante & Rech Med » dans le RP (seul le sigle « inserm » a été recherché dans l'étude),
- 4 ne sont pas dans le corpus initial (chargées après octobre 2015 ?)

82 publications ont été attribuées à l'Inserm pour lesquelles les APCs n'ont pas été réglés par l'organisme.

[12] Publications et montant des APCs attribués à l'Inserm d'après l'étude bibliométrique